NEWCASTLE UNIVERSITY

DOCTORAL THESIS

# Activity Recognition in Naturalistic Environments using Body-Worn Sensors

*Author:*

Nils Yannick Hammerla

*Supervisor:*

Dr. Thomas Plötz

*A thesis submitted in partial fulfilment of the requirements*

*for the degree of Doctor of Philosophy*

*in the*

Digital Interaction Group

School of Computing Science

December 2014

NEWCASTLE UNIVERSITY

# *Abstract*

Faculty of Science, Agriculture and Engineering
School of Computing Science

Doctor of Philosophy

**Activity Recognition in Naturalistic Environments using Body-Worn Sensors**

by Nils Yannick Hammerla

The research presented in this thesis investigates how deep learning and feature learning can address challenges that arise for activity recognition systems in naturalistic, ecologically valid surroundings such as the private home. One of the main aims of ubiquitous computing is the development of automated recognition systems for human activities and behaviour that are sufficiently robust to be deployed in realistic, in-the-wild environments. In most cases, the targeted application scenarios are people's daily lives, where systems have to abide by practical usability and privacy constraints. We discuss how these constraints impact data collection and analysis and demonstrate how common approaches to the analysis of movement data effectively limit the practical use of activity recognition systems in every-day surroundings. In light of these issues we develop a novel approach to the representation and modelling of movement data based on a data-driven methodology that has applications in activity recognition, behaviour imaging, and skill assessment in ubiquitous computing. A number of case studies illustrate the suitability of the proposed methods and outline how study design can be adapted to maximise the benefit of these techniques, which show promising performance for clinical applications in particular.

# *Acknowledgements*

Pursuing a PhD and writing this thesis would not have been possible without the help and support by my family, so I cannot thank you enough. I also want to thank Diana Nowacka for her support and for putting up with my stressed-out self over the last couple of weeks of writing. This thesis is the result of many collaborations and I want to thank all the people that contributed (and are still contributing) to the success of these projects. I would like to thank my mentor and now also supervisor Thomas Plötz. Without Thomas I would have never found my way to Newcastle and would have missed out on working in such an exciting environment and on so many interesting projects. Thomas nicely balances my utterly chaotic approach with his sense for structure which has led to a lot of excellent work over the years. Next I would like to thank Patrick Olivier for his advice and support ever since I started in Culture Lab as research assistant, back when I wasn't even sure that I really wanted to do a PhD. It turned out to have been an excellent decision and it is the research group that made it that very enjoyable experience that it has been. I would like to thank Peter Andras for his help and insights over the years of this PhD, and I wish him a successful start in his new position at Keele University. Over the years I worked with many collaborators on projects so diverse that I couldn't fit them all into this PhD for the sake of the story-line! I would like to thank James Fisher for the good time we had conducting the study on Parkinson's Disease, and in particular for handling the ethics application that would have taken me forever to complete. A big thank you also for Lynn Rochester and Richard Walker who helped tremendously in the design of that very study and who provided the invaluable medical insights that inspired my analysis approach. I thank all the people in the lab for the good time. Particular thanks go to Cas Ladha, Karim Ladha, and Dan Jackson for providing me endless technical support in using their sensors that made all that data-collection possible in the first place. Also thanks Cas for the good work on all the projects that we worked (and are still working) on together. I would further like to thank Gregory Abowd, Agata Rozga and the staff from the Marcus Autism Centre for giving me the opportunity to work on a very interesting and relevant project in a fantastic environment. Gregory allowed me to join his group for a couple of months which has been an excellent experience and he has always been there to give me advice. Thanks to Rob Thompson for taking the time to proofread this thesis and for correcting all my germanism! Lastly I would like to thank the SiDE research hub for the funding I received without which none of this would have been possible.

# Contents

# Chapter 1. Introduction

The frequency, intensity and nature of our physical activity and behaviour reflects our physical and mental wellbeing, our health and lifestyle. Human movement is studied in a large variety of disciplines such as medicine, rehabilitation or sports. Increasingly human movement is further used as a novel interaction modality in human computer interaction for context-aware interactive applications. A shared aim of all of these fields is to develop automated means of capturing human movement at great detail allowing practitioners to gain a quantitative impression of the way we move. In most cases it would be very beneficial to capture quantitative information about our movement and behaviour in our daily lives, instead of gaining e.g. snapshots thereof in clinical consultations that may be infrequent and not representative of our natural behaviour. This requires a sufficient robustness towards unforeseen behaviour under realistic, everyday, or in other words *naturalistic* conditions. Even though many systems have been devised to capture and analyse human movement, it is not at all common to evaluate such systems in naturalistic conditions, effectively preventing their wide-spread adaptation in peoples' daily life.

Movement sensing and analysis technology faces significant challenges in naturalistic environments such as the private home. On one hand there are very practical constraints surrounding the usability of sensing technology. To be deployed in the private home, a sensing system is required to be sufficiently usable by the target audience, not to require costly instrumentation of the person and the environment and to adhere to privacy constraints. On the other hand there are limitations regarding the type of ground truth annotations that are accessible in such naturalistic environments. It is infeasible to record and manually label 24 hours of video recording from a private home, making such gold standard labelling inherently inaccessible. Both these aspects have significant implications for the design of sensing equipment, the design of data analysis approaches and overall study design for systems aimed to capture realistic data from our daily lifes.

The goal of this thesis is to investigate these challenges in detail and to develop novel approaches for the analysis of movement that address some of the issues of naturalistic environments. We discuss characteristic constraints towards sensing systems in naturalistic conditions and how these issues impact study design and data analysis. We demonstrate how common research methodologies effectively limit the practical use of activity recognition systems in every-day surroundings. In light of these issues and based on case studies we develop a novel approach to the representation and modelling of inertial data captured in naturalistic settings with applications in activity recognition, behaviour imaging, and skill assessment. We illustrate how medical applications in particular can benefit from such an approach in a prototype system for the assessment of disease state in Parkinson's Disease based on a large dataset collected from 34 affected individuals.

## 1.1 Naturalistic environments

The term naturalistic has its roots in the social sciences where *naturalistic observation* is a common research method [Goodwin, 2009]. A naturalistic environment is a setting in which the behaviour of an individual is outside the influence the observer. In practice, this means that study subjects engage in arbitrary activities of their own choosing, perform those activities at their own pace, in their own style, unhindered by e.g. extensive instrumentation, and that such behaviour may be interrupted by unforeseen events at any time. Naturalistic environments can, however, still correspond to a constrained setting, even though such constraints must not be imposed by the observer or study protocol. Consider for example a surgical operating theatre, in which behaviour is not fully arbitrary but which may correspond to a naturalistic setting during an actual surgery. We can come to the following practical definition of a naturalistic environment:

> *An environment is naturalistic if it is not created and if the removal of the observer would not have significant effect on the behaviour of individuals within the environment*.

Naturalistic environments pose significant challenges towards the sensing and analysis of human movement, which affects many different disciplines. The real-life deployment of context-aware systems represents one of the major goals in ubiquitous computing, which has identified many of the issues surrounding such environments. For example, Intille et al. focus on challenges towards activity recognition systems that aim to detect,

segment and identify a pre-defined set of physical activities in sensor data: *"Differences between expectations of how people will behave and how they actually do behave in the complexity of real settings contribute to product failures"*, and that *"Simulation of realistic natural behavior in the laboratory is difficult, because to do so requires reconstructing the environments themselves."* [Intille et al., 2003b]. Other work such as from Poppe et al. focusses on challenges towards the evaluation of such recognition systems: *"Evaluating HCI systems in laboratory settings is likely to cause unnatural behavior of the users. This makes proper evaluation of the system difficult, if not impossible."* [Poppe et al., 2007]. To address these concerns systems have been proposed that continuously adapt to the user's behaviour in naturalistic settings [Choudhury et al., 2008, Blanke and Schiele, 2009]. However, it can be challenging to deploy systems that rely on user cooperation to gather contextual information in populations that aren't technologically adept or are unable to provide feedback due to cognitive decline. This is why particularly systems aimed for clinical applications are usually deployed in well-controlled environments, even if they would benefit most from real-life deployments. For example, a recent review on wearable sensors in Parkinson's Disease only found 3 out of 36 studies to be based on naturalistic data, even though the authors focussed especially on this aspect [Maetzler et al., 2013].

Challenges surrounding naturalistic environments effectively prevent practical, real-life deployment of sophisticated movement analysis systems, even if their application in our daily life would lead to significant improvements in the wellbeing of people that e.g. suffer from degenerative conditions or decreased mobility. There is enormous potential for the clinical use of data collected in non-clinical, ecologically valid environments such as the private home, if these challenges can be overcome, as e.g. summarised by Maetzler et al. (on clinical consultations in Parkinson's Disease):

> *"Measuring such disease-related outcomes objectively (fairly, without bias or external influence), continuously (without interruption), unobtrusively (not involving direct elicitation of data from the user), and with high ecological validity (approximating the real world that is being examined, for example, at the patients' homes), could boost the efficiency and clinical relevance of those visits, and improve patient management.*" [Maetzler et al., 2013].

## 1.2 Activity Recognition in Naturalistic Settings

Issues surrounding real-life deployments are not unique to the field of ubiquitous computing. A good example is the field of speech recognition, which aims to segment and identify speech in audio recordings [Huang et al., 2001]. Systems were developed based on speech samples recorded under idealised recording conditions from a small number of (mostly male) study participants. The high performance achieved under such ideal settings gave a false impression of robustness. In practice, speech recognition often failed in noisy environments such as an office or at least required significant adaptation to the user. Today speech recognition is mature and embedded in many applications, which can be attributed to efforts to collect and annotate large amounts of naturalistic data [Paul and Baker, 1992], and the development of suitable computational tools that are capable of exploiting that data for training (e.g. [Deng et al., 2013]).

As we will show in this thesis, naturalistic settings pose a number of additional challenges compared to e.g. speech recognition, simply due to constraints surrounding the physical sensing of movements in real-life conditions. For practical deployments it is crucial that sensing approaches abide strict privacy and usability constraints of the target population. This can lead to a lack of reliable gold-standard annotation that is usually required during the design of feature extraction and classification approaches in ubiquitous computing. This is summarised by Choudhury et al.

> *"The deployment must protect the user's privacy as well as the privacy of those with whom the user comes in contact. The sensors must be lightweight and unobtrusive, and the machine learning algorithms must be trainable without requiring extensive human supervision.*" [Choudhury et al., 2008]

In practice it is straight-forward, even in naturalistic settings, to collect large amounts of data as long as data does not need to be reliably labelled. However, the most common technical approach to processing such movement data relies on a pipeline approach in which components are tuned to maximise classification performance, and which are mostly unable to exploit large amounts of unlabelled data [Bulling et al., 2014]. The development of activity recognition systems therefore relies on small, possibly non-representative studies to design the recognition system. This suffers from an obvious risk of over-fitting each component to the specific study setting. In order to improve the robustness of such systems novel approaches to the representation and classification are therefore much desired.

## 1.3 Problem Statement

The development of sophisticated and sufficiently robust movement analysis systems to capture quantitative data about human movement in naturalistic environments is a major goal in ubiquitous computing and medical engineering. Obtaining a detailed impression of the way people move in ecologically valid surroundings would have significant effect in clinical applications, with the potential to improve wellbeing for significant parts of the ageing population. In practice, such naturalistic environments pose significant challenges to the sensing and analysis of human movement relating to privacy and usability constraints. While systems have been developed that allow adaptation of recognition systems in cooperation with the user under realistic conditions, such systems can be unsuitable for people that suffer from cognitive decline or decreased mobility. Particularly clinical applications would therefore benefit from novel computational techniques that substitute for existing best-practice approaches to the representation and classification of body-worn sensor data. The goal of this thesis is to investigate whether recent advances in machine learning, which have shown promising performance in similar settings, namely deep and feature learning, are suitable for data captured in naturalistic surroundings; and to explore how study design and performance evaluation can be modified and extended to maximise their benefit.

## 1.4 Contributions and Structure of this thesis

This section gives a short description of each chapter of this thesis and its contributions. As much of the research in this thesis is the result of collaborations it will further highlight where such work was published and what contributions the author of this thesis made to each paper.

**1) Introduction** This chapter highlights how activity recognition systems in ubiquitous computing face significant challenges in naturalistic surroundings. It is imperative that these challenges are overcome in order to develop practically useful tools for the many fields that would benefit from quantative human behaviour monitoring in real-life conditions.

**2) Sensing and Analysis of Human Movement** Many different approaches to the sensing of human movement are employed in ubiquitous computing, and differ significantly in their applicability in naturalistic surroundings such as the private home. This chapter introduces the terms movement, activity and behaviour which are used to give the reader an overview of related work in activity recognition, skill assessment and movement analysis. Each different sensing approach is investigated with respect to usability, infrastructure requirements, resolution and ambiguity. The chapter concludes in a summary description of the challenges of naturalistic surroundings towards sensing systems and motivates the use of inertial, body worn sensing equipment such as accelerometers and gyroscopes.

**3) Automated Assessment of Problem Behaviour in Individuals with Developmental Disabilities** This chapter describes a system for the automated assessment of problem behaviour, episodes of symptomatic behaviour such as aggression, disruption and self-injury in children with autism. It serves as an example implementation of the pipeline approach common in ubiquitous computing in a typical exploratory clinical application. This chapter illustrates common pitfalls for activity recognition systems developed on simulated data-sets, and how the addition of small scale but naturalistic case studies and publicly available data-sets can alleviate such concerns. This work was published in

> Thomas Plötz, Nils Y Hammerla, Agata Rozga, Andrea Reavis, Nathan Call, Gregory D Abowd (2012) *Automated Assessment of Problem Behaviour in Individuals with Developmental Disabilities*, Proceedings of the 2012 ACM Conference on Ubiquitous Computing (Ubicomp), p. 391-400

and nominated for the best paper award. In this work, the author of this thesis designed the technical approach and the (novel) approach for evaluation of the resulting system, contributed to the design of the study protocol and data collection, and contributed to writing.

**4) Feature Learning for Activity Recognition in Ubiquitous Computing** The most common approach to feature extraction in ubiquitous computing corresponds to manually selecting a set of statistical descriptors aimed to preserve characteristics of frames of inertial sensor data. This approach has certain shortcomings beyond being labour

intensive, as it is prone to overfitting to artificial data collection protocols and datasets. Recent advances in machine learning, namely deep and feature learning provide means to automatically exploit large amounts of unlabelled data to estimate suitable descriptors for input data. This chapter investigates the suitability of such methods for activity recognition in ubiquitous computing, concluding that their performance is superior to many other techniques. Further they alleviate some of the issues with naturalistic settings through their ability to exploit large amounts of data, easily collected in such surroundings. This chapter was published in

> Thomas Plötz, Nils Y Hammerla, Patrick Olivier (2011) *Feature Learning for Activity Recognition in Ubiquitous Computing*, Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), p. 1729

In this work, the author of this thesis developed and implemented the feature learning approaches from related literature and contributed to writing.

**5) A Novel Approach to the Representation of Inertial Data – the ECDF Representation** The statistical features typically extracted from inertial sensor data effectively aim to preserve statistical characteristics of frames of data. Due to the characteristics of inertial data many common tools such as histograms and simple measures such as means and standard deviation give an unsuitable impression for classification. This chapter describes a novel approach to the representation of accelerometer data that efficiently preserves statistical characteristics of multi-variate time-series data. This chapter was published in

> Nils Y Hammerla, Reuben Kirkham, Peter Andras, Thomas Plötz (2013) *On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution*, Proceedings of the 17th annual international symposium on wearable computers (ISWC), p. 65-68

In this work, the author of this thesis conceived the approach to the representation of inertial data, designed the experiments and led on writing.

**6) Dog's Life: Activity Recognition for Dogs** Many of the challenges of activity recognition and movement analysis are not limited to human behaviour, but also pose an interesting application and recognition challenge if applied to companion animals. The behaviour of dogs is particularly interesting as they engage in a multitude of activities

throughout their daily life. As dogs are not aware of being recorded on video or that their movement is captured at all, their behaviour is inherently naturalistic, therefore serving as a suitable test-bed for activity recognition in naturalistic environments. This chapter presents the first activity recognition system for dogs that is based on the technical approach described in chapter 4 and chapter 5. While it illustrates the techniques suitability for this application, it further highlights issues with data annotation in naturalistic settings which have significant effect on the performance evaluation and training of classification engines. This chapter was published in

> Cassim Ladha, Nils Hammerla, Emma Hughes, Patrick Olivier, Thomas Plötz (2013) *Dog's Life: Wearable Activity Recognition for Dogs*, Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (Ubicomp), p. 415-418

In this work, the author of this thesis conceived and implemented the technical approach, experiments and evaluation, and contributed to writing.

**7) ClimbAX: Automated Skill Assessment for Climbing Enthusiasts** Even activity recognition systems aimed at constrained environments can benefit from recording and exploiting naturalistic data during their development and evaluation. This chapter describes an automated system for the detection of climbing activity and the estimation of skill parameters specific to climbing performance. It gives insights into how a combination of both naturalistic and scripted or semi-naturalistic data collection can be utilised to develop and evaluate robust activity recognition systems, providing a novel study design particularly suitable for clinical application settings. This chapter was published in

> Cassim Ladha, Nils Y Hammerla, Patrick Olivier, Thomas Plötz (2013) *ClimbAX: Automated Skill Assessment for Climbing Enthusiasts*, Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (Ubicomp), p. 235-244

In this work, the author of this thesis conceived and implemented the technical approach, experiments and evaluation, and contributed to writing.

**8) Assessing Disease State in Parkinson's Disease in Naturalistic Surroundings** Following the recommendations and insights obtained in the case studies of this thesis, this chapter applies the technical approaches developed in chapter 5 and chapter 6

to a clinical application. This chapter describes a system for the automated assessment of disease state in Parkinson's Disease (PD) that is explicitly designed for naturalistic surroundings such as the private home. Inspired by the insights into study design obtained in chapter 8, this work relies on a two-part study on 34 participants with PD in both a naturalistic setting and a clinical environment. This work was published in

> Nils Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, Thomas Plötz (2015) *PD Disease State Assessment in Naturalistic Environments using Deep Learning*, AAAI 2015, AAAI

In this work, the author of this thesis conceived and implemented the technical approach, conceived the novel approach to study design, assisted in data collection and led on writing.

**9) Summary and conclusion** This chapter provides a summary of the insights, recommendations and results of the research presented in this thesis. It highlights how future applications could benefit from the technical approaches described in this work and how future work could extend such approaches to improve performance.

# Chapter 2. Sensing and Analysis of Human Movement

This chapter explores different sensing approaches and their suitability for the use in naturalistic surroundings, introduces different application settings where e.g. activities are detected based on such sensor recordings, and summarises the state-of-the-art approach to analyse sensor data from the dominant modality in ubiquitous computing, inertial body-worn sensors. The challenges of naturalistic surroundings towards each aspect are discussed, motivating a novel approach to the representation and modelling of inertial time-series data that is developed over the course of this thesis.

## 2.1 Definitions

Action, movement, activity and behaviour are terms used interchangingly in literature from the field of pervasive computing and medical engineering. No established definitions of these terms have been devised so far. However, a number of taxonomies have been proposed.

Bobick suggests *movement*, *activity* and *action* to characterise human behaviour in video recordings. According to Bobick, *Movement* is a motion "whose execution is consistent and easily characterized by a definite space-time trajectory in some configuration space', [Bobick, 1997]. As an example consider walking, a highly repetitive motion that shows a very predictable behaviour from one instance to the next. *Activity* on the other hand is a "statistical sequence of movements", [Bobick, 1997], and hence no longer a single, primitive movement. The top level of the hierarchy suggested by Bobick is the *action*, which is defined to "include semantic primitives relating to the context of the motion". To illustrate the three classes, consider a tennis match, where players run toward a ball (*movement*), perform a specific type of swing to hit the ball back (*activity*) and the overall interaction between the two players and the ball characterises the game of tennis (*action*).

An alternative taxonomy is presented by Moeslund et al. in [Moeslund et al., 2006]. The authors describe three classes; i) *motor primitives* correspond to basic motions of e.g. limbs, such as lifting an arm; ii) *actions* refer to sequences of motor primitives; and iii) *activities* which are "larger scale events that typically depend on the context of the environment, objects, or interacting humans" [Moeslund et al., 2006].

Both proposed taxonomies show similarities in describing three classes, with the final class being defined by the semantics of the behaviour observed. However, at the lowest level, Bobick defines *movement* based on it's characteristic repeatability, instead of relying on *motor primitives* as in the taxonomy proposed by Moeslund et al. The small variations when *movements* are repeated, in other words the individual *style* with which e.g. people walk, is crucial to the field of movement analysis, which aims to characterise *how* people execute a movement. Activity recognition in pervasive computing, on the other hand, aims to estimate *what* activities people engage in, without composing activities from low-level motor primitives. As work from both fields is presented in this thesis, the taxonomy by Bobick therefore appears more suitable. However, the terminology of *activity* and *action* used by Bobick can lead to confusion. In order to remain consistent throughout this thesis, this section provides brief sketches of three classes that closely follow Bobick's taxonomy [Bobick, 1997]: *movement*, *activity* and *behaviour*.

**Movement** is a body motion whose execution is consistent between multiple instances and requires little to no conscious effort. Examples for movements include walking, running, waving, but can also include postures such as standing. Movements are mostly repetitive and are performed with an idiosyncratic *style*. Parameters that describe this style can reflect the impact of certain neuro-degenerative conditions such as Parkinson's Disease, as e.g. of interest in *gait analysis*, and can give an impression of motor skill of the individual in e.g. professional sports.

**Activity** refers to motions that do not possess the characteristic repeatability and self-similarity of movements. These motions are often called *gestures*, where examples include "open drawer" or "drink from cup". The variations observed between executions of these activities can be large and are mostly affected by the environment, instead of reflecting an individual movement *style*. Some activities can also be seen as statistical sequences of movements, but not all activities from the literature fit this description. Examples for sequential activities are "brushing teeth" or "doing dishes", both of which are composed of characteristic and repeatable hand movements.

**Behaviour** refers to statistical sequences of activities observed over longer periods of time. For example, a person's morning routine can be seen as a sequence of activities, such as "taking shower" or "dressing". Another example is a kitchen environment where different recipes yield different observed sequences of activities, where recipes can be seen as individual behaviours. Crucially, behaviour shows a self-similarity and repeatability similar to that of movements, simply on a larger time-scale and higher abstraction level.

## 2.2 Sensing Movement

Many different sensing technologies have been developed that allow sensing of movement and behaviour of individuals and groups of people. In general, 4 classes of approach can be identified that allow sensing at different levels of detail; i) *Holistic* approaches capture the entirety of the body and aim to infer individual aspects from that overall representation; ii) *Reductionistic* approaches sense the movement of fundamental components, such as limbs, at varying degrees of detail and compose the overall motion from those parts; iii) *Ambient sensing*, where the environment of an individual such as a home is instrumented; and iv) *Body-worn sensing* where small lightweight, yet high resolution sensors are attached immediately to the human body. For a given application, care has to be taken to choose the most suitable sensing approach. A number of key characteristics can be identified (related aspects from [Zhou and Hu, 2008] in parenthesis):

**Resolution** (Accuracy)
Each sensing modality allows sensing at different levels of resolution. Resolution refers to both the number of degrees of freedom, i.e. the parameters that are measured by the sensing approach, and the temporal and the spatial resolution at which those degrees of freedom are captured. Some sensing approaches can resolve the individual down to the lowest possible scales while other approaches give a low-resolution impression of movement.

**Ambiguity** (Computation) (see [Poppe, 2010])
In some cases, sensors may provide high resolution data that reflects the movement of a subject but remains ambiguous, i.e. the sensor data collected allows more than one interpretation. This ambiguity can be alleviated by relying on prior knowledge in the

form of biomedical knowledge or e.g. statistical models that smooth interpretations over time. Ambiguity comes at added computational cost as e.g. bio-mechanical models of the body are required to resolve the ambiguities.

**Infrastructure** (Compactness / Cost)
Sensing approaches that show very high resolution require significant investments in infrastructure. For example, motion capture relies on a large number of cameras placed at well defined positions around a subject to fully utilise its potential. Simple sensors that are attached to the body do not require any additional infrastructure, which can be beneficial in naturalistic conditions.

**Usability** (see [McNaney et al., 2011])
Even if a sensing approach seems to be ideal to capture the behaviour of interest, it may be difficult to use, difficult to maintain or affect subjects in a way that alters their behaviour. As an example, a sensing solution that relies on multiple sensors placed across the body of a subject may well provide sufficient resolution to capture the behaviour of interest, but may be uncomfortable for the subject. Usability also captures aspects such as expected stigma for e.g. medical accessories, where people fear wearing the device might lead to embarrassment.

The design of a sensing system, which can incorporate multiple of these approaches is always a trade-off between the factors described above. Even the best performing movement analysis approach would not be useful in practice if the requirements of the target audience were inadequately, or not at all, addressed, leading to low compliance or abandonment.

### 2.2.1 *Holistic and Reductionistic sensing*

The most common approach to gain an understanding of movement and behaviour is to capture a representation of the whole human body (holistic) or to use precise tracking of fundamental body-parts which are used to construct on overall representation of the body (reductionistic). Both of these approaches typically rely on some level of computer vision.

Holistic approaches have been applied successfully in well-controlled or otherwise constrained (e.g. clinical) environments, where examples include the assessment of gait

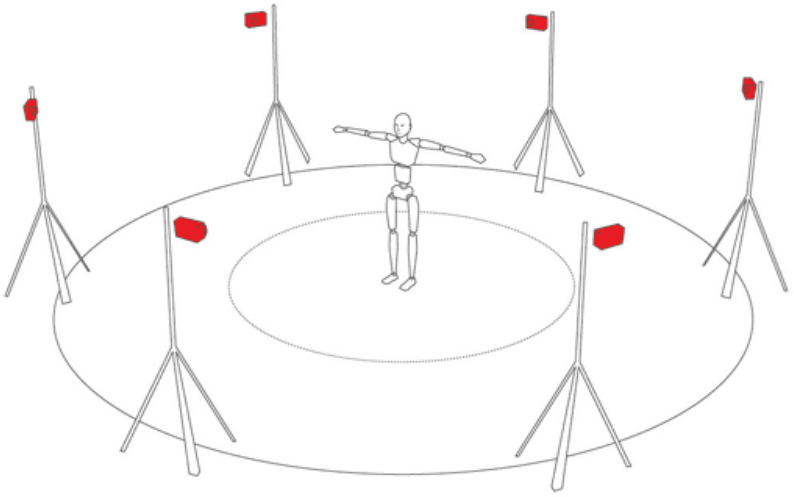


FIGURE 2.1: (Idealised) Sensing setup for reductionistic sensing approaches. A number of e.g. infra-red cameras are arranged in a known configuration around the subject. Small reflective markers are attached to the subject and tracked at high temporal and spatial resolution by the camera setup. From such measurements, a full representation of a skeleton can be obtained with very high resolution and little to no ambiguity.

from video [Lee and Grimson, 2002] and pose estimation from stationary video cameras [Lu et al., 2000, Murphy-Chutorian and Trivedi, 2009]. Holistic approaches have also been employed for human activity recognition from video [Oliver et al., 2002, Robertson and Reid, 2006, Aggarwal and Ryoo, 2011]. Characteristic to the use of video images is that subjects may be obscured, leading to some level of ambiguity. Typically such setups require extensive infrastructure for data transfer and storage of recordings in addition to power supply. Analysis typically relies on some biomechanics model of the human body [Moeslund et al., 2006]. Resolution depends on the location of the camera with respect to the activities of a subject, which may be limited in naturalistic environments.

Reductionistic sensing typically relies on the exact estimation of the location and orientation of fundamental parts, such as limbs, through the optical tracking of infrared-reflective markers attached to key points on the human body (motion capture). The most common approach is to utilise a large set of cameras arranged in an approximate circle around the subject (e.g. 12 cameras, see figure 2.1), that each illuminate the subject with high intensity infrared light. This allows the cameras to track the position of each visible marker with very simple computational means, leading to tracking information at very high frequency and resolution with negligible ambiguity. This has made such sensing technology popular in clinical or otherwise constrained environments, such as gait analysis, or precise tracking of wrist movement [Murgia et al., 2004].

Reductionistic sensing remains the ideal setup to investigate specific medical hypotheses in movement analysis due to the very high sensing resolution. However, this resolution comes at the cost of a significant infrastructure and very low usability. The very expensive cameras have to be arranged in a fixed configuration and calibration procedures have to be performed frequently. Therefore motion tracking is hardly suitable to capture naturalistic environments such as peoples' private homes.

### 2.2.2 *Ambient sensing*

An alternative approach to instrumenting the individual is to instrument entities surrounding the subject. In e.g. instrumented environments, an impression of the movement and behaviour of a subject is obtained by closely monitoring the movement of tools that are handled, e.g. kitchen utensils or low-level sensing infrastructure such as RFID readers, movement, or proximity sensors.

Ambient sensing is particularly popular in activity recognition, where high-level behaviour is inferred from a collection of sensors placed in the environment. Recently there have been plenty of proposed instrumented environments, where some notable examples are *aware home* [Kidd et al., 1999, Abowd et al., 2002], the *MIT PlaceLab* [Schilit et al., 2003], the *gator tech smart house* [Helal et al., 2005], and the *Ubiquitous home* [Yamazaki, 2005]. In some cases, just a single room, most notably the kitchen [Hooper et al., 2012, Olivier et al., 2009, Wagner et al., 2011] is instrumented to allow e.g. context aware interaction. A recent review of *smart homes* can be found in [Chan et al., 2008]. Another area where ambient sensing is employed are professional environments such as car manufacture [Stiefmeier et al., 2008], office environments [Wojek et al., 2006], or instrumented surgical theatres [Ahmadi et al., 2009, Dosis et al., 2005], among many others [Cook and Das, 2007].

The resolution of ambient sensing approaches can be similar to that of body-worn sensors when e.g. tools are handled directly by a person. However, ambient sensing introduces a new analysis challenge, that of detecting when a sensor may provide useful information in a given application scenario. Consider that for example in a kitchen, not all tools are handled simultaneously, yet all instrumented tools may continuously sample movement information. Reducing the number of sensors that are utilised during analysis is therefore beneficial, yet selecting them may be non-trivial. This challenge is usually referred to as *opportunistic sensing*, which is a developing topic of pervasive computing [Roggen et al., 2010, Kurz et al., 2011]. Similar to holistic approaches, ambient

sensing relies on an extensive infrastructure. However, the sensors utilised in ambient sensing (mostly) do not pose a threat to privacy, which renders such sensing setups more suitable for the private home and other naturalistic scenarios.

### 2.2.3 *Body-worn sensing*

In body-worn sensing, small movement sensors are attached to the body to measure the motion directly, without the need for any external sensing infrastructure such as cameras. These sensors are sufficiently small and lightweight to be attached to practically any location on the human body without causing much discomfort or affecting the motion of interest. Usually just a few discrete sensing units are placed on the body even though in the future these sensors might be embedded in sensor-saturated garments [Van Laerhoven and Gellersen, 2004]. The sensors can sense displacement and rotation at high resolution, although they are not yet sensitive enough to provide detailed three dimensional trajectories [Foxlin, 2005]. In contrast to marker-based tracking, body-worn sensor data – even where a large number of sensors are attached to a subject – are not fused to obtain an exhaustive representation of the human body (e.g. the exact posture a subject is in and the orientation of their limbs). Instead, sensors are placed to immediately capture motion of interest, such as on the ankle to estimate swing duration in human gait [Liu et al., 2009]. This is a crucial difference to motion capture technology, where such aspects are usually estimated after a full representation of the human body is obtained. There are some approaches that aim to substitute for motion capture using a large number of body-worn sensors [Vlasic et al., 2007, Tao et al., 2007, Pons-Moll et al., 2010]. Until sensing resolution and sensor noise can be improved, such approaches are unlikely to yield a sensing setup with similarly high resolution as motion capture.

Limiting the number of sensors worn by a subject, even though the data captured may be more ambiguous, has advantages that often outweigh the limited resolution and reliability. Deploying just a few sensors leads to a setup with very high usability, yet sufficient resolution to capture key characteristics of human movement and behaviour [Bergmann and McGregor, 2011]. Arguably, wrist-worn setups are the most popular in pervasive computing [Bulling et al., 2014], due to the high compliance of subjects wearing the sensors, even in ecologically valid settings such as the private home.

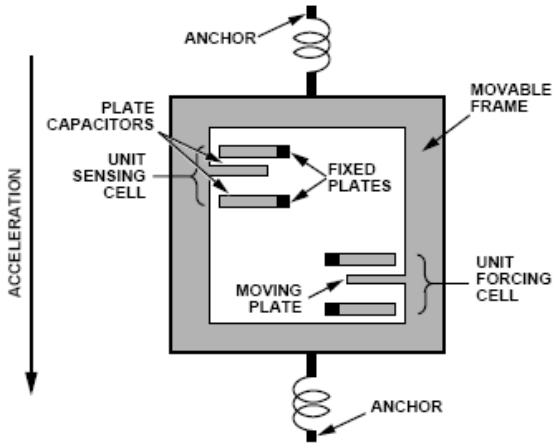



FIGURE 2.2: Working principle of MEMS accelerometer (one axis sensing). A movable frame acts as test-mass whose movement affects measurable properties of capacitors, that allow inference of the proper acceleration applied to the sensor. Three of such accelerometers are arranged perpendicularly to form a tri-axial acceleration sensor on a single die (from [Rob O'Reilly and Harney, 2014]).

**Accelerometers** Accelerometers are devices that measure *proper* acceleration. In contrast to the actual change of velocity of the device, accelerometers measure the effect of the weight of a test-mass relative to a frame of reference of the overall sensing device [Yazdi et al., 1998]. Due to this reliance on a test-mass, they belong to the family of inertial sensing approaches. Accelerometers do not measure *force* applied to the sensor directly, but instead quantify the effect of this force on the sensor, i.e. a possible displacement. Accelerometers are employed in a wide variety of (industrial) applications, where they are used to e.g. measure shock (airbag), vibrations leading to mechanical wear, or to facilitate inertial navigation [Yazdi et al., 1998]. Over the last decade or so, micro machined (MEMS) accelerometers [Judy, 2001] have became incredibly popular for all sorts of consumer devices such as mobile phones and entertainment devices. Their pervasive industrial and commercial application has rendered them very cheap to obtain, which is why they are readily available for applications in body-worn sensing and used throughout the fields of movement analysis and pervasive computing [Bulling et al., 2014].

Most MEMS accelerometers rely on capacitive sensing to measure the proper acceleration a test-mass is subjected to [Godfrey et al., 2008]. This test-mass, usually a movable frame (see Figure 2.2), is held in place by springs that allow movement of the frame relative to the sensor housing along a single dimension. Small fingers extend from the movable frame into capacitors that remain fixed relative to the sensor housing. Movement of the test-mass changes the capacity of the capacitors. From this change in capacity the acceleration can be estimated using simple mathematical means [Yazdi et al.,

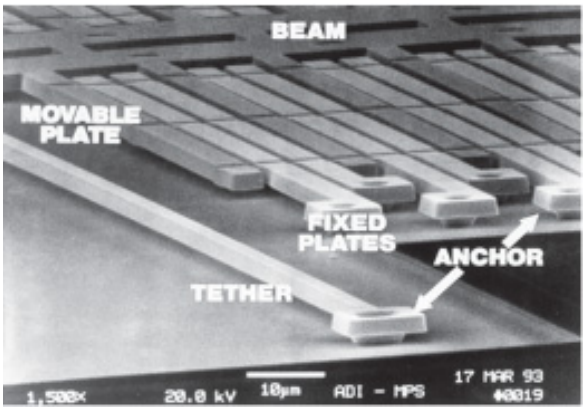



FIGURE 2.3: Annotated picture of a modern MEMS accelerometer (from [Rob O'Reilly and Harney, 2014]).

1998].

The term accelerometer is not totally unambiguous, as it can refer to a single accelerometer with one degree of freedom as described above, or it can refer to multiple accelerometers placed perpendicularly on a single die. Such tri-axial accelerometers can be realised with three independent test-masses, although more modern implementations share their test-mass between the three perpendicularly arranged capacitive sensing structures. At rest, each accelerometer measures the effect of gravity along the sensing direction. This measurement of gravity is beneficial for some situations as it allows the device orientation to be inferred, but is undesirable in many others, as the measurements of the sensors are always a result of the addition of actual acceleration and gravity [Figo et al., 2010]. This constant bias of gravity can make applications such as dead-reckoning difficult to perform and represents a major challenge to the analysis of inertial data [Foxlin, 2005]. Example accelerometer data is illustrated in figure 2.4.

Body-worn sensors relying on accelerometers typically contain a sensing unit, on-board storage and a micro controller for management of the device. Some devices also contain wireless transmission elements to allow networks of such sensors [Korel and Koo, 2010]. MEMS accelerometers are very efficient and require little power to operate reliably. Typical modern sensor devices allow continuous sensing for up to 14 days at 100 Hz. It is important to note, however, that even tri-axial accelerometers do not capture all of the degrees of freedom of the sensor device. While accelerometers allow the estimation of displacement to some degree of certainty, they do not capture any rotation of the sensing device. To capture this rotation, additional sensors have to be included.

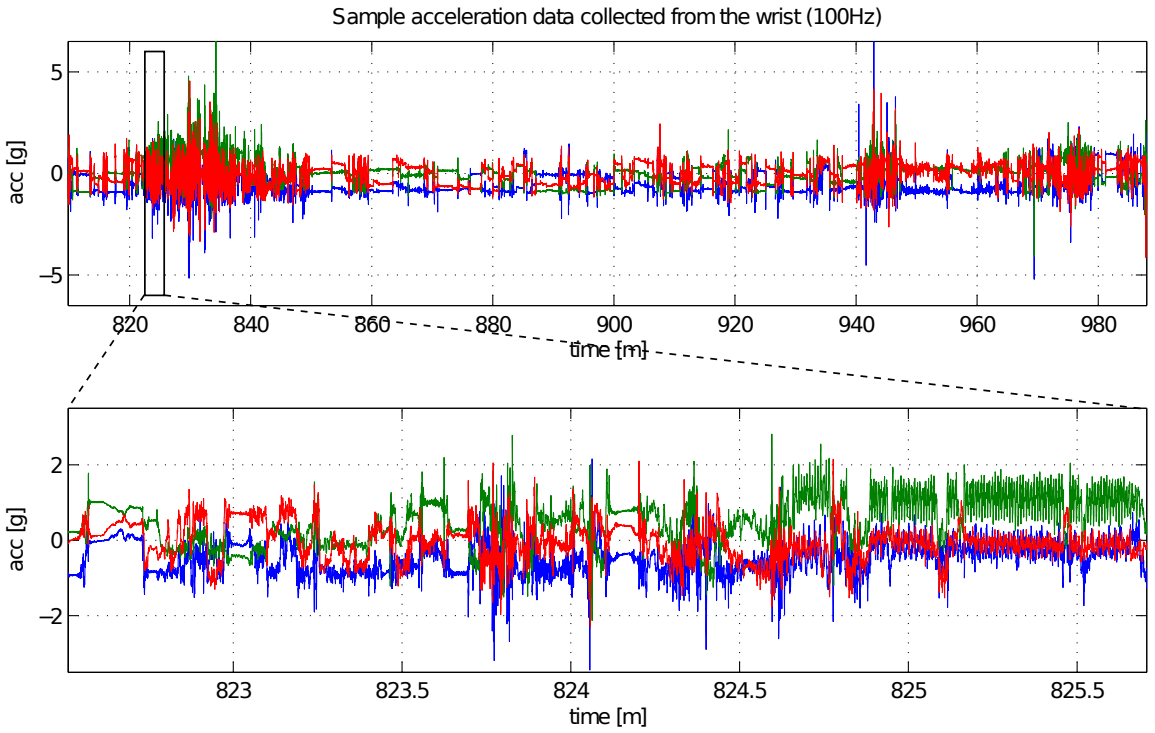



FIGURE 2.4: Sample from a long term recording of human movement captured with an accelerometer attached to the wrist. The lower plot shows an enlarged part of the signal that shows some arbitrary movement and some walking (repetitive parts) towards the right side of the plot. Colours refer to the perpendicular axes of the accelerometer. The impact of gravity on the measurements (deviation from zero mean) is clearly visible.

### 2.2.4 *Sensing in naturalistic surroundings*

Sensing approaches that are invasive towards the privacy of people, for example video cameras or audio recordings, are inherently difficult to deploy in naturalistic settings. Even if video cameras are only used as sensors, i.e. that their recordings are never stored but just used to e.g. detect the presence of an individual, it may be difficult to convince users that their privacy remains intact [Senior et al., 2003]. Deployment of such technologies is therefore likely to affect the behaviour of people in instrumented environments.

Beyond the concerns regarding privacy it can be the sheer cost of deploying extensively instrumented environments to wide parts of the population that represents a practical issue towards their use in every-day life. For example, ambient sensing systems usually rely on a single prototype instrumented environment such as the *aware home* [Abowd et al., 2002]. This raises the issue with data collection in such prototype settings. Even if volunteers are willing to inhabit such homes while their every activity is video recorded

for the sake of annotation (as in e.g. [LaMarca et al., 2005]) it is questionable how naturalistic their behaviour actually is, and how, therefore, such a system would perform if deployed to actual private homes. Ambient sensing therefore aims to adapt recognition systems over time to the behaviour of users, which is one of the main challenges in such approaches [Kleinberger et al., 2007].

Body-worn sensing, on the other hand, does not require extensive infrastructure and does not pose a significant threat towards the users privacy. Their suitability for naturalistic environments, however, depends on the way such technology is deployed in practice. The design of the equipment has to cater towards the requirements of the targeted population, as even simple buttons may pose a significant barrier to people with Parkinson's Disease [McNaney et al., 2011]. Even if very high resolution data can be obtained by attaching multiple sensors to each limb [Vlasic et al., 2007, Slyper and Hodgins, 2008] it is hardly a practical solution for people with reduced mobility, or simply a fear of public stigma [McNaney et al., 2011]. Nevertheless body-worn sensing, used in moderation and kept to a minimum, appears most suited for capturing human movement in every-day surroundings.

The concerns regarding the use of cameras leads to a very practical limitation of sensing in naturalistic settings. The lack of (high-quality) video recordings renders reliable annotation of data captured in naturalistic settings a very challenging task, if it is accessible at all. If video recordings are obtained it can be difficult to identify individual activities and their precise boundaries, even if precise definitions are provided to the annotators (see chapter 6). One approach to alleviate this issue is to obtain labels in cooperation with each individual, for example through the use of diaries (see chapter 8) or specific experience sampling methods which cue the user for recent activities [Intille et al., 2003a]. Another approach is to devise adaptive systems [Van Laerhoven and Cakmakci, 2000] that ask for annotations at times that are suitable [Fogarty et al., 2005] and for activities that are most informative for training a recognition system, referred to as *active learning* [Stikic et al., 2008b, Longstaff et al., 2010].

The suitability of methods that rely on the cooperation by the user depends on the specific application. Usually such techniques are developed to adapt systems to the idiosyncrasies of the user, for example through the use of mobile phones Lane et al. [2011]. Populations that suffer from cognitive decline or reduced mobility may however have difficulty with this type of interaction, as even the act of pressing simple buttons may be significant barrier McNaney et al. [2011]. Even if such methods can be applied the resulting annotation is still subject to boundary issues or class confusion and overall of

less quality compared to video annotations, which has to be considered when designing analysis approaches.

## 2.3 Applications

The sensing techniques presented in chapter 2.2 have been applied for a wide variety of applications in both ubiquitous computing and medical engineering. While those applications share a number of techniques the fields remain mostly separate, illustrated by the vastly different focus of recent review articles such as Bulling et al. [Bulling et al., 2014] and [Godfrey et al., 2008]. This section describes two application scenarios in detail. *Movement analysis and motor skill assessment* aim to quantify the *style* with which activities and movements are performed by an individual. *Activity and Behaviour recognition* instead generalise across those idiosyncrasies to detect and segment activities and more high-level behaviour.

### 2.3.1 *Movement analysis and motor skill assessment*

Movement analysis is concerned with characterising the style with which humans execute specific movements such as walking. The term "movement analysis" is used in the fields of bioengineering and rehabilitation to describe analytical approaches that characterise this movement style based on established parameters, mostly inspired by medical prior knowledge. The most common movement that has been analysed in detail is that of gait, with the aim to estimate *how* people walk or run, and how certain degenerative conditions such as Parkinson's Disease [Sofuwa et al., 2005, Hausdorff, 2009] or stroke [Von Schroeder et al., 1995, Mulroy et al., 2003], affect this movement. Gait analysis has also been used to characterise the impact of interventions such as hip arthroplasty [Madsen et al., 2004]. Gait analysis is a mature field and employs virtually all sensing modalities, ranging from instrumented environments and body-worn sensors to high resolution motion capture technology. The aim of movement analysis is to obtain clinically relevant parameters, with the possibility of applying such technology in every-day life.

Another type of movement that has been studied extensively is the *sit-to-stand* (STS) transition. The (in)ability to perform the STS movement, i.e. getting up from a chair,

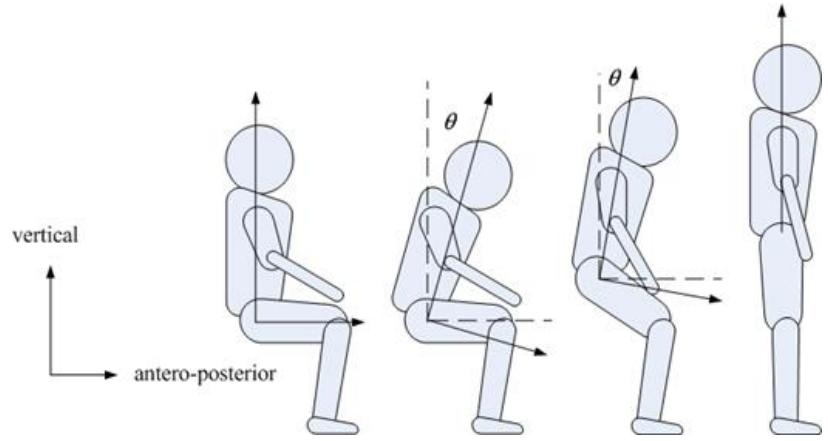



FIGURE 2.5: Sequence of phases in the sit-to-stand-transition. From left to right: i) *flexion-momentum phase*, ii) *momentum transfer phase*, iii) *extension phase*, iv) *stabilisation phase*.

can have significant impact on the quality of life of older people, and can lead to "institutionalisation, impaired functioning and mobility in activities of daily living, and even death." [Janssen et al., 2002]. Approaches to characterise the STS movement are a prime example for the field of movement analysis. An apparently simple movement is captured in laboratory environments using very elaborate sensing techniques that aim to record even the slightest deviation in execution of the movement. Sensing involves (multiple) video cameras, high resolution motion capture, body-worn sensors and an instrumented environment [Janssen et al., 2002]. The movement itself is split into well-defined *phases* that are initiated with certain events, such as the foot leaving the floor (see figure 2.5). Based on the high resolution sensing information the movement is segmented into phases from which specific parameters such as timing or stability are extracted through comparison between individuals.

Of course, such detailed information, particularly about the impact of certain conditions or to illustrate the effect of rehabilitation, is of immense value to the medical community. However, the approach to capture movements at such sensing resolution has an inherent shortcoming: it is very unlikely that the movements can be sensed at a similar resolution in everyday surroundings, such as the private home. Yet the analysis approach relies on exact estimation of the different phases of the movement in order to estimate the parameters of interest, which is unlikely to be very reliable when some uncertainty is introduced at the sensing level. In order to address this shortcoming, recent work aims to develop reliable sensing solutions that allow capturing of movement at a detail similar to that in lab environments in the private home (e.g. [Vlasic et al., 2007, Tao et al., 2007, Pons-Moll et al., 2010]).

Motor skill assessment is similar to movement analysis in the shared aim to characterise the style with which movements are performed. When people learn a new activity, such as cycling, it is common to divide the learning process into three subsequent phases [Fitts and Posner, 1967]. In the first, *cognitive phase*, people explore possible solutions to a given problem by following different strategies. Strategies that are efficient in reaching (sub) goals are retained while inefficient ones are discarded. In the subsequent *associative phase*, the strategies that are discovered in the first phase are further refined by small alterations aimed to improve efficiency. If the activity is trained continuously, the final *autonomous phase* may be reached in which performing the activity requires little to no conscious effort. Motor learning basically describes the process in which *activities*, as defined above, evolve into *movements* by extensive training, with approaches aiming to characterise this progress directly [Hammerla et al., 2011]. For example, the serve of a novice tennis player hardly fits the description of a movement, simply because it misses the characteristic repeatability. If that novice trains for many years, subsequent serves will become more and more self-similar, and are finally being performed with an idiosyncratic style. The serve therefore evolves from being an activity, into being a repeatable movement. The progress of an individual from novice to professional is reflected in the style with which the activity is performed. Characterising this style and mapping the resulting parameters onto some performance scale is of interest in the field of motor skill assessment.

In difference to movement analysis for medical applications, the tiny deviations between executions of activities such as surgical suturing are not always known a priori, but are discovered during analysis of the performance of multiple subjects. Typically, some level of ground-truth annotation for the motor skill forms the basis for an automated process to discover the specific parameters that characterise the performance of an individual. The sensing setup is usually less extensive and aimed at realistic, naturalistic deployments in real-life applications. So far, most applications of skill assessment proposed in the literature aim to characterise people's motor performance in sports, such as running [Strohrmann et al., 2011, 2012], tennis [Ahmadi et al., 2010, 2006], swimming [Bächlin et al., 2009], weight-lifting [Chang et al., 2007, Adelsberger and Troster, 2013, Velloso et al., 2013], golf [Grober, 2010], rowing [King et al., 2009b], rock climbing [Ladha et al., 2013], and snow sports [Holleczek et al., 2010, Michahelles and Schiele, 2005]. In professional environments, systems have been proposed to assess the state of hand-operated tools [Rehorn et al., 2005] and the already mentioned assessment of surgical skill in [King et al., 2009a, Smith et al., 2001, Trejos et al., 2008]. Similar approaches

are furthermore employed to assess progress in rehabilitation [Möller et al., 2012, Kranz et al., 2012].

So far, each individual application of skill assessment in pervasive computing relies on an individually crafted analysis, specific to the requirements of the activity of interest. Some work focusses on detecting mistakes in the execution of e.g. weight-lifting [Velloso et al., 2013] or balance-board exercises in rehabilitation [Möller et al., 2012] in order to provide contextual feedback, aiming to guide the user to an improved execution of the activities. In cases where the movements are more complex, such as in rock climbing [Ladha et al., 2013] or golf [Grober, 2010], explicit performance parameters are extracted from the sensor signals that motivated by prior knowledge, which is described in more depth in chapter 7. An approach suitable for a wide range of applications has, so far, not been devised. However, the increasing self-similarity of activities with training should form a suitable basis for immediate assessment of motor skill. A first approach was presented in [Hammerla et al., 2011], which aims to measure the motion efficiency using principal components analysis (PCA).

### 2.3.2 *Activity and behaviour recognition*

Where movement analysis and skill assessment aim to characterise movements and activities, the automatic recognition of such activities is the aim of (human) activity recognition (HAR). Typically some level of sensing information is gathered, based upon which a corpus of application specific activities is differentiated automatically.

The goals and applications of HAR are very diverse. The information collected through HAR about the activities and the behaviour of a user can aid computing systems to proactively adapt in context-aware computing [Abowd et al., 1998]. Early work in HAR relied on computer vision, where gestures or activities are detected in video recordings from a stationary camera or from still images. Examples for such systems are the recognition of american sign language [Starner et al., 1998a, Grobel and Assan, 1997] and applications in sport such as tennis, football or ballet [Efros et al., 2003].

Body-worn sensors have been used to detect an incredibly wide range of activities for different applications. Early work aimed at estimating energy expenditure [Trost et al., 2005] and modes of transport such as running, cycling [Le Masurier et al., 2003]. Household activities are particularly popular with activities in the kitchen related to cooking, cleaning, and personal hygiene [Hooper et al., 2012, Ward et al., 2002, Bao

and Intille, 2004, Ravi et al., 2005, Logan et al., 2007]. In professional environments, systems recognising activities related to quality control in automobile manufacturing [Stiefmeier et al., 2008], and activities in an operating theatre [Padoy et al., 2008, Bardram et al., 2011] are some examples. HAR systems have been applied to different sports [Avci et al., 2010], such as climbing [Ladha et al., 2013], tennis [Ahmadi et al., 2010] or golf [Grober, 2010], where detected activities (and their boundaries) are utilised to inform some level of skill assessment procedure (see Section 2.3.1).

One major research goal of HAR is to facilitate novel applications surrounding the older population, addressing issues such as independent living and medical diagnosis [van Kasteren and Krose, 2007, Van Kasteren et al., 2008]. Examples include "smart" environments such as the kitchen, which provide contextual cues aimed to support people with Dementia [Olivier et al., 2009], the assessment of severe (symptomatic) behaviour in autism [Plötz et al., 2012b] (see also chapter 3), and rehabilitation [Lo et al., 2007].

The type of activities that form the corpus for a HAR system depends on the specific requirements of the proposed application. The field of HAR is very diverse in both the technical approach to sensing and recognition, along with very diverse application scenarios. A recent, excellent review by Bulling et al. identifies a number of key challenges in the field of activity recognition [Bulling et al., 2014]:

**Application challenges** refer to problems faced when implementing HAR systems in real-world environments. The challenges towards sensing and analysis of movement captured in naturalistic settings fall into this category, but there are other technical problems that arise with practical deployments. An example are the *variability in sensor characteristics* that stem from a variety of (external) sources and may provoke unforeseen behaviour by the HAR system (e.g. misclassification). For example, inertial body-worn sensors may be impacted by environmental aspects such temperature and experience additional noise or drift [Yazdi et al., 1998]. Different models of mobile phones may differ strongly in their sensor characteristics such as sensitivity, spatial or temporal resolution, where e.g. depending on the phone "the accelerometer returns samples to an application unpredictably between 25-38 Hz" [Lane et al., 2010]. Such environmental aspects are difficult to simulate in a lab-like environment. Sensors attached to the body can further be in unexpected orientations, or their orientation or position can change over time [Kunze et al., 2005]. Beyond these hardware-related aspects there are *tradeoffs in human activity recognition system design*. Some applications

may require real-time processing of sensor signals, such as interactive systems, which imposes additional requirements to be considered at design time [Yan et al., 2012].

**Common research challenges** refer to problems common to fields in pattern recognition, some of which are particularly apparent in HAR systems. This includes *intraclass variability*, which refers to the same activity being performed very differently between individuals, which can be used for identification [Chang et al., 2009]. Being of prime interest in movement analysis and skill assessment, this idiosyncratic performance of activities can have a negative impact on HAR systems as generalising across individuals may be difficult. Low *Interclass variability*, on the other hand, refers to activities that show a very large pairwise similarity, making it difficult to differentiate between the two, for example walking up and down stairs Kwapisz et al. [2011]. Lastly, the *NULL class problem* refers to the difficulty of defining a suitable set of background activities. The activities of interest often occur rather sporadically. Making sure that the activities are differentiated from all other, possibly similar activities likely to occur, requires a significant amount of background data to be collected with an adequate diversity. It can further be difficult to estimate the boundaries of when activities occur in light of an arbitrary background, as further elaborated in chapter 6.

**Challenges specific to HAR** include aspects relating to the definition of activities of interest and the collection of HAR specific datasets. For most applications of pattern recognition, such as speech recognition, the definition of individual classes (here: phonemes) are well motivated in (physiological) prior knowledge. Segmenting and labelling of occurrences of these classes is therefore straight-forward with little to no ambiguity. In HAR it is much more difficult to establish definitions for each activity of interest. For example, when exactly does the activity of "walking" start? Even such a simple example requires exact definition, which is not always easy to establish (see chapter 5). A number of taxonomies have been introduced in the literature in order to find a standardised corpus of activities. The most popular example are the activities of daily living (ADLs) [Katz, 1983], which include typical house-hold activities and activities related to personal hygiene. Work where daily-life of e.g. older people is simulated in a lab-like environment often employs the ADLs in their study protocols [Hoff et al., 2001, Maetzler et al., 2013]. Another approach to establish a set of activities is to use physiological measurements such as heart-rate to find activities of similar energy expenditure

Ainsworth et al. [2011]. Such corpora serve as an inspiration for many practitioners when finding suitable background activities to alleviate the null class problem.

## 2.4 Analysis Pipeline

Most applications of inertial sensing rely on a standard processing pipeline, or *Activity Recognition Chain*, which "is a sequence of signal processing, pattern recognition, and machine learning techniques that implements a specific activity recognition system behaviour" [Bulling et al., 2014]. Figure 2.6 illustrates the individual steps that are are performed in sequence and comprises a total of 5 steps. During *Data recording*, measurements from one or more sensors are captured over a long period of time. This data is then transformed during *preprocessing*, which aims to alleviate the impact of sensor noise, artefacts, or impact of the environment. The resulting signal is then split into short, continuous sections that are likely to contain the activities or movements of interest during *segmentation*. During *feature extraction*, characteristic properties of each segment are extracted that aim to abstract from the raw sensor recordings and allow the activities of interest to be differentiated. Finally, a *classification* engine is used to produce an hypothesis for each extracted segment, often by assigning a probability to each activity class of interest.

Many variations of this analysis pipeline have been proposed which may skip or extend each individual step. Features are not extracted explicitly in some settings where classifiers are applied immediately to the segmented data. This is common for artificial neural networks (ANNs), where a feature extraction is performed implicitly, parametrized by the learned weights within the network [Hinton and Salakhutdinov, 2006]. Other approaches that estimate the similarity of the segmented signal to a number of prototypes, using e.g. dynamic time warping [Pham et al., 2010], also do not extract explicit features. In applications where segmentation is difficult as e.g. boundaries of activities are not well defined, more complex statistical learning approaches such as Hidden Markov Models (HMMs) do not rely on a segmentation step. A possible segmentation of the input data is instead part of the hypothesis produced in the classification step [Minnen et al., 2007]. Some systems add another step after classification, where e.g. the outputs are smoothed using a statistical model such as a Markov Chain, which incorporates prior knowledge about likely sequences of otherwise independent activities and can lead to drastic improvements in recognition performance. This is of particular importance for the modelling of behaviour.

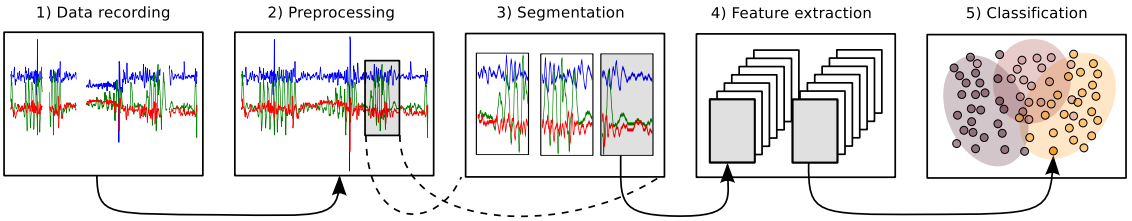



FIGURE 2.6: Typical analysis pipeline for inertial time series.

### 2.4.1 *Preprocessing techniques*

The data captured from inertial sensors is subject to noise, may show undesirable artefacts and is influenced by environmental aspects such as humidity and temperature. If data is captured from multiple sources and across multiple modalities, each sensor stream will have an individual sampling rate. Furthermore temporal frequency may be reduced on purpose to conserve energy. The recordings from $n$ sensors are commonly described in vector notation [Bulling et al., 2014]

$$\mathbf{D} = \left(\mathbf{d}^1, \mathbf{d}^2, \ldots, \mathbf{d}^n\right) \tag{2.1}$$

Each $\mathbf{d}^i$ may be of different length due to differences in sensor characteristics. The first step during preprocessing of inertial data is therefore to interpolate the recorded data to a fixed frequency and provide a joint reference time shared across all degrees of freedom of the sensing setup. This requires synchronisation between the different sensors, often relating to an external source such as video recordings. This synchronisation to a joint temporal basis $T$ can be difficult and represents a research challenge in its own right [Plötz et al., 2012a]. The data-set is transformed into

$$\mathbf{D}'_t = \left(\mathbf{d}^1_t, \mathbf{d}^2_t, \ldots, \mathbf{d}^n_t\right), \text{ for } t \in T \tag{2.2}$$

In cases where the sensor orientation may change over the course of the recordings (e.g. mobile phones), or if the orientation of the sensor may be misleading or otherwise undesirable, it is common to calculate the vector magnitude ($\mathbf{M}_t = ||\mathbf{D}'_t||$). Another method popular in preprocessing is the use of filtering to e.g. limit the analysis to movement within a small band of frequencies.

### 2.4.2 *Segmentation*

Segmentation aims to find semantically continuous regions in the sensor data that are likely to contain the activities of interest, at an adequate granularity. It does however not aim to obtain the identity of each activity, but simply aims to identify episodes that are likely to contain an activity or movement of interest. A segment $\mathbf{s}_i = (T_1, T_2)$ is characterised by the start and end times of such a continuous region [Bulling et al., 2014]. A segmentation $\mathbf{S}$ is simply the collection of these individual segments:

$$\mathbf{S} = (\mathbf{s}_1 \dots \mathbf{s}_k) \tag{2.3}$$

Discovering suitable segments and their boundaries is a significant research challenge which is sometimes referred to as *activity spotting* [Ward et al., 2005, Junker et al., 2008]. It is simply difficult to identify the boundaries between different physical activities and possibly an arbitrary background, which is apparent even in human annotators as discussed in chapter 6. In *energy-based segmentation* explicit segmentation is obtained by thresholding simple metrics on the sensor signal, where chapter 3 illustrates such an approach. It is driven by the assumption that the activities of interest show a uniquely different distribution of e.g. energy to background activities, which imposes additional requirements towards the evaluation of such a system, as the robustness of this assumption has to be demonstrated. Other application scenarios allow the use of an additional modality to obtain an initial segmentation, such as audio recording in a workshop working with wood [Lukowicz et al., 2004], that may be improved upon iteratively.

The most common approach to segmentation in HAR avoids selecting explicit segment boundaries. In *sliding-window segmentation* a window of fixed duration is moved across the stream of sensor data, extracting segments independent of their content with a certain degree of overlap between subsequent segments:

$$\mathbf{S} = (\mathbf{s}_0,\ \mathbf{s}_{\Delta t},\ \mathbf{s}_{2\Delta t} \dots) \tag{2.4}$$

$$\mathbf{s}_t = \left(\mathbf{D}'_{\mathbf{t}} \ \dots \ \mathbf{D}'_{\mathbf{t+w}}\right),\ \Delta t < w \tag{2.5}$$

This approach is characterised by the (fixed) segment duration $w$ and the overlap between subsequent windows $1 - (\Delta t / w)$. This systematic procedure assumes that the elementary units to be analysed fit into a frame, or that at least the data contained in

a frame allows such elementary units to be differentiated reliably. The choice of segment duration $w$ is not at all straight-forward and can have significant effect on the performance of the HAR pipeline [Huynh and Schiele, 2005]. Typical choices of $w$ are around one second for ambulatory movement (e.g. [Stiefmeier et al., 2008, Plötz et al., 2011b]). If prolonged periodic activities such as walking or running are of interest it is beneficial to rely on longer segments, i.e. 5 seconds [Reiss and Stricker, 2012] or up to 32 seconds [Stikic et al., 2008a]. The best segment duration $w$ is commonly obtained through cross-validation experiments.

## 2.5 Feature Extraction

Even small frames of inertial sensor data can contain a large number of samples due to the high temporal resolution of modern MEMS sensing technologies. In addition, the data can be ambiguous as a subset of degrees of freedom are captured by the sensor (e.g. just rotation in gyroscopes). In sliding window approaches, the window placement affects the appearance of the extracted frame, where the same activity of e.g. opening a drawer, is captured by a number of subsequent frames which leads to characteristic parts of the signal to occur at different relative positions in the frame. It is therefore impractical to utilise the raw recordings directly in a classification engine, as a large pairwise distance is not necessarily a good indicator for different frame identity, but could stem from the ambiguities introduced at the sensing and pre-processing level. Instead, a process has to be devised that abstracts from the raw sensor recordings, which *i)* preserves characteristics crucial to the differentiation of activities of interest, and *ii)* addresses the inherent ambiguity of the recorded signal. This process is called feature extraction and a very large number of different methodologies have been devised in the literature (see e.g. [Huynh and Schiele, 2005, Figo et al., 2010] for reviews). Formally, feature extraction projects the (segmented) sensor data into some sort of feature space.

$$\mathbf{X}_i = F\left(\mathbf{D}', \mathbf{s}_i\right) \tag{2.6}$$

The dimensionality of the feature space $\mathbf{X}$ is usually of (much) lower dimension than the samples contained within a segment. A number of different taxonomies have been described in the literature to characterise feature extraction approaches. Figo et al. differentiate three classes of features extracted from inertial sensor data [Figo et al., 2010]: *i)* Time domain features, which include mathematical and statistical attributes such as

mean, variance and correlations; *ii)* Frequency domain features that correspond to signal deconstruction techniques such as Fourier coefficients and wavelet transforms; and *iii)* Discrete domain features, where the (real-valued) signal is transformed into a symbolic representation, allowing the application of methods from other fields of pattern recognition such as bioinformatics. While this taxonomy probably captures the majority of feature extraction approaches used in pervasive computing, it fails to include more recent approaches that rely on more sophisticated machine learning methodologies. Bulling et al. describe 4 different families of feature extraction approaches [Bulling et al., 2014]: *i)* Signal based features, which correspond to the time domain features from Figo et al.; *ii)* Body model features that are calculated from a skeleton or otherwise holistic representation of the (human) body; *iii)* Event-based features which correspond to the occurrence of specific, well-defined events such as eye-blinks; and finally *iv)* Multi-level features that rely on e.g. clustering methods to gain an impression of the underlying patterns in a frame of sensor signal.

In order to provide a background for the work presented in this thesis, two categories of features will be described in detail: *i)* the use of hand-picked statistical attributes calculated for each frame, corresponding to the time-domain features described by Figo et al.; and *ii)* (statistical) dimensionality reduction or feature deconstruction techniques, broadly fitting into the multi-level feature category by Bulling et al.

### 2.5.1 *Statistical features*

The most common approach to obtain a feature representation for frames of inertial data is to construct a set of (statistical) measures, selected from a large, established set of mathematical functions, usually referred to as *time-domain features* [Figo et al., 2010]. Such feature sets typically include basic statistical measures which are calculated for each degree of freedom of interest. Many systems rely on extracting the mean of each sensing axis (sometimes exclusively), which can lead to good performance if multiple sensors are placed along the subjects body [Maurer et al., 2006]. However, if more complicated activities are of interest, a single measure per axis may not retain sufficient characteristic differences to allow reliable recognition. In such cases, more measures are added to the set of features, usually in a manual process driven by experience and intuition of the practitioner. The function $F$ that represents the feature extraction in

equation 2.6 is composed of $d$ individual, independently evaluated functions:

$$F\left(\mathbf{D}', \mathbf{s}_i\right) = \left[\, f_j\left(\mathbf{D}', \mathbf{s}_i\right) \,\right], \text{ for } j = 1 \ldots d \tag{2.7}$$

Common measures used in activity recognition systems include

- **Mean, median, variance, standard deviation, energy, entropy**
  These signal characteristics aim to summarise the distribution of values across each degree of freedom and are used widely in ubiquitous computing [Bao and Intille, 2004, Kwapisz et al., 2011]. They are informative for virtually all sensing modalities [Lester et al., 2006], and their initial application was largely influenced by their success in other areas that handle time-series. Such statistical measures are independent of the temporal structure of inertial time-series and usually driven by assumptions about the underlying data distribution, which may be violated in the case of e.g. raw accelerometer data. Such issues are further elaborated in chapter 5.

- **Correlation, cross-correlation, autocorrelation**
  Measures such as correlation take into account the relationship between multiple degrees of freedom of the recorded sensor data. Autocorrelation (cross-correlation of a signal with itself) can give insights into repetitive motion within an analysis frame. These measures are added to the feature set if the temporal structure of the sensor signal has to be retained to an extent. Example applications include repetitive activities such as walking or brushing teeth, where correlation contributes significantly to the classification accuracy [Ravi et al., 2005]; weight-lifting exercises [Chang et al., 2007]; stereotypical movements in Autism [Albinali et al., 2009]; or to automatically synchronise wearable sensors with video recordings [Plötz et al., 2012a].

HAR systems that rely on a hand-picked set of time domain features often show particularly good performance in various applications [Lara and Labrador, 2013]. However, there are significant shortcomings to this approach. Selecting the features is a very time-consuming process as there are no established procedures that a practitioner can follow to obtain a good feature extraction process. Instead, selecting individual features is driven by experience and intuition to maximise performance in cross validation settings (and sometimes performed automatically [Pirttikangas et al., 2006, Choudhury et al., 2008]). As will be discussed below, selecting a suitable performance metric is

furthermore not trivial for typical datasets captured in pervasive computing. Different performance metrics can lead to very different feature representations.

### 2.5.2 *Dimensionality reduction techniques*

An alternative approach to the extraction of explicit features from frames of inertial data is to address the problem from a mathematical viewpoint. If the segments of fixed length are extracted from the preprocessed data $\mathbf{D}'$ using e.g. a sliding window procedure (see section 2.4.2), feature extraction corresponds to a transformation from an $m$ dimensional input space to a $d$ dimensional feature space. As all the segments are of the same length, the sensor data within each can be encoded into a single $m$-dimensional row vector by e.g. concatenating each degree of freedom (sensor axis), forming a data-matrix $\hat{\mathbf{D}} \in \mathbb{R}^{n \times m}$:

$$F(\mathbf{D}', \mathbf{S}) \; : \; \hat{\mathbf{D}} \in \mathbb{R}^{n \times m} \; \rightarrow \; \mathbf{X} \in \mathbb{R}^{n \times d} \tag{2.8}$$

A large number of approaches have been developed to estimate this transformation $F(\mathbf{D}', \mathbf{S})$ directly, driven by heuristics that reflect some understanding of what characteristics are crucial. These approaches do not rely on experience or intuition of the practitioner but instead on the suitable choice of heuristic and other parameters. As the dimensionality of the feature space is usually lower than in the input space ($d \ll m$) they are referred to as *dimensionality reduction techniques*.

The most popular dimensionality reduction technique relies on the assumption that variance is the most characteristic difference between input samples. Principal Component Analysis (PCA) aims to find a subspace of the original input space, where the variance along each perpendicular component is maximised. Effectively it corresponds to a linear transformation which projects input data into a new coordinate system, in which the largest variance is found along the first dimension, the second largest along the second dimension, and so on [Jolliffe, 2002]. The dimensionality of the transformed space can be reduced simply by retaining just the first few dimensions that capture the majority of the variance found in the data. Mathematically PCA is defined as a set of *principal components*, unit-length weight vectors $\mathbf{w} = (\mathbf{w}_1 \ldots \mathbf{w}_m) \in \mathbb{R}^{m \times m}$, which are the basis for the principal subspace $\mathbf{X}$:

$$\mathbf{X} \; = \; \hat{\mathbf{D}} \cdot \mathbf{w} \tag{2.9}$$

A variety of different formulations lead to different optimisation problems in the discovery of the principal components [Jolliffe, 2002]. The first principal component has to

maximise the variance in the principal subspace and hence satisfy

$$\mathbf{w}_1 = \underset{||\mathbf{w}_1||=1}{\arg\max} \; ||\hat{\mathbf{D}} \cdot \mathbf{w}||^2 \tag{2.10}$$

$$= \underset{||\mathbf{w}_1||=1}{\arg\max} \; \mathbf{w}^{\mathrm{T}} \hat{\mathbf{D}}^{\mathrm{T}} \hat{\mathbf{D}} \mathbf{w} \tag{2.11}$$

where $(.)^{\mathrm{T}}$ denotes matrix transpose. The square matrix $\hat{\mathbf{D}}^{\mathrm{T}}\hat{\mathbf{D}}$ is proportional to the empirical covariance matrix of $\hat{\mathbf{D}}$. The principal components correspond to the *eigenvectors* of that covariance matrix, where their *eigenvalues* reflect the variance along that component. PCA is therefore equivalent to an eigenvalue-decomposition problem and a variety of efficient algorithms have been devised in the literature [Bishop et al., 2006].

PCA is a very popular tool for dimensionality reduction as well and visualisation, and has been used extensively in ubiquitous computing [Mannini and Sabatini, 2010], for e.g. feature extraction [Mantyjarvi et al., 2001, Plötz et al., 2011a] (see also chapter 4), or gait detection [Sprager and Zazula, 2009]. In many cases PCA is applied to reduce the dimensionality of another feature representation, for example time-domain features [Long et al., 2009] or time-delay embeddings [Frank et al., 2010]. However, the application of PCA is not free of practical issues. The reliance on eigenvectors illustrates a very strong assumption of PCA, that the variance in the input space occurs along *linear* components. The reliance on linear transformations can be problematic in practice, as most naturalistic input data is probably of non-linear nature. Despite these shortcomings, PCA still shows promising performance HAR and chapter 4 illustrates how PCA can outperform other feature extraction approaches.

There have been variants of PCA proposed in the literature that alleviate this reliance on linear components by introducing a *kernel*. Examples include kernel-PCA [Mika et al., 1998] and locally-linear embedding (LLE) [Roweis and Saul, 2000], which aim to preserve some local characteristics of the input data defined by the kernel. These approaches are however not very suitable for realistic data-sets, as they require significant computational effort, effectively limiting their practicality in the majority of application scenarios.

Alternatively, feature learning refers to a family of approaches from machine learning that aim to automatically infer a representation of the input data. Feature learning does not rely on an explicit non-linearity introduced as a kernel in an otherwise linear approach. Instead, feature learning assumes that the input data is generated by interactions of (a large number of) factors. The problem of finding a feature transformation

then corresponds to finding the most suitable factors that could generate the data observed in the input set. This is fundamentally different to analytical approaches such as PCA, as instead of trying to preserve some aspects of the input set, feature learning aims to learn a generative (probabilistic) model of the input data.The most popular family of methods are the Restricted Boltzmann Machines that are used to build large-scale neural networks called Autoencoders [Hinton and Salakhutdinov, 2006]. Additional related work will be introduced in section 4, where different feature learning approaches are evaluated on a number of datasets from pervasive computing.

### 2.5.3 *Feature extraction for data from naturalistic settings*

Both statistical feature representations as well as feature representation estimated automatically by e.g. dimensionality reduction depend on the quality (and quantity) of data used during their design and training. In the case of statistical features the selected subset of features is investigated with respect to their performance, iteratively improving upon the representation during system development to ensure maximum performance [Choudhury et al., 2008]. In the case of dimensionality reduction the estimated features are estimated on a set of training samples (see chapter 4).

For both approaches it is best practice to evaluate the resulting recognition system in cross validation experiments in order to investigate the generalisation performance Bulling et al. [2014]. Inherently this process encounters the risk of over-fitting feature representations and classification approaches (see below) to specific training-sets. This is a problematic strategy if the data-sets that are used for system development are based on artificial or scripted data collection procedures or recorded from people outside the target population. Cross validation in these settings effectively prevents overfitting to a specific sub-part of that very data-set but gives by no means an indication of the performance of the system under more naturalistic environments.

As discussed in section 2.2.4, naturalistic environments effectively prevent the capturing of reliable, high-resolution labelling. Despite this it may be straight-forward to collect large amounts of data in naturalistic settings if sensing systems are sufficiently usable. Statistical features or other feature extraction methods that rely on recognition performance to guide their design are unable to exploit this rich source of unreliably or even unlabelled data. The use of such features must therefore still be guided by e.g. a preliminary, well-controlled study setting in which labels for activities can be obtained and therefore does not alleviate the concerns regarding over-fitting of such an approach.

Dimensionality reduction or feature learning techniques on the other hand do not suffer from these limitations. In their training stage no labels are required at all and their application to very large amounts of data is simply a problem of devising efficient algorithms that minimise the required computational effort. Their use therefore effectively alleviates concerns regarding the over-fitting of such a feature representation if sufficient amounts of naturalistic data are recorded.

## 2.6 Classification and Inference

The next step in the activity recognition pipeline (see Figure 2.6) is that of classification, aiming to differentiate different activities from a known corpus using some form of inference method. This is a task typical for the field of Pattern Recognition and Machine Learning, where a very large number of inference methods have been developed and applied [Bishop et al., 2006]. An exhaustive review of the field of pattern recognition is beyond the scope of this thesis. This chapter will instead give an overview of the methods that are typically applied in the field of HAR.

The most common approaches to classification rely on a *supervised learning* methodology. Such approaches aim to model the posterior probability $p(y|\mathbf{x}_i)$ of class identity $y$ for a feature vector $x_i$, based on a large training dataset that contains $n$ pairs of feature vector and label $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1\ldots n}$. The class with the highest posterior probability for $x_i$ then corresponds to the predicted identity [Bishop et al., 2006]:

$$\hat{y}_i = \arg\max_y p(y|\mathbf{x}_i) \tag{2.12}$$

Where some formulations also include $\Theta$ to explicitly parametrize the model (e.g. coefficients for linear model) [Bulling et al., 2014]. Not all classification approaches used in HAR calculate probabilities, such as k-nearest-neighbour (kNN). However, for each of such methods probabilistic formulations can be found, which can be helpful to prevent e.g. over fitting [Bishop et al., 2006].

### 2.6.1 *Discriminative approaches*

Classification approaches that model the posterior probability $p(y_i|\mathbf{x}_i)$ directly are referred to as *discriminative* approaches. Intuitively such methods aim to find a *decision*

*boundary* which separates the classes in the feature space, even though this is sometimes done implicitly [Shawe-Taylor and Cristianini, 2004]. A new sample point is classified simply by looking at which side of the decision boundary it falls, where the distance to that decision boundary is related to the confidence of this classification. Discriminative approaches can pose significant computational requirements during training, as the discovery of suitable decision boundaries can be a challenging optimisation problem [Bishop et al., 2006].

Constraints on the shape of the decision boundary is a common tool to make finding an optimal decision boundary tractable. Of particular importance here are linear models such as logistic regression, where the decision boundary is expressed as a linear combination of the feature variables and coefficients. This leads to a convex optimisation problem, where a uniquely optimal solution can be obtained quickly in a gradient-based learning process [Bishop et al., 2006].

Support Vector Machines (SVMs) extend linear models by introducing a *kernel*, which allows non-linear decision boundaries while retaining the computational advantages of a convex optimisation [Suykens and Vandewalle, 1999, Shawe-Taylor and Cristianini, 2004]. In simple terms, input samples are projected from a low-dimensional feature space into a (much) higher dimensional kernel space, in which a linearly separating hyperplane is estimated. Projecting this boundary back into the low dimensional feature space yields a non-linear decision boundary, whose properties are affected by the choice of kernel function. Figure 2.7 illustrates the decision boundaries estimated by different kernels on a toy data-set. SVMs have been applied in a wide range of settings in AR [Frank et al., 2010, Ravi et al., 2005, Bulling et al., 2014, He and Jin, 2008, 2009, Sun et al., 2010, Wang et al., 2005]. [Plötz et al., 2012b] and chapter 3 show how SVMs can be applied to differentiate problem behaviour in individuals with developmental disabilities.

Partitioning of the feature space using a sequence of simple classifiers is an alternative to explicitly constrained decision boundaries. Examples for such methods include boosting, where a partitioning is found in a sequential optimisation process that weighs data according to the prediction performance in previous steps [Lester et al., 2006, Blanke and Schiele, 2009, Zinnen et al., 2009]. Particularly interesting for embedded application is the use of decision trees, a tree-like decision structure that is derived in a very efficient learning process [Quinlan, 1993]. Decision trees require comparisons in the order of $\log(d)$ with $d$ being the dimensionality of the feature space at inference time,

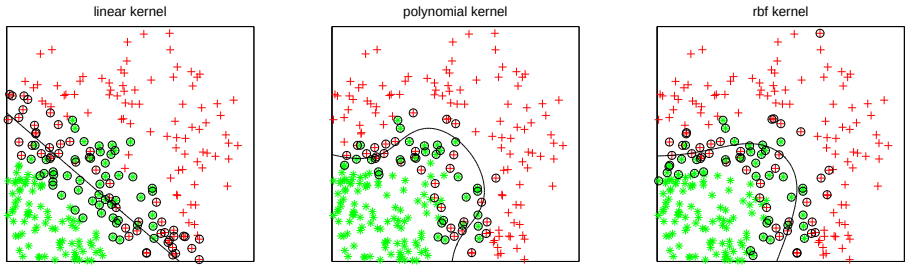



FIGURE 2.7: Examples for different kernels in SVMs on a toy data-set that is not linearly separable. Circles indicate support vectors that implicitly define the decision boundary (in black). The choice of kernel and it's (hyper-) parameters is crucial for the performance of SVMs and may require a computationally intensive grid-search in parameter space.

which makes them particularly suitable for embedded systems. Chapter 5 shows the application of decision trees to a variety of data-sets. Decision trees and their extensions have been applied to a wide range of problem settings in HAR [Bao and Intille, 2004, Albinali et al., 2009, Győrbíró et al., 2009].

In an Artificial Neural Network (ANN), a large network of interconnected simple units (neurons) parametrises the function from the input features to the output classes. Each unit performs a simple calculation based on it's input and applies a non-linear activation function to obtain an output, which is then sent to the other units it is connected to [Bishop et al., 1995]. In contrast to other discriminative approaches, ANNs are universal function approximators that can, in theory, model any decision boundary given that a sufficient number of units are used in one *hidden* layer in the network [Hornik et al., 1989]. This comes at the cost of a more challenging optimisation problem, where training may converge prematurely in a local optima which can make the application of (large) ANNs very challenging.

One advantage of ANNs is that it is straight-forward to model multi-class classification problems by adding a so called *softmax group* as the final output layer. A softmax group contains one unit for each output class, and the output activation of each unit is normalised by the sum over all activations in the output layer. Effectively the output of the ANN then corresponds to a distribution over class labels and simple measures for discrepancy between the distribution and the actual label can be used to train the network [Bishop et al., 2006]. In addition to the predicted class (the unit with the highest activation), the output of the network reflects a degree of confidence in that prediction. For cases where the ANN is "sure", the output for the predicted class will be close to 1, with all other units being close to 0. For cases where the ANN is "unsure" there may

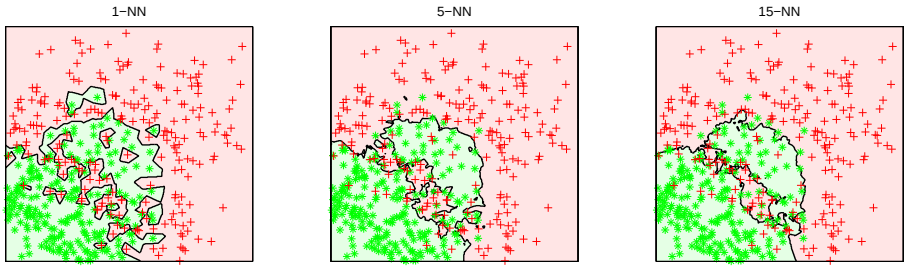



FIGURE 2.8: Plots illustrating instance based learning with varying size of neighbourhood. With increasing number of neighbours the space becomes less fragmented, which may prevent over-fitting.

be significant weight on other units in the output. This allows the ANN to express a *confidence* in each prediction, which can be beneficial for many multi-class classification problems with high class confusion (see chapter 8).

Recent advances in machine learning, so called *deep learning*, aim to initialise large multi-layer ANNs using generative models such as RBMs or autoencoders [Hinton et al., 2006, Hinton and Salakhutdinov, 2006]. This has made these *deep ANNs* very popular for settings where large amounts of unlabelled data are easily accessible, such as speech recognition [Hinton et al., 2012a, Deng et al., 2013]. Even though ANNs have been applied in HAR and movement analysis in a variety of settings [Best and Begg, 2006, Cole et al., 2010, Győrbíró et al., 2009, Keijsers et al., 2000, Pirttikangas et al., 2006], application of deep ANNs are still relatively rare [Plötz et al., 2011a]. Chapter 4 investigates the use of such deep ANNs for feature extraction for a variety of data-sets. Furthermore chapter 7 and chapter 8 utilise deep ANNs for practical applications.

### 2.6.2 *Instance-based learning*

In instance-based learning, any computation such as modelling a decision boundary, is left to the classification stage when a novel sample point is considered. Such *lazy learning* relies on keeping a large amount of samples (or points derived from samples) within memory [Wilson and Martinez, 2000]. At inference time, an hypothesis about the class identity of a new sample point is constructed using pairwise similarity between data-points and a notion for the local structure, or neighbourhood surrounding the new sample point.

Given a new sample, the probability of being classified as class $y_i$ is simple the fraction of "similar" samples belonging to that class $y_i$:

$$p(y_i|\mathbf{x}_j) = \frac{1}{|n(\mathbf{x}_j)|} \sum_{k \in n(\mathbf{x}_j)} c_{ik}, \tag{2.13}$$

$$c_{ik} = \begin{cases} 1 & \text{if } y_i = y_k \\ 0 & \text{otherwise} \end{cases} \tag{2.14}$$

Instead of parametrizing a decision boundary, such *instance-based learning* approaches encode it implicitly in their notion of pairwise similarity and the size of the neighbourhood, which is why they are referred to as *non-parametric* approaches. In practice it is common to utilise euclidean distance as a similarity metric and the closest point from the training set in the neighbourhood of each point $n()$ (1-nearest neighbour) to perform classification. Increasing the number of points in $n()$ increases the smoothness of the decision boundary which can prevent over-fitting (see Figure 2.8).

A large variety of instance-based learning approaches have been proposed, which mostly differ in their notion of similarity and neighbourhood. For example, nearest centroid approaches estimate class centroids for each class using simple heuristics. Different methods to reduce the number of retained points from the training set (prototypes) have been introduced that follow scoring mechanisms [Wilson and Martinez, 2000, Garcia et al., 2012] to construct a set of prototypes. Particularly k-nearest neighbour classification with euclidean distance as similarity metric is popular in AR [Bulling et al., 2014]. Dynamic time-warping has also shown suitability for accelerometer data in e.g. a kitchen scenario [Pham and Olivier, 2009] along with approaches that explicitly match parts of a signal to a database of labelled movement data [Van Laerhoven and Berlin, 2009].

### 2.6.3 *Generative Modelling*

In contrast to directly modelling the posterior, *generative* approaches model the joint probability distribution of feature vectors and labels $p(\mathbf{x}, y)$ or the conditional generative distribution $p(\mathbf{x}|y)$ along with prior $p(y)$. Intuitively such methods do not try to find some (optimal) decision boundary, but instead aim to model how the data from each class was generated. The joint probability can be transformed into the posterior

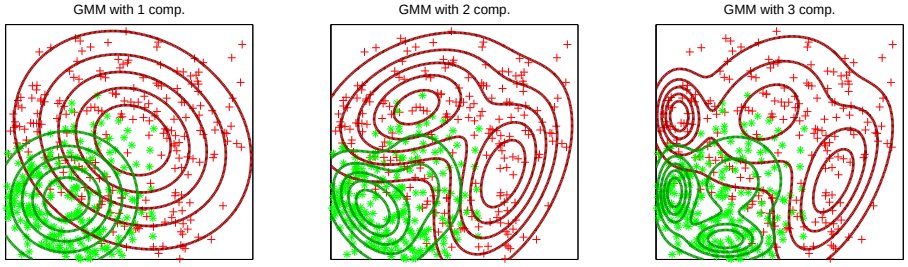



FIGURE 2.9: Gaussian Mixture Models (GMMs) fitted to toy data-set with varying number of components (per class). Contours indicate probability density of estimated model.

over class labels given feature vectors, $p(y|\mathbf{x})$, using Bayes theorem:

$$p(\mathbf{x}_i, y) = p(\mathbf{x}_i|y)p(y) = p(y|\mathbf{x}_i)p(\mathbf{x}) \tag{2.15}$$

$$p(y|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|y)p(y)}{p(\mathbf{x}_i)} \tag{2.16}$$

$$p(y|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|y)p(y)}{\sum_y p(\mathbf{x}_i|y)p(y)} \tag{2.17}$$

$$\hat{y}_i = \arg\max_y \frac{p(\mathbf{x}_i|y)p(y)}{\sum_y p(\mathbf{x}_i|y)p(y)} \tag{2.18}$$

The prior $p(y)$ can be used to incorporate knowledge such as the expected frequency of occurrence of specific activities, which may not be reflected adequately in the training data. Estimating the generative model $p(\mathbf{x}_i|y_i)$ is robust towards class imbalance, as capabilities can be controlled by the family of models that are utilised to model each class. Furthermore generative models can be very efficient when fitting to a training set, as no complex optimisation process is required to e.g. find a decision boundary.

A very simple example for a generative model is NaiveBayes, which assumes complete independence of each feature variable. Due to this assumed independence, the probability of $x_i \in \mathbb{R}^d$ is then simply

$$p(\mathbf{x}_i|y) = \prod_{k=1}^{d} p(x_{ik}|y), \tag{2.19}$$

where $x_{ik}$ corresponds to the $k$-th feature of $\mathbf{x}_i$. For categorial variables, estimating $p(x_{ik}|y)$ is a simple matter of counting the co-occurrence of feature value and class label. Due to it's strong assumptions, NaiveBayes is often outperformed by discriminative methods. It is however suitable to gain an understanding of the quality of the feature

space, which makes it a common tool for data exploration. NaiveBayes has been employed extensively in HAR [Ravi et al., 2005, Bulling et al., 2014], where it is often used as a baseline algorithm (see also chapter 4).

Instead of modelling each input dimension independently, mixture models rely on a number of weighted component distributions to model the probability of a training sample $p(\mathbf{x})$, or of a training sample given its class membership $p(\mathbf{x}|y)$ [Bishop et al., 2006]. The component distributions typically all belong to the same parametric family, where Gaussian approaches are the most common due to their favourable mathematical properties. The overall model is parametrised by the mean vectors, covariance matrices and mixture weights, collectively represented as $\Theta = \{w_i, \mathbf{\Sigma}_i, \boldsymbol{\mu}_i\}_{i=1..k}$. If a Gaussian Mixture Model (GMM) is fitted to model $p(x|y) = p(x|\Theta_y)$, classification corresponds to selecting that mixture that provides a higher likelihood for a specific test sample $x$:

$$p(\mathbf{x}|\Theta) = \sum_{i=1}^{k} w_i \, \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i), \tag{2.20}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \, \mathbf{\Sigma}_i^{-1} \, (\mathbf{x}-\boldsymbol{\mu}_i)\right), \tag{2.21}$$

$$\sum_{i=1}^{k} w_i = 1, \tag{2.22}$$

$$\hat{y}_i = \arg\max_y p(\mathbf{x}_i|\Theta_y)p(\Theta_y) \tag{2.23}$$

There are different approaches available to fit a GMM to a data-set, where the most popular approach is maximum likelihood (ML) estimation of the parameters [Bilmes et al., 1998, Figueiredo and Jain, 2002]. As the name suggests, the aim of ML is to maximise the likelihood of data under the model. It starts with an initial model and iteratively improves the model in an Expectation-Maximisation (EM) algorithm. Figure 2.9 illustrates GMMs with different numbers of components fitted to example data. Some work in HAR has utilised GMMs to classify e.g. kitchen activities [Plötz et al., 2011b] and generic activity recognition [Allen et al., 2006]. GMMs are utilised extensively as building block of more complex models. One particular example of such approaches applied to time-series data are *Hidden Markov Models* (HMMs).

An HMM consists of a set of internal states, each of which are capable of generating data according to an associated *emission* model. In the most popular family of HMMs, such emissions are modelled as GMMs (*continuous* HMM), whose components may be

shared between the emissions of different states (*semi-continuous* HMM) [Riedhammer et al., 2012]. Often the different component distributions of emission models in HMMs are constrained to have a diagonal covariance. This significantly reduces the number of parameters that have to be estimated when fitting the emission model to training data at the cost of reduced representational power. To counter this effect the number of components in the model is increased and may be in the thousands, which can be beneficial if the data distribution is highly non-Gaussian [Plötz and Fink, 2009]. HMMs assume that a sequence of input vectors is generated by the emissions from a *hidden* sequence of states. The number of states and their arrangement (e.g. transition probabilities) are designed using prior knowledge and may represent some higher level abstraction, e.g. gestures in HAR, letters in handwriting recognition or phonemes in speech recognition (where these models originate [Huang et al., 1990]). Crucial to their application in HAR is that HMMs can be used to infer an hypothesis about activities as well as their boundaries, where a likely segmentation of the input data is effectively part of the produced hypothesis (as the most likely state-sequence to generate the input sequence). A detailed explanation of HMMs and the different algorithms used for inference and testing goes beyond the scope of this thesis and the reader is referred to [Bishop et al., 2006] for a comprehensive review of Markov models for pattern recognition.

HMMs have been employed widely in the field of HAR and movement analysis. Some examples include sign-language recognition [Starner et al., 1998b], tracking of weight-lifting exercises [Chang et al., 2007], assessment of self-stimulatory behaviour in autism [Westeyn et al., 2005], gesture spotting and discovery of characteristic actions from sensor data [Minnen et al., 2007, Junker et al., 2008], and selection of suitable sensor placement [King et al., 2007]. Their capabilities make them particularly suitable for recognition of behaviour at a higher abstraction level, for example, to gain an understanding of *behaviour* compared to plain physical activities. Recently, further extensions to HMMs, namely Conditional Random Fields (CRFs) have become popular in HAR [Atallah and Yang, 2009b, Lee et al., 2011].

### 2.6.4 *Performance Metrics*

Evaluating the performance of the activity recognition pipeline is crucial to its design, as the different components are selected from a large set of possible combinations guided by performance. The selection of the components of the activity recognition pipeline is therefore tightly linked to the choice of evaluation metric. On one hand evaluation

of HAR systems faces the same challenges as similar settings in other tasks of pattern recognition. On the other hand, the sequential nature of activities and sensor recordings introduce additional issues, which become apparent in practical deployments of HAR systems [Minnen et al., 2006, Ward et al., 2011]. Most of the performance metrics listed in table 2.1 have been adopted in HAR. The most common performance metric in HAR is the overall accuracy, i.e. the fraction of correctly classified instances. Additionally, other methods are popular for performance evaluation and illustration:

**Confusion matrices** are used to get an overall impression of the confusion (i.e. misclassification) in a HAR system. It is constructed as follows: In a square matrix, row $i$ corresponds to all instances labelled to belong to class $c_i$, and each column $j$ contains each instance classified to belong to class $c_j$. Diagonal elements represent correct predictions, while off-diagonal entries indicate confusion between classes. Since data-sets in HAR are often dominated by the (less interesting) background activities (or NULL class) it is common to normalise the absolute numbers in each cell to represent *confusion probabilities*. Colour-coded illustrations of these normalised confusion matrices are common tools to visualise classification performance.

**mean-, or average f1-scores** have become increasingly popular, due to their robustness towards imbalanced datasets. This imbalance between the classes, for example the dominance of background-activity in a data-set, can lead to misleading high performance figures when utilising accuracy, while particularly the mean f1-score remains informative in such cases. Based on the confusion matrix, these measures are calculated by constructing a virtual two-class classification problem for each class following a *one-vs-all* procedure.

**ROC curves** or *receiver-operator-characteristics* are a common tool to illustrate the impact of parameters on classification approaches. Depending on the application, different performance characteristics are deemed desirable. For example, a system aimed to support manual video annotation may be aimed at finding occurrences of some target activities such as gesturing. In such a setting, it is desirable to devise a system with very high sensitivity, even at the cost of additional false predictions (i.e. low specificity). In other settings the contrary may be true, that false positives need to be avoided at all costs. When reporting results for such systems it is often desirable to estimate a performance metric that illustrates the overall difficulty of the problem, independent of the

trade-off in performance characteristics. ROC curves provide such a means to visualise or quantify performance, by illustrating the relationship between the true positive rate (sensitivity) and false positive rate (1-specificity) in an intuitive graph.

### 2.6.5 *Classification in naturalistic settings*

In principle, classification approaches suffer from similar limitations as feature extraction methods (see chapter 2.5.3) in that they are prone to overfitting to artificial or otherwise non-representative study settings. However this risk of over-fitting is well known in the field of machine learning and powerful regularisation strategies have been employed to prevent it. An example is to randomly turn off neurons in a hidden layer during neural network training (*dropout*) to prevent co-adaptation [Hinton et al., 2012b], or to introduce a prior over hyper-parameters in kernel-based methods for bayesian model selection [Cawley and Talbot, 2007].

The suitability of classification approaches for naturalistic settings rather depends on how robust their performance is towards unreliable or imprecise labelling, which can be the result of practical constraints towards sensing systems in naturalistic settings (see chapter 2.2.4). Particularly the boundaries of activities may be imprecise, leading to frames of accelerometer data whose label does not reflect the activities it contains. In instance-based learning such wrongly labelled instances in a training-set lead to problems during inference, as new samples that fall in the neighbourhood of such points are more likely to be classified incorrectly. In practice the effect of these imprecise boundaries can be addressed by increasing the size of the neighbourhood, which may however be detrimental to the overall performance of the system. Both discriminative and generative modelling avoid this problem to an extent in that the impact of an individual (wrongly labelled) sample on the model is minimal. The problem of *label-noise* is however also an issue for methods such as SVMs [Biggio et al., 2011] or boosting, which is susceptible to overfit to outliers [Krause and Singer, 2004, Karmaker and Kwek, 2006].

While all classification approaches suffer from noise in the labelling, it seems that the effect is minimal for methods that rely on *soft* labels [Thiel, 2008]. Soft labels here refers to the representation of a class membership not as categorical variable but instead as a real-valued vector where each entry reflects the probability that the sample belongs to a class, as is common in neural networks. A further advantage of neural networks is that very large amounts of unlabelled data can be used to train multi-layer networks with millions of parameters in a greedy, layer-wise procedure referred to as *deep learning*

[Hinton et al., 2006, Deng et al., 2013]. It therefore appears that, in particular, neural networks are suitable for data from naturalistic settings, as they are robust towards label-noise while being able to exploit the large amounts of unlabelled data accessible in naturalistic surroundings.

## 2.7 Summary

This chapter investigated different sensing approaches and their suitability for naturalistic surroundings, introduced different application settings where e.g. activities are detected based on such sensor recordings, and summarised the state-of-the-art approach to analyse sensor data from the dominant modality in ubiquitous computing, inertial body-worn sensors.

The sensing approach that is most suitable for naturalistic surroundings are body-worn sensors, as they do not require extensive (costly) infrastructure and have high spatial and temporal resolution at moderate ambiguity. If sensing systems relying on body-worn sensing are designed with the requirements of the target population in mind they can show very high usability which leads to high compliance. However, it is practically difficult to obtain reliable annotation of data collected in e.g. the private home, which has implications for the use of such data in designing automatic recognition systems.

The data collected from body-worn sensors is typically processed in a pipeline approach, whose components are tuned to maximise performance in cross-validation experiments on the collected movement data. Both the design of feature extraction and classification approaches rely on precisely annotated training data, where the practical lack of reliable annotation from naturalistic settings has a significant impact. To an extent the generalisation ability of automatic recognition systems can be assessed by augmenting a small, artificial data-set with additional naturalistic data, where chapter 3 gives a practical example surrounding assessment in Autism.

In summary, systems aimed at naturalistic surroundings should not be based on manually selected feature representation but instead should utilise automatic inference methods that allow processing of large amounts of unlabelled movement data. The label noise typical in these applications is best addressed in classification methods that utilise

soft labels, where the categorical class membership is substituted with a vector representing class membership probabilities. One method that combines both of these aspects is deep learning, where large amounts of unlabelled data are utilised to greedily initialise multiple layers of feature detectors, which are fine-tuned using standard gradient-based back-propagation. The research presented in chapter 4 represents the first exploration of the suitability of deep learning and feature learning for typical application scenarios in HAR in ubiquitous computing, and chapter 5 provides insights into a novel representation for accelerometer data – the ECDF representation – which has favourable properties towards the application of feature learning techniques. The approach is evaluated in two subsequent case-studies that represent novel applications of ubiquitous computing for the recognition of canine activities (chapter 6) and automatic skill assessment and detection of rock-climbing activity (chapter 7).

The insights from this chapter, the technical approach and the lessons learned for activity recognition in naturalistic settings inform the development of a novel approach to the assessment of disease state in Parkinson's Disease, based on a large study conducted in naturalistic settings.

| | | |
|---|---|---|
| **sensitivity** | $\frac{\texttt{tp}}{\texttt{tp + fn}}$ | Fraction of positive instances that are predicted as positive. Also known as true positive rate or recall. |
| **specificity** | $\frac{\texttt{tn}}{\texttt{fp + tn}}$ | Fraction of negative instances that are predicted as negative. Also known as true negative rate. |
| **precision** | $\frac{\texttt{tp}}{\texttt{tp + fp}}$ | Fraction of positive instances of those predicted as positive. Also known as positive predictive value. |
| **accuracy** | $\frac{\texttt{tp + tn}}{\texttt{tp + fp + tn + fn}}$ | Fraction of correctly classified instances (two-class) |
| **f1-score** | $2 \cdot \frac{\texttt{prec} \cdot \texttt{sens}}{\texttt{prec + sens}}$ | Geometric mean of precision and recall. |
| **accuracy** | $\frac{\sum_{i=1}^{c} \texttt{tp}_i}{N}$ | Number of correctly classified instances with classes $C = \{c_1..c_c\}$ and $N$ instances in the whole set. |
| **average f1-score** | $\frac{2}{c}\sum_{i=1}^{c}\frac{\texttt{prec}_\texttt{i} \cdot \texttt{sens}_\texttt{i}}{\texttt{prec}_\texttt{i} + \texttt{sens}_\texttt{i}}$ | Average f1-score for multiple classes $C = \{c_1..c_c\}$. |
| **weighted f1-score** | $\frac{2}{N}\sum_{i=1}^{c} n_i \frac{\texttt{prec}_\texttt{i} \cdot \texttt{sens}_\texttt{i}}{\texttt{prec}_\texttt{i} + \texttt{sens}_\texttt{i}}$ | Weighted f1-score for multiple classes $C = \{c_1..c_c\}$, where each class has $n_i$ instances, summing to $N$ instances in the whole set. |

TABLE 2.1: Typical performance metrics for classification problems, based on class confusion. For two class problems one class is denoted *positive* and one *negative*, leading to the the terms *true positive* (`tp`), *false positive* (`fp`), *true negative* (`tn`), and *false negative* (`fn`). For multi-class problems, each metric is calculated in a *one-vs-all* approach, where *positive* refers to instances from the class and *negative* to all instances from the other classes (see text for details).

# Chapter 3. Automated Assessment of Problem Behaviour in Individuals with Developmental Disabilities

This chapter explores automatic means to assess problem behaviour in individuals with autism using body-worn inertial sensing. It represents a typical application of the activity recognition pipeline discussed in chapter 2.4 and is useful to illustrate the challenges of naturalistic settings with an explicit example. Typically for clinical applications, data collection is challenging as access to study participants may be limited. Specifically this study highlights issues with: i) data collection from vulnerable and possibly uncooperative populations; ii) over-fitting the components of an activity recognition pipeline to a very specific study setting; and iii) the evaluation of HAR systems if representative data is limited by augmenting data collection with case studies and existing data-sets captured in similar surroundings.

## 3.1 Introduction

Many individuals with developmental disabilities, including those on the autism spectrum, engage in problem behaviors [Eisenhower et al., 2005, Hartley et al., 2008, Lecavalier et al., 2006]. Behavior problems, such as temper tantrums, destructive behaviors, aggression toward others, and self-injury, are part of the clinical description of autism [Cooper and Michels, 1994, Karjalainen, 1992]. Beyond the potential for harm or injury to the individual or those nearby, negative consequences of these behaviors extend to many aspects of the individual's life. They disrupt family functioning and increase caregiver stress and anxiety [Herring et al., 2006, Lecavalier et al., 2006], interfere with learning and socialization [Horner et al., 2002], and negatively impact long-term prognosis [Howlin et al., 2004]. Thus, many treatment programs have been developed to

reduce the frequency and severity of problem behaviors in children with developmental disabilities [Machalicek et al., 2007, Matson and Lovullo, 2008]. While treatments themselves differ in approach, all require the collection of accurate data on the frequency and severity of problem behaviors to understand why and when they occur and to determine if there is a change in the behavior as a result of treatment.

The two main methods for measuring problem behaviors include standardized, validated parent- or teacher-report checklists [Achenbach and Rescorla, 2000, Aman et al., 1985, Rojahn et al., 2001], and direct observations [Foster and Cone, 1986, Hanley et al., 2003]. The former provide quick and cost-effective means of gathering data and are widely used in research settings. However, they do not capture precise frequencies of occurrence of the behavior. Thus, the standard procedure for measuring problem behavior in clinical settings consists of having an observer track and record the frequency of the behavior based on precise pre-determined definitions. While such observations yield rich data regarding frequency and context of problem behavior, there are drawbacks. Definitions of problem behaviors can be subjective and somewhat arbitrary, requiring extensive training and reliability assessments. In addition, certain behavior types can be especially difficult to recognize based on what they look like, while others are difficult to track accurately and objectively. There is no way to objectively assess the intensity of a behavior by human observation alone, even though this is the very characteristic of behavior that may improve with treatment. Finally, direct observation is time intensive and expensive to conduct, and thus, can only be employed to gather small samples of behavior.

Accurate assessment of problem behavior is both key to successful treatment planning and evaluation and the main drawback of current methods of manual observation and tracking. Therefore, our goal is to explore how technology and computational analysis, i.e., activity recognition using body-worn sensors, can support the gathering of objective, accurate measures of the frequency of problem behaviors. Direct sensing and assessment has the potential for enhancing current clinical practice by providing analysis that is more objective and consistent, and less expensive and time intensive than manual assessments. The complexity of problem behavior and the large variance in its manifestations implies non-trivial challenges to sensor data analysis. The same holds for the design of a safe, robust, and reliable sensing system for a vulnerable population. This paper describes the first system of its kind, which replicates experts' assessments of problem behavior in clinical settings. As such it represents the first milestone towards

our ultimate goal of developing a sensing and analysis system for continuous unsupervised behavior assessment in everyday life situations.

We observed current practice at a treatment clinic where behavior is assessed for the frequency of aggression (directed at others), disruption (directed at the environment), and self-injury (directed at self). Based on these observations we designed and developed a sensing system based on tri-axial accelerometers worn on the individual's limbs. Computational analysis is based on unsupervised segmentation of sensor data streams into behavior episodes that are then classified using an activity recognition system based on a novel, problem-specific feature representation capturing energy characteristics and sensor orientations, and statistical classifiers.

We rigorously tested the developed system in three sets of practical experiments. First, we evaluated its sensitivity by analyzing a large dataset of simulated assessment sessions where experienced staff members at the clinic engaged in typical problem behaviors while wearing the sensing system. The automatic analysis detected severe behavior episodes with a precision of $> 95\%$ (recall: 41.5%) and an average accuracy of approximately 80% for differentiating among aggression, disruption, self-injury, and movements unrelated to problem behavior. Second, we evaluated the system on a standard activity recognition dataset (OPPORTUNITY challenge [Roggen et al., 2010]), which contains data recorded using a comparable sensing system and covering activities of daily living (ADL) that —by definition— do not include problem behavior episodes. Our system achieved a negligible number of false positive predictions. Third, we evaluated our system in a real clinical assessment session with an autistic child who occasionally engages in problem behavior. Our automatic analysis largely replicates the results of expert assessment.

## 3.2 Clinical Assessment of Problem Behavior – Current Practice in Behavior Clinics

The work presented in this paper was conducted in close collaboration with a local behavior treatment clinic. In the following section we describe the clinic's behavior assessment practices, which are representative of typical procedures in such facilities and thus form the foundation for our research.

When individuals with developmental disabilities engage in problem behaviors, caregivers typically seek professional help to address these behaviors. The first step is to

objectively assess the frequency and severity of the problem behavior, its topology (characteristics), and to understand its causes and functions so that an appropriate, targeted treatment plan can be devised. Once treatment commences, there is a need to gather data to determine whether the child is responding to the treatment. Upon treatment completion it is common practice to follow-up with the family to ensure that treatment gains are being maintained and generalize to the child's everyday life. The key variable underlying this entire process consists of expert assessments of frequency and topology of the target behavior, rooted in the gold-standard practice of direct observation [Foster and Cone, 1986].

### 3.2.1 *Functional Behavioral Assessment*

The key variable in matching interventions to individuals and their particular problem behavior is the *function* of that problem behavior [Hanley et al., 2003, Patel et al., 2000, Smith and Iwata, 1997]. Function refers to the antecedent variables —both internal to the individual and external in the environment— that evoke and the consequences that maintain the behavior. Common functions for problem behavior include desire for caregiver attention, access to preferred items, or escape from/avoidance of demands to engage in non-preferred activities.

At the outset of treatment, identifying the function of an individual's problem behavior is often accomplished using so-called *functional behavioral assessment* (FBA [Gresham et al., 2001, Iwata and Worsdell, 2005]). During this procedure, the individual is observed in test conditions in which potential antecedents of problem behavior are introduced (e.g., attention is withheld). Rates of problem behavior that occur during these conditions are compared to control conditions that don't contain variables that might evoke problem behavior (i.e., child is provided with attention). The function of problem behavior is determined by identifying those test conditions in which the rate of problem behavior is elevated relative to the control condition.

An FBA is usually conducted within specialized clinic facilities and with highly trained staff, both of which are necessary to collect the requisite observational data and ensure the safety of all involved. Current practice consists of sessions conducted within treatment rooms equipped with one-way mirrors, microphones, and cameras to allow unobtrusive data collection. One staff member remains in the room with the child in order to administer the various test conditions, while another observes from an adjacent room through a one-way mirror and flags occurrences of target behaviors. The latter

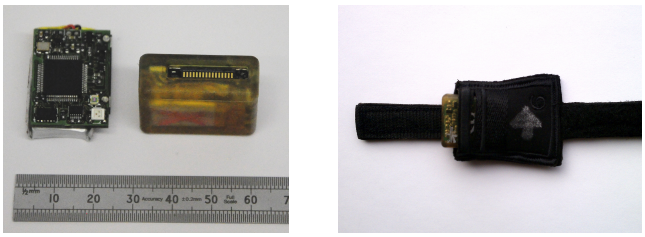
| Behavior | Operational Definition |
|---|---|
| Aggression (AGG) | **Biting:** top and bottom teeth come into contact with any part of a person's body<br>**Grabbing:** squeezes/pinches/grabs person's body part/clothing with one/both hands<br>**Hair Pulling:** grabbing another person's hair with one or both hands resulting in moving the person's head from its original position<br>**Hitting:** hands/forearms contact any part of a person's body from distance of $\geq$ 6"<br>**Object AGG:** throwing object within 2 feet of a person from distance of $\geq$ 6"<br>**Kicking:** foot/leg contacts any part of a person's body from distance of $\geq$ 6"<br>**Pushing:** forcefully moving a person from their original location using one/both hands |
| Self-Injurious Behavior (self-injury) | **Self-Biting:** jaw opens and teeth come into contact with any part of body<br>**Body Slapping/hitting:** hits/slaps any part of his/her body with an open palm or closed fist from a distance of $\geq$ 6"<br>**Face slapping:** slaps face with and open palm from a distance $\geq$ 6"<br>**Head banging:** head forcefully comes into contact with the ground or any other hard surface from a distance of $\geq$ 6"<br>**Head Hitting:** hits head with open/closed fist or with object from distance of $\geq$ 6"<br>**Self Kicking:** foot contacts another part of body from distance of $\geq$ 3 " |
| Disruption | **Body Slamming:** runs into objects from 6 " or greater<br>**Furniture:** tipping furniture 45 degrees from its original position<br>**General:** hands/feet/body come into contact with floor/wall/object from $\geq$ 6 "<br>**Object Disruptions:** pushing or swiping objects from surfaces or throwing an object not within 2 feet of a person<br>**Property Destruction:** rips or tears an object |

TABLE 3.1: Operational definitions used for assessment of problem behavior.

are operationally defined to allow for consistent scoring (Table 3.1). A second observer annotates $\geq$ 20% of sessions for inter-observer agreement calculation.

### 3.2.2 *Tracking Treatment Progress and Outcome*

Once functions of targeted problem behaviors are identified and treatment begins, there is a need for ongoing data collection to monitor the child's progress and, as needed, to make the necessary adjustments to the treatment plan. Tracking of the occurrence of problem behaviors typically takes place during therapy sessions, following the aforementioned assessment procedure. Once a child has completed treatment, follow-up services with families are conducted to determine whether treatment gains are being maintained. These follow-up services are provided in the families' homes and communities post discharge. During these visits, a therapist observes and records data on caregiver implementation of the treatment components and on problem behavior using paper and pencil methods. If needed, additional training is provided in the form of didactic instruction, modeling, rehearsal, and performance feedback.



(a) data logger (size in mm) (courtesy of axivity.com)

(b) sensor strap with attached data logger

FIGURE 3.1: Sensing system consisting of tri-axial accelerometers (left: data loggers), and straps for sensor placement on limbs (right)

## 3.3 Automatic Assessment of Problem Behavior

Logging and evaluating frequency of occurrence of specific problem behaviors is central to assessing whether treatment strategies are effective, and whether treatment gains generalize outside of the clinic. We have identified key challenges with current data collection and analysis methods that we believe Ubicomp systems are uniquely positioned to address:

1. Relying on direct observation to gather data during treatment sessions places a strain on staffing.
2. The need for a high level of agreement between observers necessitates precise definitions of problem behavior that can be subjective and somewhat arbitrary, such as the need to define distance metrics to help observers agree on what constitutes sufficient movement for a hit or kick.
3. Precise measurement is problematic for behaviors that occur at a very high rate (e.g., 1Hz) or at a very low rate (e.g., 1 per week), or for covertly occurring behaviors.
4. Relying on parent-reports may present an inaccurate picture of the extent to which treatment gains generalize to the child's home and school.

We developed an assessment system consisting of on-body sensing (see chapter 2.2.3, illustrated in figure 3.1) and automatic analysis, which has the potential to replicate and augment current clinical assessment practices to yield more accurate, objective, and reliable measurement of frequency and typology of problem behaviors.

### 3.3.1 *Wearable Sensing System*

Problem behavior (Table 3.1) is typically linked to intensive and characteristic physical movements by the individual engaging in the behavior. Thus, our methodology is based on direct recordings of movements using wearable sensors.

The application scenario of recording potentially aggressive and disruptive behavior of vulnerable individuals places specific constraints on a wearable sensing system. Robustness and durability obviously represent major constraints. Furthermore, the system should be designed in a way that maximizes the likelihood of being tolerated by the potential wearer, not to mention safety issues that require effective elimination of potential injuries. In order to capture as much detail on behavior as possible, the use of a single sensing system is inappropriate. Even when optimizing on-body placement of a single data logger [Atallah et al., 2011], chances are high that certain types of problem behavior will be missed. Finally, continuous operation over multiple days needs to be ensured for integration into everyday routine with sporadic clinic consultation only.

*Sensor Straps* Our sensing system is based on four small data loggers that continuously record tri-axial acceleration signals (see below). In order to attach the devices, we designed straps for wrists and ankles that effectively keep the sensors in place even during rough treatment. The straps are made of hypoallergenic and robust fabric with attached Velcro® locks designed to obstruct one-handed removal. After fastening the straps, the sensors, which are housed in a small pocket in the strap, are secured and kept in place with fixed orientation. All borders of the straps are finished with a seam made of extra-strong yarn to ensure durability. The straps are black and very thin (less than one-inch wide).

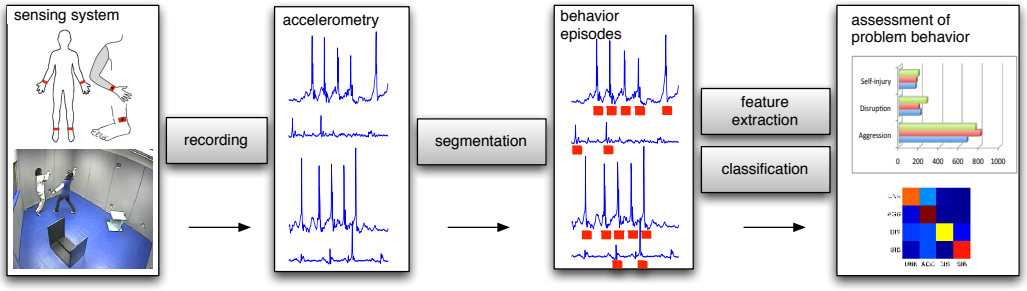



FIGURE 3.2: Analysis of problem behavior based on tri-axial acceleration data – system overview (see text for description; best viewed in color)

*Data Loggers* The sensing system used for capturing behavior data is based on Axivity AX3 data loggers, each consisting of a 16bit micro-controller, a micro-electro mechanical systems tri-axial accelerometer and a large block, single layer chip NAND flash [Axivity, 2013]. Also included are ambient light and temperature sensors (not used in this work), and a real time clock, which is stabilized by a 20ppm oscillator. The device is powered by a rechargeable Lithium-Polymer battery. It is hermetically encapsulated in a tough macromelt polymer, which is shock-proof, food safe, wipe clear and sterilizable using alcohol. Following a full charge the device can log continuous data from all sensors at a rate of 100Hz for a period of 15 days (approx. one week for 200Hz). We chose AX3 data loggers especially due to their robustness and durability, which is necessary for the potentially rough treatment of the devices if assessed individuals actually engage in problem behavior.

### 3.3.2 *Computational Behavior Assessment: System Overview*

Figure 3.2 gives an overview of the analysis system for problem behavior assessment. The sensors attached to the limbs continuously record tri-axial acceleration data and store it to on-board memory (*recording*). To allow for discrimination between different kinds of complex behaviors and for identifying the exact moment of impact, we sampled with a rather high sampling rate of approx. 200Hz within a range of ±16g. Manufacturing tolerances of the data loggers result in differences in the absolute sampling rates of the particular sensors involved. Furthermore, over time, inevitable sensor drifts have to be compensated. Such drifts, caused, for example, by temperature or humidity differences, slightly change the effective sampling rate of the data loggers. To ensure constant and identical sampling rates for all sensors used over the analyzed recording period, all data are resampled to a fixed rate of 100Hz using cubic interpolation.

Recorded sensor streams are then analyzed for behavior episodes (*segmentation*; behavior episodes are underlined in red in Figure 3.2). Feature representations of these automatically extracted segments are fed into a statistical classification system, which discriminates among behavior episodes of aggression, disruption, self-injury, and other (*classification*).

### 3.3.3 *Detection of Behavior Episodes – Segmentation*

The assessment system will be used for the analysis of large amounts of sensor data recorded in sessions of considerable length. In order to effectively process these streams of sensor readings we employ an explicit, lightweight segmentation procedure that identifies behavior episodes before classifying them regarding their type. Behavior episodes represent human activities that are defined by continuous movements resulting in sufficiently large sensor displacements and orientation changes, and include both the problem behaviors we are interested in measuring as well as “regular" activities like a stride or a hand wave. The goal of the segmentation step is to highlight these building blocks of human behavior by filtering the input data and reducing it to segments that can then be analyzed in more detail in the next step of the recognition pipeline (see chapter 2.4). The key idea in our segmentation is to first identify certain characteristic points within the continuous sensor data streams. Based on these *seed points* we identify the boundaries of the surrounding behavior episode in order to capture not only the impact but also to include characteristic motions before and after the specific behavior.

As most problem behaviors involve high amplitude movements (e.g., punch or kick), a main criterion for segmentation are peaks in the short-term signal energy. However, some disruptive events, like tipping over furniture, are mainly composed of characteristic changes in limb inclination that may be missed if energy was the only criteria used. Therefore, our segmentation procedure additionally considers limb inclination changes in terms of relative sensor orientation changes. By abstracting from absolute values it becomes robust with respect to factors such as sensor displacement. Based on the spherical representation of the acceleration signals $\mathbf{x}^{\mathscr{S}} \in \mathbf{R}^3_{\mathscr{S}(\sigma,\phi,\mu)}$ —$\sigma, \phi, \mu$ denote radial distance, inclination, and azimuth— we calculate short term energy $E_1$ and magnitude of orientation change $E_2$ (with $\Delta \sin\{\sigma, \phi\}$ as first derivatives of spherical angles $\sigma$ and $\phi$) using a sliding window procedure. The weighted sum $\mathscr{E}$ of both components serves as a 1D representation, covering signal energy and sensor orientation changes in a compact way:

$$E_1 = 1/N \sum_{i=1}^{N} (x_i^s)^2 \tag{3.1}$$

$$E_2 = 1/N \sum_{i=1}^{N} \gamma_i \tag{3.2}$$

$$\text{with} \quad \gamma_i = \sqrt{(\Delta \sin \sigma_i)^2 + (\Delta \sin \phi_i)^2} \tag{3.3}$$

$$\mathscr{E} = \alpha E_1 + \beta E_2 \tag{3.4}$$

Weights can be derived in cross-validation experiments or explicitly set to incorporate prior knowledge to personalize the procedure (e.g., for slim vs. more corpulent individuals). For our experiments we set $\alpha = 1.5$, $\beta = 1$, and $N = 32$ as the frame-length for the sliding window procedure.

Local maxima in the $\mathscr{E}$-representation of the input data are used as seed points. For peak detection we utilize a hysteresis approach with data-driven threshold estimation. Starting from a particular seed point the surrounding segment is extracted by aggregating adjacent samples until the lower cut-off point is breached. Since the $\mathscr{E}$-representation encodes both energy and orientation change information in a combined signal, this aggregation is very effective. The $\mathscr{E}$-magnitude of energy maxima typically exceeds those of orientation changes by far. Thus, seed points usually correspond to energy maxima, e.g., the moment of impact during a kick. These events are surrounded by orientation changes, i.e., foot approaching before the actual kick and moving back afterwards. Consequently, local minima in the vicinity of seed points represent the boundaries of behavior episodes. Imperfect peak detection combined with this aggregation often results in the generation of segment duplicates, which are reduced to a unique set using straightforward post-processing.

### 3.3.4 *Feature Extraction*

The main criteria for the design of a feature representation of the acceleration input data as it is fed into the subsequent statistical classifier are: *(i)* independence of the resulting representation on the length of the analyzed signals (since we are avoiding explicit sequence models; see below); and *(ii)* the need to capture *characteristic* differences between the activity classes of interest. Especially in the case of the latter and in the light of the target application domain, it is worth reconsidering what kind of differences an automatic analysis system would need to deal with. For example, with *aggression*, the majority of activities correspond to the person hitting someone else. The "target" of the aggressive act typically reacts and may deflect or block the hitting limb. Kicking is usually accompanied by orientation changes for the sensors attached to the feet. In contrast to the rather soft target of aggressive behavior (i.e., human body), *disruption* is directed towards more rigid objects like furniture. The recorded signals show characteristic shapes and/or oscillations after impact. In the case of *self-injurious behavior*, the actor and target are the same individual, and this typically results in more

**Algorithm 1** Feature extraction (segment-wise)

**Input:** accelerations $\mathbf{x} \in \mathbf{R}^{3\times l}$ for segment $s$ ($l$ = segment length), and orientation change signal $\boldsymbol{\gamma}$ (Equation 3.3); $f$ = #Fourier coeff.; $n$ = #ECDF coeff.

**Output:** features $\mathbf{c} \in \mathbf{R}^D$ for $s$; $D = f \times 3 + n + 1$

$\{\mathscr{F}_i^c(s)|i = 1 \ldots f, c = \{x, y, z\}\} = \texttt{calcFTDesc}(\mathbf{x})$

//calculate first $n$ coefficients of ECDF representation of orientation changes

$\mathbf{O} = \texttt{calcECDF}(\boldsymbol{\gamma}, n)$

$NRJ(\mathbf{x}) = 1/l \sum_{i=1}^{l} \sum_{j=\{x,y,z\}} x_{i,j}^2$

$\mathbf{c} = (\{\mathscr{F}^c|c = \{x, y, z\}\} \quad \mathbf{O} \quad NRJ)^T$

forceful impact as the "attacker" deliberately does not deflect or move back. Consequently, these events show the highest absolute energy of all problem behaviors along with unique changes in limb inclination (e.g. hitting the head). Figure 3.3 shows examples of all three classes of problem behavior and the corresponding raw acceleration data (magnitude) recorded by the body-worn sensors.

Based on these constraints and observations, we calculate features that cover: *i)* spectral characteristics of acceleration signals; *ii)* orientation change statistics; and *iii)* explicitly integrate signal energy that is normalized regarding segment length (Algorithm 1). Features are calculated for every detected segment, i.e., behavior episode, and separately for each sensor. First, $f = 33$ Fourier descriptors $\{\mathscr{F}_i^c(s)|i = 1 \ldots f, c = \{x, y, z\}\}$ are calculated for every segment $s$ and per channel $c$ of the acceleration signal $\mathbf{x} \in \mathbf{R}^3$. The actual choice of $f$ has been determined in cross-validation experiments (results not shown). The second set of features consists of a probabilistic representation of the orientation changes within the analysis window. We calculate the empirical cumulative density function (ECDF, see chapter 5) for ECDF-based representations in activity recognition) of the $E_2$ signal (Equation 3.2) and integrate the first 20 coefficients, i.e., a compact yet meaningful approximation of the ECDF, into our feature representation. Finally, the segment's energy, normalized by its duration, is added resulting in $D$=120-dimensional feature vectors per segment and sensor.

The discrimination of problem behavior is based on statistical classifiers. In order to allow for robust parameter estimation of the classification system, our feature extraction process is finalized by means of PCA-based dimensionality and de-correlation (see chapter 2.5.2). Based on the analysis of the Eigenvalue spectrum of a cross-validation dataset, we project the $D = 120$-dimensional feature vectors onto a lower-dimensional sub-space, which captures 95% of the feature space variance.

### 3.3.5 *Fine-Grained Classification of Problem Behavior*

Behavior episodes extracted from recorded sensor signals represent potential candidates for problem behaviors of interest. Two key aspects are of interest for the envisioned applications. First, how many instances of a problem behavior occur over the time of system deployment? Second, what types of different problem behavior occur? The classification step of our recognition pipeline helps answering these questions for those types of problem behavior that can be captured by our sensing system.

Feature extraction produces a compact and meaningful representation of behavior episodes with fixed dimensionality. As the characteristic differences between different types of behavior can be small, plain distance-based classification approaches such as KNN are likely to fail. Consequently, we apply more complex statistical modeling methods for the recognition of problem behavior. We explore the effectiveness of the three main types of statistical classifiers that each focus on different aspects in the statistical modeling process [Duda et al., 2001] (see also chapter 2.6): *i)* Naive Bayes (NB) classifier, a rather simple probabilistic example of generative modeling; *ii)* C4.5 decision tree classifier,

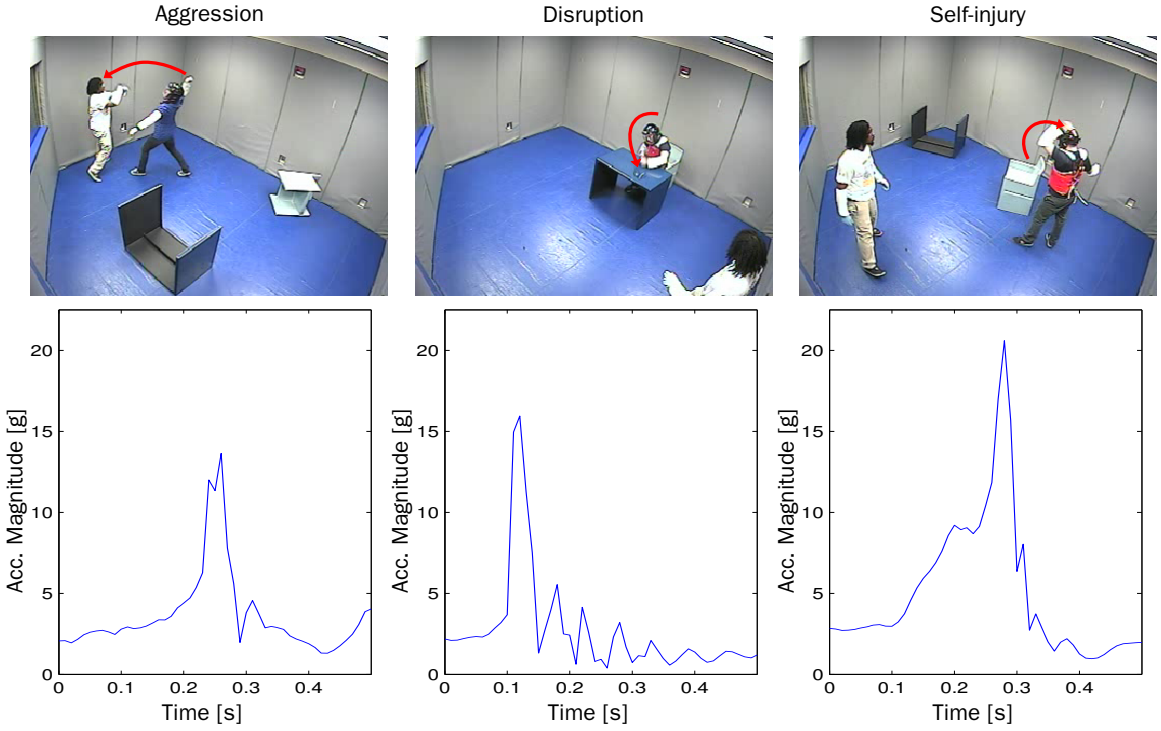



FIGURE 3.3: Examples of problem behavior (top) and their manifestation in raw sensor data (lower row: magnitudes of 3D acceleration signals). Recordings from simulation sessions with staff-members engaging in typical problem behavior (wearing protective gear).

the standard implementation of predictive modeling; and *iii)* Support Vector Machine (SVM) classifier, the most prominent example of discriminative modeling, which has proven very successful for a number of classification problems especially if only little sample data is available for training. We deliberately did not include explicit sequence models into the evaluation (e.g., hidden Markov models) since they are prone to overfitting if data are analyzed that exhibit high *intra*-class but low *inter*-class variance as is the case for the analyzed behavioral data [Fink, 2008].

## 3.4 Experimental Evaluation

The experimental evaluation of any kind of technology designed to assess the behavior of a vulnerable population is challenging. Automatic predictions need to be rigorously validated following established protocols, and based on a solid statistical basis, i.e., a representative and significantly large dataset. However, the collection of such a dataset is hard if solely focusing on recordings of actual clients. Ethical and safety issues are two major obstacles. It is hard to predict if/when an individual will engage in problem behavior, which complicates the recording of such a dataset. We address these challenges utilizing a three-stage experimental evaluation. With this procedure we are in the position to extensively evaluate and validate the developed system.

*Stage 1 (*SIMPROB*)* For system development and validation we recorded a dataset where experienced members of staff of the collaborating behavior clinic simulated assessment sessions in the clinic's facilities. They were asked to engage in a variety of problem behaviors as they experienced them in their clinical practice. This gives us a rich dataset of realistic behaviors that is used for systematic evaluation of our system's segmentation performance and classification accuracy.

*Stage 2 (*ADL*)* Arguably, the high frequency of occurrence of problem behaviors in the SIMPROB dataset is not representative for actual clinical assessment sessions. In order to evaluate the system's precision more realistically we conducted a second set of experiments. We evaluated our system on a standard activity recognition database that covers activities of daily living but —by definition— does not contain any problem behaviors. The success of our automatic assessment system is measured by the false alarm rate, i.e., by the number of falsely predicted problem behaviors.

*Stage 3 (*KID*)* In the third experiment we used the system for a real assessment session in the behavior clinic with an autistic child who engages in problem behavior. We

evaluated the system's capabilities for replicating human expert assessments according to current clinical practice in terms of segmentation's recall and overall classification accuracy.

### 3.4.1 *Data Collection and Ground Truth Annotation*

SIMPROB We recruited five members of the clinic's therapy staff (2 females, 3 males; all right-handed) to help us run 11 simulated assessment sessions within the clinic's facilities. Participants were asked to take on one of three roles: *i)* the individual who engages in problem behaviors; *ii)* the therapist who is typically in the room with the child and is the target of the child's aggressive behaviors; *iii)* the data collector who watches the assessment through a one-way mirror and records the frequency of behavior according to the operational definitions (Table 3.1). Actors changed roles to increase the variability of expression of the various problem behaviors. On average, two minutes of sensor data were collected per session for the actor simulating the child. SIMPROB contains a total of $1,214$ problem behaviors (Table 3.2).

To prevent injuries, actors wore protective gear, including a padded vest, a helmet, and limb-protectors. This equipment is routinely used at the clinic when assessments are conducted with very aggressive individuals. Live-annotation from behind a one-way mirror represents the "best-practice" in data collection at the clinic. The annotator watches the session and notes each time a target problem behavior of interest occurs (time-stamp, type). We cannot assume that this live annotation is accurate as some events might be missed by the annotator and the time-stamp is likely to be inaccurate due to human reaction times. In addition, the live-annotation does not contain information regarding the specific limb involved, which is needed for model training. In order to obtain ground truth (GT) annotation for model training and validation, a trained researcher re-annotated the sessions based on video-footage. She noted the exact moment of impact for each instance of problem behavior, and then categorized its type and the limb involved. Annotation is based on detecting and labeling problem behavior *events*, neglecting their duration, which is standard practice in this clinical assessment.

ADL Arguably SIMPROB contains an artificially high number of problem behaviors. Thus, experiments based on it are ideal for evaluating the precision of our analysis system but recall assessment would be overly optimistic. For a more realistic picture the assessment system also has to be evaluated on "regular," i.e., non problem behavior data. We discarded the idea of extending SIMPROB by letting the actors wear the sensing

| | left wrist | right wrist | left ankle | right ankle | total |
|---|---|---|---|---|---|
| SIMPROB dataset | | | | | |
| aggression | 311 | 377 | 22 | 50 | 760 |
| disruption | 91 | 132 | 12 | 34 | 269 |
| self-injury | 70 | 114 | 0 | 1 | 185 |
| total | 472 | 623 | 34 | 85 | 1,214 |
| KID dataset | | | | | |
| aggression | 14 | 17 | n/a – child did not | | 31 |
| disruption | 95 | 73 | tolerate sensors | | 168 |
| self-injury | 40 | 86 | on ankles | | 126 |
| total | 149 | 176 | n/a | | 325 |

TABLE 3.2: Summary of datasets recorded for system evaluation (GT).

system outside the simulation sessions. GT annotation was difficult to integrate into clinic routine, and impossible to obtain outside the clinic for privacy reasons. Instead, we used an alternative dataset for recall evaluation.

Within the OPPORTUNITY project, a major activity recognition dataset was recorded with a focus on activities of daily living – ADL [Roggen et al., 2010]. A total of 72 sensors of 10 modalities, embedded into objects or body-worn, were employed for recording people's morning routine, resulting in a "particularly large number of atomic activities (more than 27,000), collected in a very rich sensor environment."[Roggen et al., 2010] By definition, the recorded activities do not contain any kind of problem behavior but the complete variety of domestic activities.

We used the OPPORTUNITY challenge task B2 (Multimodal activity recognition: Gestures – test set) for evaluation on $>$ 1 hour of "regular", i.e., non problem behavior data. The annotated activities comprise opening and closing kitchen furniture and appliances, cleaning the table, moving objects, and NULL. Since the recorded morning routine has been conducted with no further constraints in a kitchen environment, the NULL class also contains a large variety of "other" activities, including walking, sitting down, standing up, unspecified hand gestures etc. It is imperative that our analysis system not confuses these regular activities with severe problem behavior. We evaluated the recall of our system by applying it "as is" to the ADL dataset. For compatibility

with our sensing system we used the acceleration data recorded by the limb-worn inertial measurement units, which represents an identical sensor placement as in our other experiments. We upsampled the ADL data from 30Hz to 100Hz using cubic interpolation. Sensor orientations at every limb were manually transformed to match those of our sensing systems. Absolute accelerations were measured in earth's gravity $g$ in both sensing systems. By means of this mapping procedure we ensured that both signal types are comparable.

KID In the third experiment we used the assessment system for the analysis of a real functional assessment session in the clinic. During this session (length: $>$ 50 min.) the child (male, aged 11, weight 63 lbs, right-handed) engaged in 325 problem behavior episodes (168 disruption, 31 aggression, 126 self injury). Manual annotation was obtained using the same procedure as for the SIMPROB dataset. This experiment directly corresponds to the envisioned clinical application case. It also reflects the practical challenges faced by wearable assessment systems such as ours, as the child only tolerated the sensing system on his upper limbs. Consequently, the evaluation is based on problem behaviors observed for the arms only. This child exhibits problem behavior according to a specific pattern. Over the course of the session he engaged in a variety of behaviors that involved playing with toys and high energy activity such as jumping and running around. The therapist that accompanied the child over the course of the session did not disrupt any severe behavior unless there was imminent danger, such as falling off a chair. Often the severe behavior occurred in batches where multiple events followed closely on each other.

### 3.4.2 *Results*

We report segmentation and classification results separately, and for all three stages of analysis (Table 3.3). For the SIMPROB and KID datasets we employed 10-fold cross-validation procedure for classifier training and system optimization. The derived system is then used "as is" for the analysis of the ADL dataset, which does not contain problem behavior, and we report absolute numbers of false positive predictions, i.e., erroneous detections of problem behavior episodes. For these false predictions we also provide classification results with respect to the problem behavior classes of interest.

| | SimProb | | | Adl | | | | Kid | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Segmentation of behavior episodes (BE)* | | | | | | | | | | |
| | Precision [%] | Recall [%] | #BE (GT) | #False Positives | | | | Precision [%] | Recall [%] | #BE (GT) |
| left wrist | 51.3 | 95.1 | 472 | 0 | No problem BE in Adl set | | | 29.6 | 80.5 | 149 |
| right wrist | 63.7 | 96.3 | 623 | 0 | (by definition); overall | | | 32.5 | 81.8 | 176 |
| left ankle | 7.9 | 94.1 | 34 | 327 | duration: >1 hour; average | | | n/a – child did not tolerate | | |
| right ankle | 19.8 | 98.8 | 85 | 350 | length FP: 1.2$s$(±0.3) | | | sensors on lower limbs | | |
| average | 41.5 | 95.4 | 303.5 | 169.3 | | | | 31.1 | 81.2 | 162.5 |
| total | | | 1,214 | 677 | | | | | | 325 |
| *Classification* | | | | | | | | | | |
| | accuracy [%] | | | | accuracy [%] | | | | accuracy [%] | | | |
| | NB | C4.5 | SVM | #BE (pred.) | NB | C4.5 | SVM | #BE (pred.) | NB | C4.5 | SVM | #BE (pred) |
| left wrist | 68.9 | 65.4 | 78.5 | 917 | n/a – no (false) prediction | | | 0 | 69.6 | 63.5 | 71.6 | 395 |
| right wrist | 63.1 | 56.9 | 77.6 | 956 | of any BE | | | 0 | 63.8 | 57.8 | 69.4 | 434 |
| left ankle | 87.8 | 89.7 | 94.7 | 532 | 29.4 | 98.3 | 99.4 | 350 | n/a – child did not tolerate | | | |
| right ankle | 77.8 | 74.9 | 87.6 | 562 | 60.6 | 59.6 | 99.7 | 327 | sensors on lower limbs | | | |
| average | 69.8 | 66.6 | 80.3 | 741.8 | 44.5 | 76.3 | 99.6 | 169.3 | 65.9 | 59.9 | 69.7 | 414.5 |
| total | | | | 2,967 | | | | 677 | | | | 829 |

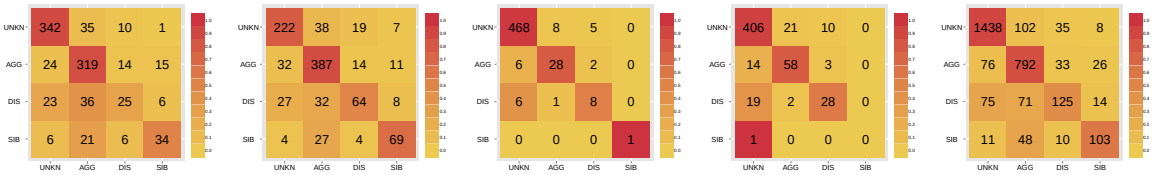
Table 3.3: Evaluation results for all three tasks. Classification accuracy regarding aggression, disruption, self-injury, and unknown based on automatically extracted segments (as reported in upper half – segmentation).

SEGMENTATION The accuracy of segmentation is reported in the upper half of Table 3.3. For SIMPROB, behavior episodes were detected with an average precision of 41.5% and average recall of 95.4% across all limbs. While high precision values were achieved for the detection of behavior episodes on wrists (51.3 and 63.7%), the segmentation lacks precision for behavior episodes involving the ankles (7.9 and 19.8%), which corresponds to over-segmentation. The over-segmentation for episodes involving the ankles stems largely from the abundance of high energy episodes during walking, as each step, particularly when running, produces sharp peaks in the signal energy. Note, however, that such false positives will be addressed in the next step of the analysis, since our classification algorithms will classify these episodes as “unknown" (i.e., not problem behavior related). The human-annotated events that are missed by our segmentation step (false negatives) typically involve low amplitude motions that do not produce a sufficient displacement of the sensors to be detected in the current sensor configuration.

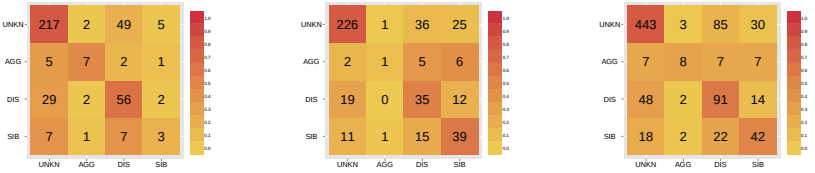
A total of 677 false positives were produced during segmentation of the ADL dataset. Since this dataset does not contain any actual problem behavior episodes we report absolute numbers. Again the lower limbs were more affected by over-segmentation (no erroneous prediction on arms). For the KID dataset, precision of detecting behavior episodes (involving arms only) is largely comparable to the SIMPROB dataset, though recall drops about 14% to 81.2%.

CLASSIFICATION The accuracy of classification of the behavior episodes extracted in the segmentation step are reported in the lower half of Table 3.3. We evaluated the effectiveness of three types of statistical classifiers: Naive Bayes (NB), Decision Trees (C4.5), and Support Vector Machines (SVM; with RBF kernel). For the latter we optimized the slack $C$ and the kernel parameter $\gamma$ in a grid-search procedure as it is standard for SVM-based applications [Schölkopf and Smola, 2002]. SVM-based classification consistently outperformed the other two modeling technique throughout all three tasks.

Overall classification accuracy for differentiating among the relevant classes of behavior episodes was, on average, 80.3% for SIMPROB, 99.6% for ADL, and 69.7% for KID. The confusion matrices in Figure 6.2 (upper row: limb-based and averaged results for SVM-based classification on SIMPROB; lower row: same for KID) indicate that our classification procedure effectively compensated for the over-segmentation effect seen in the first step of the analysis procedure, which resulted in low precisions for detection of behavior episodes.




(a) Left wrist (b) Right wrist (c) Left ankle (d) Right ankle (e) Averaged




(f) Left wrist (g) Right wrist (h) Averaged

| Abbreviation | Behavior |
|---|---|
| AGG | Aggression |
| DIS | Disruption |
| SIB | Self-Injury |
| UNKN | other (unknown) |

(i) Abbreviations used

FIGURE 3.4: Confusion matrices for SVM-based classification of extracted behavior episodes (top: SIMPROB task; bottom: KID task; all matrices row-wise normalized). Absolute numbers may differ from ground truth totals (Table 3.2) due to false negative predictions in segmentation stage.

The averaged confusion matrix for the KID task (Figure 3.4(h)) shows that we can successfully reject unknown instances and differentiate between disruption and self-injury. The successful modeling of disruptive behavior for this specific child is reasoned in the reduced complexity compared to the SIMPROB task, as just a few characteristic motions occur (mainly hitting furniture, walls). Aggression on the other hand cannot be identified as reliably, which indicates large variations for this specific category.

## 3.5 Related Work

The de-facto standard procedure for assessing problem behavior is based either on standardized parent- or teacher-reports [Achenbach and Rescorla, 2000, Aman et al., 1985, Rojahn et al., 2001], or on direct human observation in clinical settings. Although these

procedures are widely employed, and represent current best practice in problem behavior assessment, they result in data that is far from optimal. Reasoned by potentially subjective and arbitrary definitions of problem behavior (not to mention their severity), and difficulties in observing and tracking certain kinds of behaviors, an objective and accurate assessment is often difficult to achieve. These drawbacks served as the motivation for the development of the approach presented in this paper.

Few publications exist that address the use of automatic analysis techniques to assess behavior related to developmental disabilities and autism. Most of these papers focus on specific behavioral phenomena rather than assessing a broader range of behaviors. For example, Goodwin and colleagues developed a system for recognizing stereotypical movements (not problem behavior) in individuals with autism [Albinali et al., 2009, Goodwin et al., 2010]. Similar to our work they used wrist-worn accelerometers for recording data on movements of the limbs. By means of a decision tree classifier a frame-wise recognition of two types of stereotypical movements —hand flapping and body rocking— was performed with satisfying accuracy in two different environments (classroom and laboratory). Westeyn et al. described the classification of a range of self-stimulatory behaviors typically observed in individuals with autism using body-worn tri-axial accelerometers and an HMM-based analysis approach [Westeyn et al., 2005]. Interestingly, they also let an actor mimic the behavior that was the target of the analysis. However, the dataset that was collected is very small and the overall procedure of rather exploratory nature. Finally, Min and coworkers also focused on detecting self-stimulatory behavior in individuals on the autism spectrum using on-body sensing [Min and Tewfik, 2010a,b]. The focus of their work is on exploring the effectiveness of various signal processing techniques.

## 3.6 Discussion

SUMMARY The goal of this paper was to explore how technology and computational analysis can support the clinical practice of problem behavior assessment in individuals with developmental disabilities. We developed a body-worn sensing system and activity recognition techniques that effectively gather objective measures of the frequency of problem behaviors. Using our system we were able to replicate current manual assessments, i.e., clinical best practice, with high accuracy (Figure 3.5). This is very promising, especially in light of an extremely challenging application domain. Children with developmental disabilities pose substantial challenges for the wearable sensing system

(e.g., tolerance by the wearer, durability) and the analysis algorithms (e.g., substantial variability in manifestations of problem behaviors and their similarity to day-to-day activities).

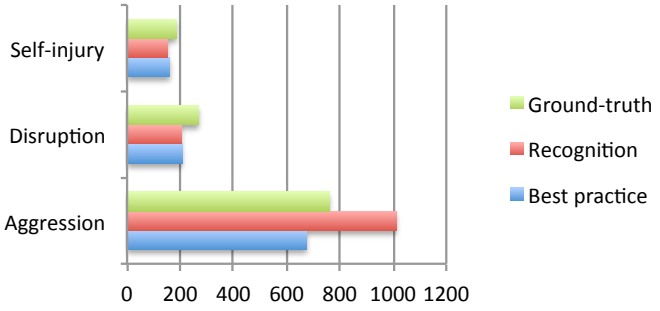



FIGURE 3.5: Comparison of summary reports (count of behavior occurrence) for GT annotation, automated recognition and BP (SIMPROB).

Our main validation experiment included simulated data performed by trained clinic staff. We focused on simulated data for system development and reliability evaluation because it afforded us the opportunity to maintain strict control of the experiments, which is important at this stage of our exploratory research. The staff were instructed to generate a reasonable number of problem behavior episodes across the three classes of problem behavior, which is an obvious advantage over uncontrolled data collection with actual patients. Given the staff's extensive training and experience in working with the target population, the simulated sessions are realistic in the sense that the participants exhibited problem behaviors typically observed and treated by the clinic. Demographic variance in the actors' themselves (e.g., gender, height, weight) further increased variability in the expression of the observed behavior data. Because of these factors, we can hypothesize that these data are a very reasonable proxy for problem behaviors of the target population.

The results of the case study, where we applied our sensing and analysis system to a real behavior assessment session with an autistic child, confirm this hypothesis. We were able to replicate the promising recognition results from the validation experiment with a moderate drop in recognition accuracy. Further evidence of the effectiveness of our assessment system was given by its evaluation on non-problem, i.e., "regular" behavior data. Our system produced almost no false positives on a major activity recognition dataset that contains a broad range of domestic activities of daily living.

LESSONS LEARNED AND FUTURE WORK The case study also unveiled further challenges that we need to face in future work. The first concerns the number and placement of

sensors on the participant. For example, the child did not tolerate the sensors on his legs. Consequently, we were unable to assess problem behaviors linked to activities of his lower limbs using our current setup. Moreover, this child engaged in a number of problem behaviors that wrist- and ankle-mounted accelerometers would not capture, including biting and head butting. Thus, one key future direction for our work is to investigate how to further minimize the number of sensors and adjust their on-body positions for robust and reliable sensing of a larger range of behaviors.

Furthermore, the need for adaptation techniques became apparent as the child engaged in problem behavior in a very idiosyncratic way. For example, his aggressive behaviors showed a large variability in expression, but occurred only sporadically over the course of the session. On the other hand, his disruptive and self-injurious behaviors occurred with much higher frequency and took on very characteristic forms. Such intra-individual variability in expression of problem behavior is clinically meaningful, yet is not being captured using current practices that focus solely on recording frequencies of occurrence. There is much potential for automated analysis systems to quantify such variability.

While the focus of the current analysis was on problem behaviors, we acknowledge that body-worn accelerometers are appropriate for detecting other types of clinically meaningful behaviors that involve body movements, particularly repetitive and stereotyped behaviors (e.g., hand flapping, body rocking) often exhibited by individuals with autism. In the current analysis, these behaviors would have been classified as non-problem behavior related, and as such, placed in the unknown class. However, our analysis can be extended to differentiate these clinically meaningful behaviors from incidental movements and activities also classified as unknown.

Comparing our additional ground truth annotation to current clinical practice, which involves a human annotator flagging the occurrence of problem behaviors as they happen, reveals several sources of inaccuracy in this practice. First, the observed behaviors can occur at a high frequency, at times faster than a human can manually track. Second, the observer may actually be occluded from seeing the behavior. Both of these sources of error can be improved upon by our automated classification. On the other hand, the automated technique can be inaccurate in cases where the problem behavior is very similar to other (non-severe) behavior. We saw this for the classification of disruption behaviors by foot-mounted sensors (e.g., kicking), which are very similar to ordinary walking behaviors. Furthermore, disruption represents the most challenging class of behavior, likely because it involves a more diverse set of activities than, for example, self-injury

(e.g., it includes both hitting the furniture/wall but also swiping objects, tipping furniture, throwing furniture). One way we hope to address both of these shortcomings is to analyze the sensor data across multiple limbs.

The purpose of automated techniques as they were presented in this paper is to enable clinical researchers to explore new areas of inquiry into behavior analysis, beyond simple frequency counting. For example, our colleagues hypothesize that automatically reinforced behaviors (i.e., the sensation or stimulation provided by the behavior is in itself reinforcing) are going to be more consistent than the same behaviors expressed for a different function, such as for attention. We saw evidence supporting this hypothesis in the analysis of the KID dataset where behavior episodes linked to aggression showed large variance in their sensor data manifestations. These hypotheses, and others like them, can now be formulated and tested in terms of the classification capabilities that our computational approach encourages. Finally, in addition to refining our procedure for clinical assessments, we are working towards our goal of an automatic assessment system for settings outside the clinic. Such a system would enable clinicians to gather data on the occurrence of problem behavior in natural environments, which would allow them to track whether treatment gains observed in a clinical setting generalize to the child's day-to-day life.

## 3.7 Implications for activity recognition in naturalistic surroundings

Exploratory studies aimed to investigate the suitability of novel technology such as movement sensors to augment or substitute for clinical assessments typically rely on simulated or scripted behaviour in well-controlled environments. Two reasons for such a study setting can be identified: i) Access to the (possibly vulnerable) target population is difficult or at least restricted, where a demonstrated reliability is required to recruit large numbers of affected individuals; and ii) gold-standard labelling by experts is required to demonstrate the suitability of the approach. While such a setting is suitable to investigate the general technical feasibility it by no means gives a reliable impression of how such a system would perform in more realistic, possibly naturalistic surroundings. The work in this chapter illustrates how this concern can be alleviated by incorporating additional sources of information, namely a case study on an individual with Autism and a publicly available data-set of "regular" physical activities.

### 3.7.1 *Benefits of incorporating a case study*

The technical approach is developed on simulated problem behaviour. Arguably the trained staff is capable of reproducing the range of problem behaviour they observe during typical treatment sessions, which makes this data suitable for the general exploration of the problem setting. However by incorporating data from a child that actually expresses this problem behaviour a number of lessons were learned that would have remained inaccessible if just simulated data was considered in this work.

**Usability of the sensing system** played a key role in our inability to recruit additional individuals in this study. As discussed above, individuals with autism are often sensitive towards objects attached to their body as already new clothing may cause significant distress. When designing the sensing system we were mostly concerned with participants removing the sensing equipment which led to relatively large wrist-bands (at least for the wrist of a child). This illustrates the importance of considering the requirements of the target population when designing the sensing approach, which is crucial for compliance and the quality and quantity of data collected in a study.

**Unknown idiosyncrasies** became apparent in the data collected in the case study. While clinicians already suspected that purposeful problem behaviour such as self-injury may show increased self-similarity it was unknown at the time if that was actually the case. Obviously these aspects are inaccessible if just simulated data is considered. However such idiosyncrasies may be both beneficial and possibly detrimental to the performance of an assessment system. On one hand the large self-similarity can be exploited when designing the system, increasing the likelihood that some level of adaptation of a recognition system would improve the automated assessment. On the other hand the exact way in which an individual performs problematic behaviour such as self-injury may not have been covered in the simulated set of activities, which would mean that all occurrences of this activity may possibly be missed.

**Preventing over-fitting to artificial study settings** is a major concern with hand-crafted feature extraction procedures and finely tuned recognition approaches, particularly on the small data-sets typical for simulated activities. Incorporating a case-study from an individual with Autism showed that, in this case, the performance of the system

is largely retained. If no such data is incorporated the results may lead to a false sense of robustness of the system.

### 3.7.2 *Benefits of incorporating "regular" background activity*

While the problem behaviour simulated by trained staff may be a good proxy for naturalistic behaviour of people with autism it is very difficult to capture suitable background activities in such a setting. As the study was performed in a clinical environment there were not many scenarios for regular physical activities that would provide some level of suitable background activities. As the long-term goal of this work is automated assessment in the private home it is crucial to incorporate some data from this setting in the evaluation of this work. To our advantage there are numerous publicly available data-sets that involve household activities and we chose to incorporate one that best reflected our sensing specific sensing setup (sensors on wrists and ankles). This allowed us to alleviate concerns regarding false-positive detection of problem behaviour, as such data-sets by definition do not include problematic behaviour typical for individuals with autism.

However, it would have been much more convincing to include actual data from people with autism to provide background activities. Collecting such data would be a significant challenge in practice, as continuous supervision would be required to avoid including problem behaviour. Incorporating this data-set therefore corresponds to a compromise between the quality of the background activity and the cost to capture such data.

### 3.7.3 *Summary*

Based on the results and insights obtained in the work presented in this chapter we can come to the following recommendations for data captured in exploratory studies:

- Usability requirements of the target population
  It is crucial to abide by usability requirements of the target population when designing a sensing system, as this may otherwise have a detrimental effect on the quality and quantity of data that it is possible to collect.

- Case studies
  Augmenting a simulated data-set with data from small case studies can effectively

alleviate many concerns regarding the robustness of human activity recognition systems if otherwise only scripted or artificial data would be considered.

- Background activities
  Particularly in clinical settings it is difficult to capture background activities, which may instead be obtained from existing, publicly available data-sets that allow a demonstration of the suitability of the technical approach.

One aspect that remains unadressed by incorporating additional sources of study data is the over-fitting of manually selected feature representation and specifically tuned classification engines towards the artificial study setting. Even though added (semi-) naturalistic data can aid in investigating this issue it remains open how to obtain a feature representation that is robust even towards naturalistic settings beyond trial and error. Even in this study it would have been straight-forward to obtain large amounts of data from individuals with autism as long as no gold standard labelling was required. The next chapter will introduce means that allow exploitation of such large amounts of unlabelled movement data to obtain reliable feature representations for activity recognition systems.

# Chapter 4. Feature Learning for Activity Recognition in Ubiquitous Computing

The work presented in chapter 3 illustrated a common application of the HAR pipeline to differentiate severe behaviour in autism. It was shown that the small scale data-set collected had to be augmented using background data from an established HAR dataset as background activities. The main reason for this augmentation was to investigate the robustness of the individual components of the HAR pipeline, where particularly the feature extraction and the classification approach are prone to over-fitting. The good performance of the resulting system can be attributed to medical prior knowledge and an iterative process to discover suitable components of the proposed approach. However, in many cases this prior knowledge is not available or may be misleading in naturalistic settings, complicating this manual design process significantly.

Particularly the design of a suitable representation that would allow robust recognition is a recognised issue [Figo et al., 2010]. So far, no well-established feature extraction technique has been devised that would alleviate this problem by providing a well-motivated representation for human movement for cases in which prior (medical) knowledge is unavailable. However, it is possible to obtain very large amounts of unlabelled data in naturalistic settings, which are arguably a representative source for the type of data that needs to be characterised by a feature representation. Recent development in machine learning, namely deep and feature learning [Hinton et al., 2006] have shown great promise in deriving robust feature extractors based on very large amounts of unlabelled data. Feature learning methods assume that the characteristics of training data can be discovered by learning how to generate the data, and that a subset of those characteristics are then suitable to differentiate different classes Hinton [2007]. While these methods have been applied to other areas they have so far not been applied in ubiquitous computing. The research presented in this chapter represents an initial exploration of basic feature learning methods applied to inertial time-series, which show very promising performance across different application domains.

## 4.1 Introduction

Activity recognition is a classical (multi-variate) time-series or sequence analysis problem, for which the task is to detect and classify those contiguous portions of sensor data streams that cover activities of interest for the target application. The predominant approach to AR is based on a sliding window procedure, where a fixed length analysis window is shifted along the signal sequence for frame extraction. Consecutive frames usually overlap to some degree but are processed separately. Preprocessing then transforms raw signal data into feature vectors, which are subjected to statistical classifiers that eventually provide activity hypotheses (see chapter 2.4).

As for any pattern recognition task, the keys to successful AR are: (i) appropriately designed feature representations of the sensor data; and (ii) the design of suitable classifiers. The ubicomp literature describes a wide variety of creatively applied classification approaches (see chapter 2.6). By contrast, comparatively little systematic research has addressed the problem of feature design, with almost all previous work using heuristically selected general measures (see chapter 2.5.1). The lack of systematic research on features has been identified as one of the major shortcomings of current AR systems [Lukowicz et al., 2010]. For example, it is questionable whether the next generation of applications, such as behavioural analysis, or skill assessment can be realised based on the use of heuristically selected features alone. Such problems require quantitative analyses of the underlying data which are beyond the capabilities of current procedures for discriminating within limited sets of activities and rejecting unknown samples. Also, as highlighted in chapter 2.2.4 it appears that are not particularly suited for the analysis of data captured in naturalistic surroundings, which is characterised by the lack of detailed annotations.

The most straightforward approach to feature design is to investigate the nature of the data to be analysed and to develop a representation that explicitly captures its core characteristics. For ubicomp AR problems, no all-encompassing model exists to afford the expert-driven design of a universal feature representation. However, recent developments in the general machine learning field have the potential to overcome this shortcoming by automatically discovering universal feature representations for such ubicomp sensor data.

We present a general approach to feature extraction and investigate the suitability of feature learning for ubicomp activity recognition tasks. We utilise a learning framework, which automatically discovers suitable feature representations that do not rely on application-specific expert knowledge. We use unsupervised feature learning techniques, namely (variants of) principal component analysis and deep learning, and show how the automatically extracted features outperform standard features across a range of AR applications. Such an automatic feature extraction procedure has important implications for the development future applications since no manual optimisation is required. The deep learning approach allows for in-depth analysis of the underlying data since the new representation implicitly highlights the most informative portions of the analysed data. This is likely to be important for new classes of activity analysis such as skill assessment.

## 4.2 State-of-the-Art

A recent survey of preprocessing techniques for AR [Figo et al., 2010] distinguished the principal classes of calculation scheme according to the domain of the preprocessing: (i) time domain; and (ii) the frequency domain. The most widely used feature extraction scheme calculates statistical metrics directly on the raw sensor data, independently for every frame extracted by a sliding window procedure. Commonly used metrics include the mean, standard deviation, energy, entropy, and correlation coefficients. Feature extraction in the frequency domain is usually based on Fourier coefficients calculated for the analysis frames (see chapter 2.5). Huynh and Schiele conducted an experimental evaluation of the capabilities of feature representations, namely statistical metrics and Fourier coefficients [Huynh and Schiele, 2005]. They concluded that Fourier coefficient based representations are more appropriate than statistical metrics.

Whereas the majority of published work utilizes standard features a small number of alternative approaches have been proposed. Recently, time-delay embeddings have been used for activity and gait recognition [Frank et al., 2010]. Time-delay embedding is a technique borrowed from physics, where it is used to describe the state of complex systems by means of phase space analysis. This novel representation of sensor data has proved as significant utility in the analysis of repetitive (i.e periodic or quasi-periodic) activities. However, classifiers based on time-delay embedding representations are less appropriate for non-periodic activities. Another emerging approach is to use discrete domain features and to calculate distance measures on string representations of the

sensor data, which has a particular relevance for activity discovery applications (e.g. [Minnen et al., 2006]). However, the quantisation of the sensor data required removes detailed information that is important for the in-depth analysis of certain activities of interest.

## 4.3 Feature Learning for Activity Recognition

Feature learning is a well-studied approach for static data (e.g., object recognition in computer vision). The goal is to automatically discover meaningful representations of data to be analysed. Contrary to heuristic feature design, where domain specific expert knowledge is exploited to manually specify features, feature learning seeks to optimise an objective function that captures the appropriateness of the features. Standard approaches include energy minimisation [LeCun et al., 2006], manifold learning [Huo et al., 2004], and deep learning using auto-encoders [Hinton, 2007].

We have developed a feature extraction framework for sequential data based on feature learning, which is integrated into a general activity recognition work-flow (Fig. 4.1). A sliding window procedure extracts overlapping, fixed length frames from continuous sensor data streams, which in our experiments were the $x, y, z$ data values for tri-axial accelerometers (as described in chapter 2.4.2, upper left part of Fig. 4.1). Frames extracted from raw data are used to estimate the parameters of the actual feature learning procedure (see "fex" block in Fig. 4.1). This feature extractor is then used to transform raw sensor data to be analyzed by the application.

Our design criteria for feature learning ("fex" in Fig. 4.1) are as follows:

1. Capable of extracting generally applicable representations – not be limited to specific AR tasks.
2. Must not rely on the availability of ground truth annotations of the training data.
3. Benefits from larger datasets, but not dependent on them.
4. Provides intrinsic information (for sub-frame analysis).
5. Must be computationally feasible and applicable in real-time application contexts.

Given these design requirements we focused on two learning techniques: PCA and auto-encoder based deep learning.

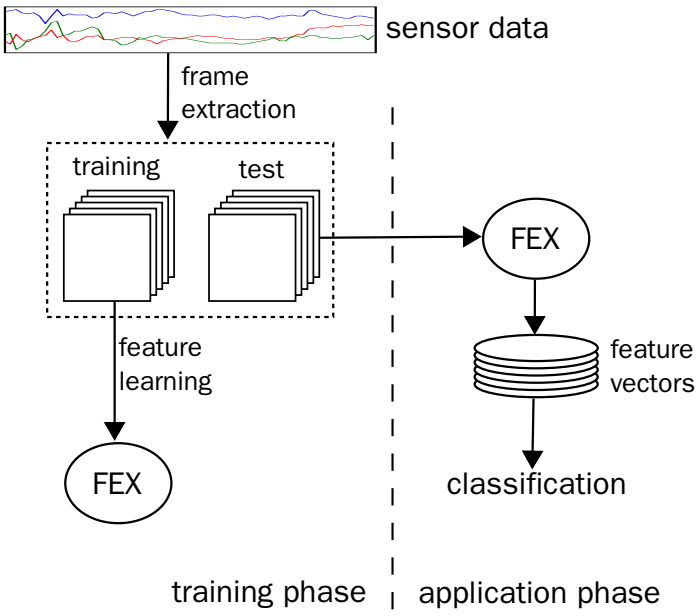



FIGURE 4.1: Feature learning for activity recognition – overview.

### 4.3.1 *PCA based Feature Learning*

PCA is a well established technique used for decorrelation and dimensionality reduction of data (see chapter 2.5.2). PCA is a basic form of feature learning since it automatically discovers compact and meaningful representations of raw data without relying on domain specific (or expert) knowledge. The eigenvectors of a sample set's covariance, which correspond to its largest eigenvalues, are utilised to span a lower-dimensional sub-space that concentrates the variance of the original data. The projection of the original data onto the variance-maximising sub-space serves as a feature representation and can be used either for visualisation or fed into a subsequent classifier. Automatic analysis of the eigenvalue spectrum of the sample covariance uncovers the appropriate target-dimensionality of the feature space.

### 4.3.2 *ECDF-based sensor data representation*

PCA suffers from the limitation that it treats each input dimensionality as statistically independent. In our application setting that assumption is violated, as the input to PCA corresponds to frames of inertial data where subsequent samples are correlated. The features extracted with PCA are therefore inherently based on the *appearance* of the data within each frame, where e.g. characteristic peaks of movement may occur at different

relative positions within each frame. This may be detrimental to the performance of PCA for feature extraction from frames of inertial time-series.

To address this issue we developed an alternative data representation based on the empirical cumulative distribution function (ECDF) of the sample data within each frame. It is inspired by approaches used in other application domains of time-series analysis, e.g., bioinformatics [Chou, 1995], where protein sequences are represented by their amino-acid compositions. The main idea of the ECDF representation is to extract a fixed set of real-valued coefficients that best represents the underlying distribution for each degree of freedom within a frame (i.e. each sensing axis for accelerometer data). To obtain the ECDF representation $f_i$ for a degree of freedom of analysis frame $i$ we first estimate the ECDF $P_c^i$:

$$P_c^i(x) = P(X \leq x) \tag{4.1}$$

To quantify this distribution we select $d$ equally spaced and monotonically increasing points $C = \{p_1 \dots p_d\}$ between 0 and 1. For each of those points we estimate the value $x_k$ for which $P_c^i(x) = p_k$:

$$C = \{p_i\} \in \mathbb{R}^d_{[0,1]},\ p_i < p_{i+1} \tag{4.2}$$

$$f_i = \{x,\ \exists j : P_c^i(x) = p_j\} \tag{4.3}$$

where we use cubic interpolation where necessary to obtain each $x$. The new representation for analysis frame $i$ then corresponds to concatenated ECDF representations of the individual degrees of freedom. Effectively this process provides an estimate for the *quantile function* for each of the selected points in $C$. The ECDF representation is described in additional detail in chapter 5, where it is further motivated and experimentally evaluated.

### 4.3.3 *Deep Learning for Feature Extraction*

Autoencoder networks (or deep belief networks) have proved to be a powerful tool for the generic semi-supervised discovery of features [Hinton, 2007]. These aim to learn a lower-dimensional representation of input data, which produces a minimal error when used for reconstructing the original data. As an alternative to PCA based feature extraction for continuous sensor streams we employed deep learning methods for autoencoder based feature learning on sequential data. The desired representation is discovered by means of a feed-forward neural network that consists of one input layer, one output

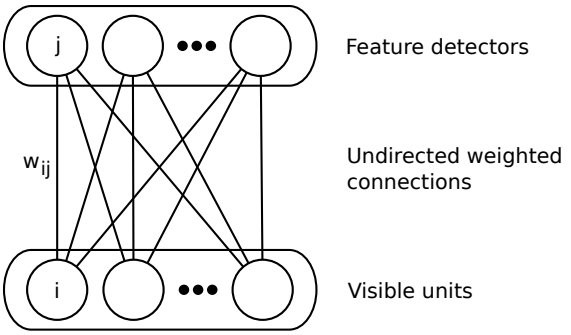



FIGURE 4.2: Schematic illustration of an RBM. Two fully connected layers form an undirected, bi-partite graphical model. The stochastic binary units in the hidden layer act as low-level feature detectors. Different kinds of visible units can be employed though the most popular choices are either Gaussian or binary.

layer and an odd number of hidden layers. Every layer is fully connected to the adjacent layers and a non-linear activation function is used. The objective function during training is the reconstruction of the input data at the output layer. The autoencoder transmits a description of the input-data across each layer of the network. Since the innermost layer of the network has a lower dimensionality, the transmission of a description through this bottleneck can only be achieved as result of a meaningful encoding of the input. This non-linear low-dimensional encoding is hence an automatically learned feature representation.

For robust model training we follow the suggestions given in [Hinton et al., 2006], i.e., we learn the layers of the autoencoder network greedily in a bottom-up procedure, by treating each pair of subsequent layers in the encoder as a Restricted Boltzmann Machine (RBM). An RBM is a fully connected, bipartite, two-layer graphical model, which is able to generatively model data (see Figure 4.2). It trains a set of stochastic binary hidden units which effectively act as low-level feature detectors for the configuration of the visible layer. Each configuration $(\mathbf{v}, \mathbf{h})$ of visible and hidden units has an associated energy:

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i \in \text{input}} a_i v_i - \sum_{j \in \text{features}} b_j h_j - \sum_{i,j} v_i h_i w_{ij} \tag{4.4}$$

Where $a_i$ and $b_j$ correspond to a per-unit bias and $w_{ij}$ is the weight between visible unit $i$ and hidden unit $j$. Each configuration has a probability that depends on its energy:

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{4.5}$$

where $Z$ is a partition function whose that sum over the probabilities over all possible

pairs $(\mathbf{v}, \mathbf{h})$ which is intractable. Learning in an RBM corresponds to lowering the energy of input samples while raising the energy of other samples, particularly those that (falsely) have a low energy. The derivative of the log probability of a training sample $\mathbf{v}$ is given by Hinton et al. [2006]:

$$\frac{\partial \log(\mathbf{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \tag{4.6}$$

$$\Delta w_{ij} = \epsilon \left( \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \right) \tag{4.7}$$

Where $\langle \rangle_d$ denotes the expected value under distribution $d$ and $\epsilon$ is a learning rate. As there are no lateral connections between the units in the layers of the RBM it is straightforward to obtain an unbiased sample from $\langle v_i h_j \rangle_{\text{data}}$ using the conditional distributions for the activation of hidden and visible layer respectively:

$$p(h_j = 1 | \mathbf{v}) = \sigma \left( b_j + \sum_i v_i w_{ij} \right) \tag{4.8}$$

$$p(v_i = 1 | \mathbf{h}) = \sigma \left( a_i + \sum_j h_j w_{ij} \right) \tag{4.9}$$

Where $\sigma$ is a suitable non-linear function such as a sigmoid. Obtaining an unbiased sample from $\langle v_i h_j \rangle_{\text{model}}$ is difficult in practice as it would require sampling a Gibbs sampling procedure starting at a training sample $\mathbf{v}$ and using the conditionals above to alternate sampling of hidden and visible layer until samples become stationary:

$$\mathbf{v} \xrightarrow{p(h|v)} \mathbf{h} \xrightarrow{p(v|h)} \mathbf{v}^2 \xrightarrow{p(h|v)} \mathbf{h}^2 \xrightarrow{p(h|v)} \ \ldots \ \xrightarrow{p(h|v)} \mathbf{v}^\infty \xrightarrow{p(h|v)} \mathbf{h}^\infty \tag{4.10}$$

What made the practical application of RBMs possible was the discovery by Hinton [2002] that an approximation for $\langle v_i h_j \rangle_{\text{model}}$ can be obtained by performing the gibbs sampling for just a few (or just one) steps which is referred to as *contrastive divergence* learning. If $d$-steps of Gibbs sampling is performed the weight-update rule becomes

$$\Delta w_{ij} = \epsilon \left( \langle v_i h_j \rangle_{\text{data}} - \langle v_i^d h_j^d \rangle_{\text{recon}} \right) \tag{4.11}$$

One RBM is trained for each pair of subsequent layers by treating the activation probabilities of the feature detectors of one RBM as input-data for the next. Once the stack of RBMs is trained, the generative model is unrolled to obtain our final fully initialised autoencoder network for feature learning [Hinton and Salakhutdinov, 2006] (see Figure

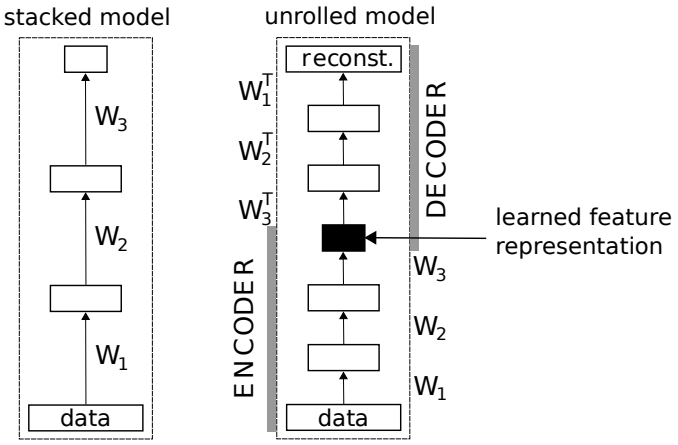



FIGURE 4.3: Schematic illustration of the RBM training procedure. A stack of 3 RBMs is trained to obtain the weight matrices $W_1$ to $W_3$ (left). The model is then unrolled to obtain a fully initialized autoencoder network (right). The innermost layer acts as bottleneck and the activation probabilities of individual units correspond to the feature representation for an input sample.

4.3).

Different methods exist to model real-valued input units in RBMs. We employ Gaussian visible units with unit variance for the first level RBM that activate binary, stochastic feature detectors (Gaussian-binary). The subsequent layers can then rely on the common binary-binary RBM. The final layer is a binary-linear RBM, which effectively performs a linear projection. For the first RBM, equation 4.9 becomes

$$p(v_i = 1|\mathbf{h}) = \mathcal{N}\left(a_i + \sum_j h_j w_{ij}, 1\right) \tag{4.12}$$

During training the sample data is processed batch-wise, where each batch ideally comprises samples from all classes in the training-set. Note that the availability of the class information is not mandatory. RBMs can also be trained in a completely unsupervised manner. However, balancing the batches with respect to the distribution of the classes, i.e. performing semi-supervised training, improves the model quality since it removes the potential for artificial biases.

## 4.4 Experimental Evaluation

To evaluate the effectiveness of feature learning for AR we conducted a number of experiments using published datasets that compared the proposed approach to state-of-the-art heuristically selected features. Sensor data was analysed by means of a (previously optimised) sliding window procedure, extracting frames of $n = 64$ contiguous samples, which overlap by $p = 50$ percent. Feature extraction was then performed on a frame-by-frame basis. The focus of our evaluation was on the capabilities of the particular feature representations. Accordingly, we did not focus on classifier optimisation but on the features themselves. In accordance with the state-of-the-art in HAR systems, we selected a standard, instance-based classification approach, Nearest Neighbour (NN), and applied it "as is" to all tasks.

Given ground truth annotations we report the classification accuracy as percentages of correct predictions provided by the NN-classifiers. The experiments were performed as $N = 10$-fold cross validations (unless mentioned otherwise). Folds were created by randomly choosing samples from the original dataset thereby respecting fold-wise balanced distributions of all classes (i.e. activities to be recognised).

### 4.4.1 *Datasets*

We selected four standard datasets for our evaluation, each of which is described in the literature and is publicly available. All datasets relate to human activities in different contexts and have been recorded using tri-axial accelerometers. Sensors were either worn or embedded into objects that subjects manipulated.

**Ambient Kitchen 1.0 (AK)** Pham et al. [Pham and Olivier, 2009] describe a dataset in which twenty participants prepared either a sandwich or a salad using sensor-equipped kitchen utensils. Modified Wii-controllers were integrated into the handles of knives, spoons and scoops, serving as a sensing platform for continuous recording of tri-axial acceleration data. In total the dataset comprises almost 4 hours of sensor data, approximately 50% of which cover ten typical food preparation activities. Given the sampling frequency of 40Hz, the sliding window procedure produced almost 55,000 frames.

**Darmstadt Daily Routines (DA)** In [Huynh et al., 2008] the analysis of activities of daily living (ADL) is addressed by means of worn sensors used to monitor the daily activities of individual subjects in a living lab-like experiment. Two tri-axial accelerometers (wrist-worn and carried in the pocket) recorded movements at 100Hz. Preprocessing and subsampling yields an overall sampling frequency of 2.5Hz. In total more than 24,000 frames were extracted for both the wrist-worn and pocket-carried sensors using our sliding window procedure. Ground truth annotation used 35 activities of different levels of abstraction. Cross-validation experiments were conducted based on class-wise balanced, random selection of frames for creating the folds. We report results only for pocket-sensor experiments, which, as reported in the original publication, yielded significantly better results than those based on the wrist-worn sensor data.

**Skoda Mini Checkpoint (Skoda)** [Zappi et al., 2008] describe the problem of recognising activities of assembly-line workers in a car production environment. In the study a worker wore a number of accelerometers while undertaking manual quality checks for correct assembly of parts in newly constructed cars (10 manipulative gestures of interest). We restrict our experiments to a single sensor, which is sufficient to identify all 10 activities (i.e. right arm). In total the dataset comprises 3 hours of recordings from one subject (sampled at 96Hz resulting in 22,000 frames). As a result of the unequal distribution of the samples across the classes we were only able to perform 4-fold cross evaluation.

**Opportunity – Preview (Opp)** The final dataset relates to a home environment (kitchen) and the analysis of ADL using multiple worn and embedded sensors [Roggen et al., 2010]. Although the activities of multiple subjects, on different days have been recorded, an official excerpt of annotated data for a single subject has recently been released. Our analysis was based on the sensor data recorded by the accelerometer attached to the right arm of the subject. We considered 10 low-level activities of interest plus an *unknown* activity category. The acceleration data were sampled with 64Hz yielding approximately 4,200 frames.

### 4.4.2 *Features Analyzed: Overview*

To analyze the performance of learned features for activity recognition we performed classification experiments that compared the capabilities of state-of-the-art representations of sensor data streams and learned features as already discussed. To allow comparison of the resulting feature representations we ensured that the target dimensionality of each was in approximately the same range. Since we used instance-based classifiers there was no requirement to use *identical* dimensionalities for objective comparisons. This stands in contrast to generative models (such as mixture densities) where small differences in the dimensionality of the underlying data can have a significant impact on the estimation procedure and hence on the capabilities of the models.

**Statistical Metrics** Probably the most common approach to feature extraction for activity recognition is to use a set of statistical measures to represent frames of contiguous multi-dimensional sensor data. Given the 192-dimensional analysis frames ($64 \times 3$) provided by our sliding window procedure, we first calculated pitch and roll values. Subsequently, for each source channel (i.e. $x, y, z$, pitch, and roll) we then calculated mean, standard deviation, energy, and entropy. Together with three correlation coefficients (estimated for all combinations of the $x, y, z$ axes) this yielded a 23-D representation of the raw signal data covered by an analysis frame.

**FFT coefficients** Characteristic differences in certain activities are apparent from changes in the particular spectra and consequently we can apply frequency transformations to extract feature representations for such classes of activity recognition problems. We performed a channel-wise Fourier analysis on the raw signal data of an analysis frame. Given the resulting spectra we selected the first $f$ coefficients per channel $(x, y, z)$ and concatenated these into a single feature vector. For our experiments we evaluated different choices of $f$. For our dimensionality range (23-39), differences in classification accuracy were negligible – for succinctness we only report the results for $f = 10$ (target dimensionality of 30).

**PCA** We performed experiments utilising PCA-based features where the projection subspace is spanned by those eigenvectors that correspond to the $c = 18$, 23, 30, and 39 largest eigenvectors. These selections of $c$ are justified by significant drops in the eigenvalue spectrum of the data and correspond to the selected target dimensionalities of the other approaches investigated. No significant changes in classification accuracy were observed for the four choices of $c$, hence we present the results for $c = 30$. Experiments

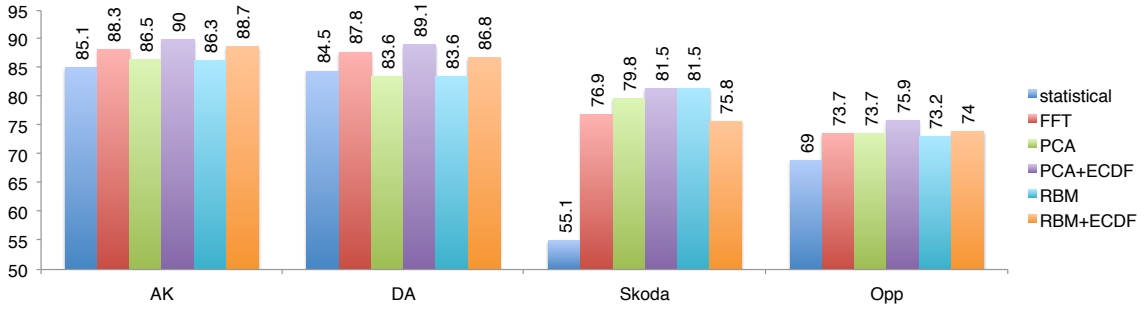



FIGURE 4.4: Classification results for experimental evaluation of learned features and heuristically chosen metrics.

were performed both for the raw sensor data and for the ECDF-based representation. Note that kernel PCA based approaches were ruled out for our unsupervised feature extraction approach due to their exorbitant turnaround times during training.

**Autoencoder (Deep Belief Networks)** Autoencoder networks contain a number of free parameters, including the network topology, i.e., the number of internal layers and its dimensionalities. To show the general applicability of the method, the learning parameters and the network layout (one for the raw data, and one for the ECDF-representation) were tuned on the AK dataset via cross-validation and then used as is for the remaining tasks. The optimised network layout consists of a 4-layer model with 1024 units in each hidden layer and 30 units in the top one (192-1024-1024-30). In all experiments, the first layer was trained for 100 epochs while the subsequent layers were trained for 50 epochs. For the DA dataset, which incorporates a large number of classes (35), the distribution of samples in each batch corresponds to that of the training set, while for the other sets each batch is split equally among all classes, holding 10 samples for each.

### 4.4.3 *Results*

**Classification accuracy** The first set of experiments was devoted to the evaluation of the classification performance as it can be achieved when using the particular feature representations. Fig. 4.4 presents the results for the four analysed datasets. Contrasting our results with those already published for these datasets, we found our results to be broadly comparable (accuracies between 74% and 90%). Interestingly, traditional statistical features performed rather poorly on the Skoda and the Opportunity datasets.

Both variants of learned features lead to statistically significant improvements of the classification accuracy (95% confidence) for all datasets analysed. These improvements on statistical features and FFT based representations are meaningful, especially when we consider that the feature representations have been learned automatically without relying on domain-specific expert knowledge. The results also demonstrate that our feature learning approach greatly benefits from the ECDF-based representation of the input data which yielded significant improvements in classification accuracy for the majority of cases.[1]

In summary, both learning techniques can be used across different AR tasks to discover compact and meaningful feature representations which outperform classical approaches. Features are discovered in an unsupervised manner. For optimization of the deep learning approach prior knowledge about the underlying distributions of the classes is exploited, resulting in a semi-supervised approach.

**Influence of Sample Set Size** The second set of experiments addressed the sparse data problem. Feature learning relies on the availability of sufficient quantities of sample data. The construction of the projection sub-space for the PCA procedure relies on a statistically robust analysis of the sample set covariance. For small datasets the empirical estimation of covariance matrices can result in singularities, which undermines the sub-space creation. Estimating parameters of the auto-encoders for the second learning approach also relies on a representative sample set. Non-representative sample sets bias the parameter estimation procedure such that the resulting features are not flexible enough to capture unknown data.

We evaluated classification accuracies which can be achieved when the training sets used for estimating the feature extraction procedures are artificially limited. Given the original $N$-fold cross validation procedures we gradually removed samples from the training set, performed feature learning as before, and ran classification experiments. Fig. 4.5 illustrates the dependency of the classification results on the amount of sample data available for training the feature extractors. For comparability the x-axis indicates fractions of the original dataset and the y-axis indicates the relative changes in classification accuracy. We ran the evaluation for all four datasets but for the sake of clarity we limit our presentation to the results achieved for the Skoda dataset which is representative

[1] The drop in accuracy for RBM+ECDF on Skoda (compared to plain RBM) is reasoned by an overfitting artifact of the unsupervised approach, which could easily be solved by employing the semi-supervised approach as used for the DA experiments.

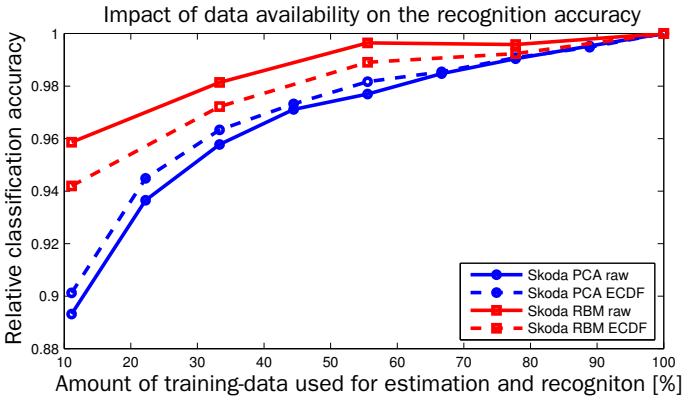



FIGURE 4.5: Exemplary evaluation of sparse data problem.

of the others (for which similar results were achieved). From the results (Fig. 4.5) it is clear that the size of the sample set does not substantially influence the capabilities of the resulting classifiers. However, it seems that PCA has a stronger reliance on the quantity of available training-data compared to RBMs. Given the results of the second set of experiments we can conclude that feature learning meets the third design criteria for practical AR applications.

**Further Analysis** The learned representations can be used for in-depth analysis of the underlying sensor data (the fourth criteria as described in section 4.3). For example, a frame-wise analysis of the reconstruction error provides insights into the quality of the performed activity. Beyond simple clustering, the default choice for quality assessment of activities, more appropriate metrics can be developed that are potentially the key to quantitative activity analysis.

Once the parameters of the feature learning scheme have been estimated (offline) the extraction of learned features corresponds to simple matrix multiplication. Consequently, the results of feature learning can be applied in online interactive applications (fifth design criteria). For some applications the classifiers might even be implemented on the sensors themselves, which would result in a substantial reduction in data transmission and in practical terms a more responsive system.

## 4.5 Conclusion

One of the major shortcomings of activity recognition for ubiquitous computing is the lack of systematic approaches to feature extraction. By explicitly addressing this shortcoming we have demonstrated the suitability of feature learning for AR providing the basis for next generation AR applications. We identified practical design criteria for such activity recognition systems with respect to which we developed an activity recognition framework that employs PCA and deep belief networks for feature learning. An alternative representation of the sensor data, based on an estimation of the frame-wise empirical cumulative distribution of the signal, has been developed. The capabilities of feature learning methods were evaluated by means of recognition experiments on four publicly available AR datasets. Automatically estimated features outperformed classic heuristic features for all the analysed AR tasks we considered. We also demonstrated that feature learning benefits from larger datasets but does not rely on them. The learning approach is computationally feasible and can be applied directly for interactive applications.

Our feature extraction framework has general applicability in ubicomp AR applications, particularly in circumstances where little is known about the target domain. The framework can be used "as is" for activity recognition tasks. Our experimental evaluation provides evidence that feature learning provides reasonable representations, which are immediately usable for further analysis tasks. The deep learning procedure provides sub-frame insights, which is important for a thorough analysis of the captured data.

Based on our findings a number of extensions can also be considered. Although it somewhat circumvents the learning approach we espouse, we could overcome the limitation of current frame-wise analysis procedures that they (typically) treat every sample independently, by explicitly incorporating derivatives into the feature representations. In addition, the linearity assumption could be relaxed during modelling. Non-linear dependencies within the temporal data could be captured by means of kernel PCA approaches for the sub-space projection procedures.

The methodological key to the next generation of activity recognition lies in the systematic analysis of the analysed sensor data. Beyond discriminating fixed numbers of certain activities of interest, domains such as behaviour monitoring or skill assessment require quantitative classifications of the underlying sequential data streams. Our study represents a starting point for systematic research in such sensor data analysis. Given

the promising results of the experimental evaluation, feature learning can be considered as having enormous potential for activity recognition.

## 4.6 Implications for activity recognition in naturalistic surroundings

In both PCA and auto encoder networks the training phase, in which characteristic features of data are discovered automatically, is independent of any labelling of the input data. As discussed in chapter 2.2.4 it is straight-forward to collect unlabelled or unreliably labelled data in naturalistic surroundings if the sensing technologies employed are suitable.

### 4.6.1 *Robust performance on all data-sets*

A particularly encouraging result obtained in the experiments in this chapter is that the performance of feature learning was robust and independent of the investigated application domain. In all cases we were able to outperform FFTs and statistical features with results improving with the amount of training data being available. This implies that these techniques are unlikely to over-fit to even large data-sets, a result which is supported by related work. In practice, existing feature extraction methodologies can be substituted easily with feature learning if the system relies on sliding window segmentation and their application effectively alleviates concerns regarding possible overfitting of manual feature representations described in chapter 2.5.3.

Even though being outperformed by PCA in some of the experiments in this chapter, use of RBMs shows promise for naturalistic settings. They can be implemented very efficiently on graphics processing units which allows very large data-sets to be processed in a short amount of time [Ly et al., 2008]. In this chapter they were applied to extract features but are equally suitable to initialise deep neural networks for classification tasks. Instead of "unrolling" multiple RBMs into an autoencoder network we can simply add a (randomly initialised) top-layer with a softmax-group that represents each class (see chapter 2.6.1). The initialised network can then be fine-tuned for the task of classification by common gradient-based backpropagation, obtaining a powerful classification approach that relies on soft-labels, which has been shown to be efficient if training-data

is corrupted by label-noise [Thiel, 2008]. This approach is applied in chapter 8 where it is described in additional detail.

### 4.6.2 *Novel representation for accelerometer data*

Another insight provided in this chapter is that an efficient transformation can be applied to inertial data to ease the task of feature learning - the ECDF representation introduced in this chapter. Further work revealed that this transformation is very suitable to alleviate some issues with hand-crafted feature representations based on statistical time-domain features. The next chapter motivates the use of this representation and gives a technical description in additional detail.

# Chapter 5. A Novel Approach to the Representation of Inertial Data – the ECDF Representation

When the ECDF representation for accelerometer data was first conceived during the work presented in chapter 3 it showed particularly good performance in combination with feature learning approaches. Investigating the properties of the ECDF representation in preliminary experiments we discovered that it is, in summary, an efficient and practical approach to extract the short-term statistical characteristics of accelerometer data. Given this insight it is no surprise that we saw very good performance when applied to accelerometer data. The state-the-art in HAR systems is to rely on statistical features, which effectively aim to describe this very distribution, just using a manually selected list of attributes and not, as in the case of the ECDF representation, an analytical approach.

This chapter explains the motivation and the technical approach of this novel representation for accelerometer data in detail and illustrates its promising performance in a variety of experiments on 6 publicly available data-sets.

## 5.1 Introduction

Selecting features that adequately model short frames of accelerometer data represents a major challenge in pervasive and wearable computing. The choice of features depends on the application domain and has significant impact upon the classification performance of activity recognition systems. The most widely used approaches fall into three overall categories: i) the use of (hand-picked) statistical attributes extracted from the signal such as means and moments; ii) plain dimensionality reduction techniques such as *PCA* or *FFTs*; and iii) matching a set of explicit patterns to the signals represented in individual frames [Figo et al., 2010, Plötz et al., 2011a, Saria et al., 2011, Berlin and Van Laerhoven, 2012] (also see chapter 2.5).

The majority of activity recognition systems in pervasive and wearable computing rely on approaches from the first category, where statistical features are selected to form a feature-set. This feature selection is most commonly performed by hand, driven by experiments, intuition and experience, although automatic approaches have been proposed (e.g. [Pirttikangas et al., 2006]). In contrast to methods from the other categories, the main objective here is to find measures that effectively quantify the distribution of accelerometer data observed in each frame. Intuitively, this distribution should instead be easily accessible with simple analytical tools. Yet typical approaches to quantify distributions, such as histograms, are difficult to apply to accelerometer data due to their characteristic statistical properties. Even more sophisticated approaches that are informed by an observed or an assumed overall distribution (such as *SAX* [Lin et al., 2007]), may fail as accelerometer data is not independently and identically distributed. A more transparent and efficient approach to reliably describe the distribution of accelerometer data is therefore much desired.

In this chapter we investigate why common methods to quantify distributions fail in the case of accelerometer data and present an alternative approach that clearly outperforms other feature extraction methods across a wide range of scenarios. It circumvents typical pitfalls of naive methods such as histograms and just relies on a single parameter. The *ECDF representation*, first presented briefly in chapter 4, is explained in detail and evaluated in extensive experiments on 6 publicly available datasets representative for the domain. Given the very low computational requirements of this novel approach and the superior performance in realistic settings it is particularly useful for mobile and embedded applications.

## 5.2 Distribution of accelerometer data

Data captured using tri-axial accelerometers correspond to the superposition of the acceleration experienced by the (rigidly mounted) sensor and the impact of the earth's gravitational field as discussed in chapter 2.2.3. Each of the three perpendicular axes of displacement are subject to a bias that is characteristic for the orientation of the device towards ground. This information about the orientation is crucial and sometimes used exclusively to differentiate human activities [Figo et al., 2010].

Figure 5.1 shows the data distribution for one body-worn sensor from *Opportunity*, a large dataset of activities of daily living [Chavarriaga et al., 2013]. Even though the

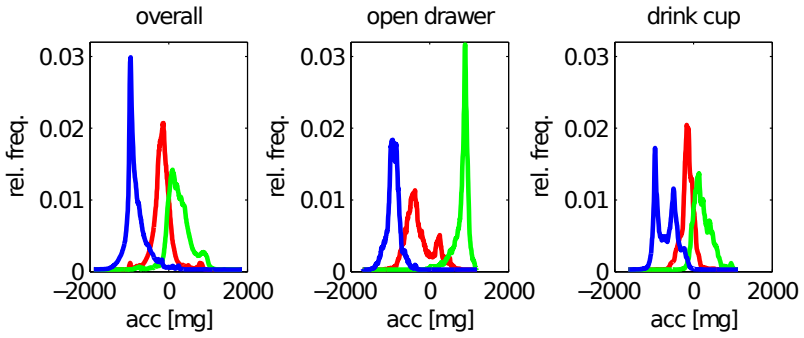



FIGURE 5.1: Data distribution of wrist worn accelerometer in the *opportunity* dataset. The leftmost figure illustrates the overall distribution per axis across the whole dataset. The other two figures show the distribution observed for different activities.

overall distribution for each sensing axis is unimodal, normality can be rejected at 0.1% significance level. If only the data from specific activities are considered, possible orientation changes even lead to bimodal distributions as seen in figure 5.1 (*drink cup*). Simple statistical measures such as mean or standard deviation do not fully capture these characteristics. To allow recognition systems to benefit from descriptors of such distributions, it is important to capture i) their spatial position (bias from gravity); and ii) their detailed shape.

A straight-forward approach to capture such characteristics is to employ simple data histograms. Here a number of bins are placed across the sampling range of the sensor (e.g. $-8g$ to $+8g$) and samples are counted according to which bin they fall into, attempting to approximate the underlying data distribution. Yet in practice choosing the number of bins, and where they are placed, is difficult and has strong implications on the quality of this approximation. If just very few bins are used, fine deviations in shape may not be captured, while a large number of bins may lead to a very sparse approximation with high inter-frame (Euclidean) distance.

More sophisticated approaches, such as *SAX* [Lin et al., 2007, Shieh and Keogh, 2008], attempt to alleviate this issue by placing the bins to represent equal probability mass, informed by an observed or an assumed underlying distribution. The idea is simple, increase the frequency of the bins where samples are expected with high frequency and reduce the granularity of the bins where samples are not expected to occur. This maximises the utility of each individual bin in order to improve the approximation of the distribution observed in an individual analysis frame. However, the distribution of accelerometer data over long periods of time differs significantly from that observed for shorter, individual activities. This violates a crucial assumption of this approach: accelerometer data is not independently and identically distributed (i.i.d.). Therefore

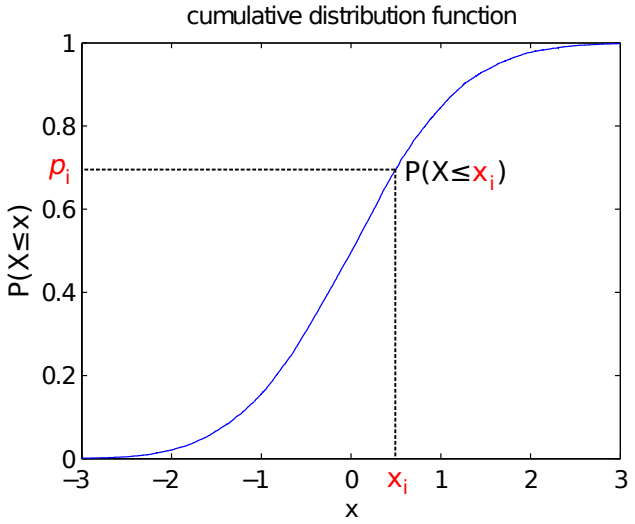



FIGURE 5.2: An example empirical cumulative distribution of random data. To capture the overall shape of the distribution we select $d$ points between 0 and 1. For each point $p_i$ we estimate $x_i$ for which $P(X \leq x_i)$ is equal to $p_i$.

placing bins informed by a distribution observed over a long period of time may actually be counter-productive for the approximation of the distribution in a short analysis frame belonging to a specific activity, which renders such approaches impractical for accelerometer data.

An alternative to explicit histograms is the use of e.g. a Gaussian mixture model to estimate the probability density in each frame [Verbeek et al., 2003]. Given that a sufficient number of Gaussians is estimated, their mixture is capable of approximating arbitrary distributions. While this approach fits with the requirements noted above, it introduces a significant amount of parameters that require careful cross validation, such as the number of Gaussians and possible constraints on their covariance. Furthermore it requires considerable computational effort to estimate the mixture components for a given frame of data, rendering its application challenging in embedded and mobile settings where resources are constrained.

## 5.3 ECDF representation

We identified above that the main challenge with accelerometer data is that it is not independently and identically distributed. This renders placement of the bins required for e.g. a histogram difficult, as it is unclear where to place them to ensure that fine deviations of the distribution are captured. However, there is an alternative view on

distributions which alleviates some of these problems, the empirical cumulative distribution function (ECDF) $P_c$:

$$P_c(x) = P(X \leq x) \tag{5.1}$$

$P_c$ effectively corresponds to the left integral of the original distribution function $P$. By definition, $x$ covers the whole range of sample values observed in the data and $P_c$ is monotonically increasing. The shape of the underlying distribution is reflected in the transition of $P_c$ from 0 to 1, with an example illustrated in Figure 5.2. This opens an efficient and pragmatic approach to quantify arbitrary distributions based on their ECDF $P_c$. Inspired by the approach from common histograms we select $d$ (equally spaced) points between 0 and 1. For each of those points $p_i \in \mathbb{R}_{[0,1]}$ we estimate the value $x_i$ for which $P_c(x_i) = p_i$. For a collection of points belonging to analysis frame $i$, its representation $f_i$ then becomes

$$C = \{p_i\} \in \mathbb{R}^d_{[0,1]},\ p_i < p_{i+1} \tag{5.2}$$

$$f_i = \{x,\ \exists j : P^i_c(x) = p_j\} \tag{5.3}$$

This approach effectively corresponds to finding the inverse of the ECDF $P_c$, which is also referred to as the distributions' *quantile function*. This $d$-dimensional representation $f_i$ fully covers both the spatial position of a distribution, as well as its overall shape. The only parameter is the number of points at which the inverse of $P_c$ is interpolated, which controls the granularity with which the shape of $P_c$ is captured in the resulting representation. This approach can be implemented efficiently using e.g. a Kaplan-Meier estimator [Cox and Oakes, 1984] to obtain $P_c$. Sample MATLAB code is provided in Listing 1, where additionally the overall mean of all axes is added to the representation to improve performance.

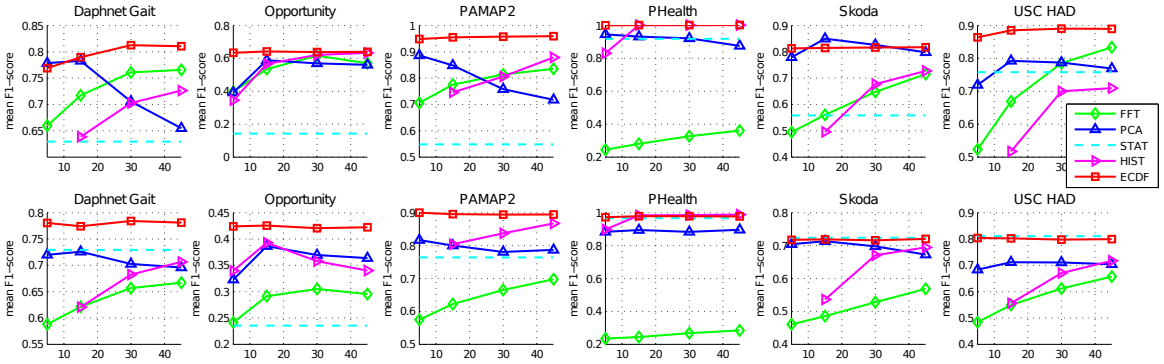



FIGURE 5.3: Results from classification experiments. Each plot shows the mean F1-score for 5 different feature extraction approaches with varying feature parameter (x-axis: number of interpolation points for ECDF; number of bins used for each histogram (HIST); dim. of projection for PCA; number of coefficients for FFT). The top row shows results for KNN classification while the lower one shows results for decision trees (C4.5). ECDF clearly outperforms the other methods with just one exception (PCA+KNN on *Skoda*). Common histograms approximate the ECDFs performance in some cases but appear unreliable in comparison. Already very few interpolation points in the ECDF suffice for excellent recognition performance. The best result on *PHealth* corresponds to 99.94%.

```
% calculate ECDF from D at n points
function X = ECDF_representation(D, n)
  m = mean(D); X = [];
  for d=1:size(D,2),
    [f, x] = ecdf(D(:,d)+
        randn(size(D(:,d)))*0.01);
    ll=interp1(f,x,linspace(0,1,n), cubic );
    X=[X ll];
  end
  X= [X m];
end
```

LISTING 5.1: Calculating ECDF in MATLAB

## 5.4 Experiments

For our experiments we selected 6 different datasets that correspond to typical problems in wearable computing. They include physical activities such as walking, sitting, running, along with gait-freezing in people with Parkinson's Disease (*Daphnet Gait* [Bächlin et al., 2010], *PHealth* [Stikic et al., 2008a]), activities of daily living as they occur in a typical domestic environment (*Opportunity* [Chavarriaga et al., 2013], *USC-HAD* [Zhang

and Sawchuk, 2012], *PAMAP2* [Reiss and Stricker, 2012]) and activities (gestures) captured at a workplace (*Skoda* [Stiefmeier et al., 2008]). All datasets consist of recordings from one or more inertial sensors. Where applicable, we selected solely accelerometers and avoided the inclusion of other modalities, such as heart-rate monitors or gyroscopes. In the case of *Skoda* we limited the experiments to a single sensor (wrist). The reader is referred to the respective publications for details on the exact recording conditions and protocols.

For each dataset a number of experiments were performed, in which 5 different feature extraction methodologies are applied: i) ECDF (this work) to extract $d$ descriptors per sensor per axis; ii) plain histograms based on $d$ bins per sensor per axis between its overall minimum and maximum (HIST); iii) Fourier analysis to extract the first $d$ Fourier coefficients per sensor per axis (FFT); iv) principal component analysis to project the data to its $d$-dimensional principal subspace (PCA); and v) a set of 23 statistical attributes such as means and moments as described in [Plötz et al., 2011b] (STAT). The recordings from each dataset are split into frames with the length described in their original publications (see above, typically around 1 second) and 50% overlap. One experiment is performed for each combination of feature method, dataset and parameter $d \in \{5, 15, 30, 45\}$ (where applicable).

In order to minimise the impact of the classification engine, we decided to employ two non-parametric classification methods. The first is plain $k$-nearest neighbour with $k = 1$, which corresponds to a typical *offline*, instance-based analysis approach where computational resources such as memory are readily available. The second approach is standard decision trees (C4.5) which are particularly popular in embedded and mobile applications where resources are constrained. We chose the mean F1-score as our primary performance metric for the 10-fold cross validation experiments (see chapter 2.6.4):

$$F_1^m = \frac{1}{n} \sum_{i=1}^{n} \left( 2 * \frac{\text{precision}_i * \text{recall}_i}{\text{precision}_i + \text{recall}_i} \right), \tag{5.4}$$

which is more robust towards uneven class distributions typical for applications of pervasive and wearable computing.

## 5.5 Results

The classification results from all experiments are illustrated in Figure 5.3. Each plot shows the performance of all five feature extraction methods on a specific dataset (see plot title) and classification approach with varying feature parameter $d$ (x-axis). The top row corresponds to k-nearest neighbour and the bottom row depicts the results from decision trees. It is immediately apparent that the ECDF approach clearly outperforms all other methods with just one exception (PCA+KNN on *Skoda*). In the case of *PHealth* the problem is effectively *solved*, with maximum mean F1-score of 99.94%. As PCA outperforms ECDF on *Skoda*, it seems that some information crucial to differentiate activities is not reflected in the distribution of accelerometer data found in individual frames. This may be attributed to some activities being inverted versions of one another, such as opening vs. closing the hood, which leads to sensor readings of similar distribution but different pattern, favouring appearance based approaches such as PCA.

The shortcomings of histograms for feature extraction are well illustrated by their recognition performance (HIST) in our experiments. In some cases, their performance approximates that of the ECDF for frequent bins (e.g. k-NN on *Opportunity*). However, in most cases the performance is rather poor, as no reasonable representation can be obtained. Particularly for $d = 5$ the method sometimes fails completely as classes become in-differentiable (omitted values in Figure 5.3). The ECDF representation appears more informative, as already few interpolation points suffice to obtain excellent recognition results. For larger $d$ we just observe modest increases in recognition performance of the ECDF approach. This is in stark contrast to methods such as PCA, where performance often drops significantly when the dimensionality is increased (e.g. on *Daphnet Gait*).

## 5.6 Application in embedded settings

The computation of the ECDF representation consists of two main steps, i) the estimation of the empirical cumulative distribution observed in a frame of data, and ii) the interpolation of its inverse at a limited number of points (see Listing 1). Both steps can be implemented efficiently with computational cost linear in the number of samples in a frame when e.g. a standard Kaplan-Meier estimator is employed [Cox and Oakes, 1984]. Calculating the ECDF representation is therefore well within the computational capabilities of low power PIC micro controllers embedded in modern sensing hardware.

On one hand this allows to reduce storage requirements for systems when deployed, where just the resulting (low dimensional) features are retained for each frame. On the other hand this efficient calculation of features, together with the good performance with an efficient classifier (C4.5), can enable real-time embedded activity recognition on existing sensing infrastructure.

## 5.7 Summary

Many successful systems that distinguish human activities in accelerometer data rely on hand-picked statistical attributes in their feature extraction, which effectively parametrize the distribution observed in frames of data. This indicates that information about how the data is distributed is crucial for recognition. A more immediate approach to quantify the distribution of accelerometer data is therefore desirable. However, the characteristic statistical properties of this distribution render the application of common methods for its quantification, such as histograms, very challenging in practice.

In this chapter we presented ECDF, an approach to preserve crucial information about the distribution of accelerometer data, such as its spatial position and general shape, in an efficient and transparent manner. We demonstrated that the ECDF representation clearly outperforms 4 other approaches to feature extraction common for the domain. Given the superior performance and the low computational requirements, this novel feature extraction approach is well suited for mobile and embedded applications. It is straight-forward to employ the ECDF in existing analysis systems to either augment or substitute current feature extraction methods.

We motivated the ECDF representation based on data collected using accelerometers. The insights obtained are, however, not restricted to applications that utilise this sensing modality. The capability of the representation to capture arbitrary distributions of data makes it a good candidate for feature extraction from other non-stationary time-series, where statistical characteristics change over time.

## 5.8 Implications for activity recognition in naturalistic surroundings

As discussed in chapter 2.5.1, the most common feature extraction approach in activity recognition corresponds to hand-crafted procedures that extract statistical descriptors from frames of accelerometer data. The ECDF representation described in this chapter alleviates the need for this manual, labour intensive process by providing analytic means to preserve crucial statistical characteristics of inertial movement data. In practice it is impossible to overfit this feature representation to a specific study setting as the ECDF features are motivated analytically and not composed of measures selected to maximise a performance score.

Chapter 2.5.3 highlighted how overfitting, and furthermore the reliance on detailed labelling effectively prevents the use of hand-crafted statistical features in naturalistic data. In cases where no reliable ground-truth can be collected for the majority of a dataset it is unlikely that a manual feature selection process yields features that are robust to the unforeseen behaviour encountered in real-life deployments. The ECDF feature representation does not require any training but provides efficient and reliable means of accessing the statistical attributes of accelerometer data. Additionally it seems suitable for use with feature learning approaches as discussed in chapter 4, that are themselves apparently very suitable for such naturalistic settings.

The next chapter describes a specific application of the techniques presented in this chapter and a basic feature learning approach from chapter 4. In order to illustrate the suitability for naturalistic settings this application relies on data captured from a novel setting - data collected from the daily life of dogs.

# Chapter 6. Dog's Life: Activity Recognition for Dogs

Capturing data from naturalistic surroundings can be difficult as it is challenging to obtain reliable labelling. As soon as some degree of artificial study protocol is introduced and subjects are e.g. recorded on video, they effectively change their activities and are likely to deviate from realistic behaviour. This makes exploration of the problem setting difficult as studies capturing naturalistic data are inherently more complex and costly to perform.

However, many of the challenges in movement analysis and activity recognition are not unique to the movements of humans. Many animals, particularly companion animals such as dogs, perform a multitude of activities throughout their day which makes for an interesting and challenging recognition problem in its own right. One crucial advantage of working with animals is that the concerns regarding naturalistic behaviour, or rather deviation from realistic behaviour in artificial settings, is largely inapplicable. Animals such as dogs are unlikely to change their behaviour, as they do not know that their behaviour is monitored. This renders the study of animal behaviour a good test-bed to investigate issues with naturalistic environments. This chapter presents an activity recognition system for dogs that utilises the approaches described in chapter 4, namely PCA-based feature learning and chapter 5 in relying on the ECDF representation of inertial movement data.

## 6.1 Introduction

Humans and dogs have lived together in close proximity for thousands of years [Clutton-Brock, 1999], which has led to strong emotional and social bonds [Serpell, 1995]. By far the largest number of dogs are kept as domesticated pets. For example, in the UK alone an estimated 31 percent of households own a dog, totalling to approx. 10.5 million animals [Murray et al., 2010]. Pet dogs often fulfil the role of companions or even

friends [Menache, 1998]. Apart from this, dogs are widely employed as service animals to perform tasks that are deemed too dangerous, difficult or arduous for humans. Examples include dogs for the blind, search and rescue animals for emergency management, sniffer dogs for narcotics and explosives, and security dogs for policing.

In both service and domestic dogs, the animal's health and well-being are major concerns that are taken seriously for ethical, emotional, and financial reasons. A common definition of animal welfare was laid down in 1979 by the British Farm Animal Welfare Council (FAWC) and encompasses 5 freedoms:*i)* hunger and thirst; *ii)* discomfort; *iii)* pain, injury, and disease; *iv)* fear and distress; and *v)* freedom to express normal behaviour. Whereas the former three have been well researched by veterinarians through direct measurements and observations, the latter are difficult to assess.

Objective observations of both frequency and variability of behaviour traits are key to welfare assessments in dogs where common practice is currently based around manual observational studies and questionnaires [Prato-Previde et al., 2003, Tomkins et al., 2011]. The difficulties these present with regard to logistics as well as upscaling to larger populations are widely accepted as a barrier to gaining behavioural insights. For example, it is difficult to closely monitor dogs in natural outdoor environments or buildings with multiple rooms for long periods of time. However, this is a common situation for many animals to encounter and the majority of domestic dogs spend long periods of time at home alone. In addition, longitudinal measurements are particularly difficult as the frequency of some behaviours (e.g., eating) is difficult to quantitatively report manually.

With a view on assessing animal welfare there is a strong desire to automate detailed behaviour analysis for dogs. However, surprisingly little work has been done so far focusing on in-depth analysis of specific, assessment relevant behaviour traits that go beyond monitoring the general physical activities of dogs. Existing approaches (e.g., [PetTracker, 2013]) focus on logging overall activity patterns and related energy expenditure of the animals. While this allows for high-level analysis, it is not suitable for detailed assessment of specific behaviours and tracking of changes therein, as is desired by both vets and concerned dog owners.

We present an automatic behaviour assessment system for dogs based on a collar-worn accelerometer platform, and data analysis techniques that recognise typical dog activities. For real-world applications the system is capable of recording data for up to 30 days,

| Behaviour (type) | Definition and Movement Characteristics (potential triggers) |
|---|---|
| Barking (A) | Vocalisation of loud sounds. Head is often elevated and thrown forwards at the moment of the bark. Often in bouts of multiple barks.(E,D) |
| Chewing (A) | Object in mouth of the dog. Typical chew motions corresponds to the lower mandibular moving rhythmically. Excludes ***Eating***.(E) |
| Digging (A) | Front paws move in conjunction with each other bimanually (consecutively or concurrently). At least one full motion circle required.(E) |
| Drinking (A) | Series of movements where the dog's tongue touches the liquid up to the swallow. A dog may often stop in between drinking to breathe. The head performs a bobbing motion. (E,D) |
| Eating(A) | Dog swallows the item it has in its mouth. Results in sequence of characteristic movements of the mandibular.(E,D) |
| Excreting(A) | When a dog excretes it will maintain a squatting position. Some dogs may take a few steps when defecating but their bodies are still held in a rigid position. (E,D) |
| Jumping (A) | Movements between the moment the paws (all four) leave the floor until they are back in contact with the ground.(E,I) |
| Laying (P) | Movements between the moment the hock and pastern are in contact with the floor and remain there for more than 1 second, until either the hocks or pasterns are no longer in contact. A dog may also lay on its side or on its back with its legs in the air.(E,I,D) |
| Pawing(A) | Front paws working independently of each other. A pawing action corresponds to repeated backwards pulls towards the dog's belly and hind legs of a single paw. (E,I) |
| Running (A) | Incorporates gaits referred to as *galopping* & *trotting* Edward M. Jr. Gilbert, Thelma R.Brown [1995] resulting in forwards motion of the dog.(E,I,D) |
| Shaking(A) | Twisting motion starting form front of the dog's head and continuing along the whole body down to the tail.(E) |
| Shivering(A) | Muscles around the core of the dog shake in small vigorous movements. (E,D) |
| Sniffing (A) | Nose angled downwards and in close proximity to the floor. Often the head will make sharp side to side movements. Can be done while the dog is in motion or stationary.(E,D) |
| Sitting(P) | Movements between the moment the rump makes contact with the floor and remains there for more than 1 second, until the moment the rump leaves floor. In contrast to ***Laying*** the belly must not touch the ground.(E,I) |
| Urinating (A) | A male dog will often raise one of his rear legs in order to ensure that urine is sprayed in a forward direction. Bitches on the other hand will often squat down so that the urine is sprayed onto the floor between their rear legs. (E,D) |
| Walking(A) | Gait defined by Edward M. Jr. Gilbert, Thelma R.Brown [1995] resulting in forward motion of the dog.(E,I,D) |

TABLE 6.1: Definitions of typical dog behaviours as they are analysed by the automatic recognition system. Types: (A) – Action; (P) – Pose. Potential triggers offer an explanation for deviation in typical behaviour: (E) –Environmental (defined as manually controllable stimuli); (I) – Injury (defined as a recoverable state affecting animals mobility); (D) – Disease (defined as a semi-permenant state that could affect animal psychology or physiology).

and is waterproofed to enable use in rough working environments. We evaluate the system based on the analysis of 16 behaviour traits in 18 dogs, incorporating 13 breeds of various sizes, ages and of both sexes. Our analysis system successfully replicates manual assessments based on hand annotated video ground truth, which demonstrates the applicability of automated dog behaviour analysis for realistic use cases.

## 6.2 Dog Behaviour

The literature on dog behaviour and well-being draws strong correlations on combinations of body movements to specific moods, intentions or desires of the animal, which are grouped as *communicative behaviours* [Daniel S. Mills, 2010, Martin and Bateson, 1993]. Such behaviours are composed of body movements that make up a complex body language the dog uses —often exclusively— to communicate with humans. The majority of such body language is deeply rooted in phylogeny of the genus. For example, when acting aggressively, a dog will often stand in a stiffened pose with its tail straight out, bearing its teeth while snarling or growling. This behaviour is designed to portray the dog in an intimidating pose such that an aggressor will be discouraged away from physical confrontation, thus preventing possible injury.

Dogs can also engage in what is termed *response behaviours* [Edward Price, 2008, Martin and Bateson, 1993]. These types of behaviours often only last short periods of time and are in response to environmental or stimulus based influences. An example of an environmental based response behaviour is shaking after a period of swimming. Response behaviours are not generally linked to dog-human interactions but are often used by vets as symptomatic indicators of injury or disease. For example, in the case of a fractured bone, a dog may spend extended periods of time lying because it is painful to walk.

In the same way humans present mal-state with abnormal behaviour, animal scientists have also linked certain behaviours as key indicators of disease and pain. Very often in the case of a disease such as arthritis that impairs mobility, the gradual initial stages of onset go un-noticed by owners. Such oversight can mean the animal goes untreated and is thus subjected to severe amounts of pain.

Table 6.1 gives a list of behaviours and potential triggers that are key in interpreting sudden or trending changes thereof.

## 6.3 Automatic Analysis of Dog Activities using a Wearable Sensing System

Monitoring and tracking the health and well-being of dogs requires the analysis of the broad range of their everyday activities (Table 6.1). Since most of a dog's activities are linked to substantial physical movements, our automatic analysis system is therefore based on a worn accelerometry sensing platform (see chapter 2.2.3). For almost all of a

dog's activities its head plays an important role, either for directly performing the particular activities (e.g., barking, chewing, drinking), or for balancing full body movements (e.g., walking, running, shaking). It is for these reasons, as well as practical considerations such as comfort and minimal obstruction to the dogs activities, that the collar was chosen as the best site for the sensor.

### 6.3.1 *Sensing Platform*

The platform chosen for recording dog activities is an AX3 logging accelerometer manufactured by Axivity [Axivity, 2013]. It contains a tri-axial MEMS accelerometer coupled to a PIC24 micro controller. The accelerometer can be sampled at a range of frequencies between $2.5 - 3,200$Hz and we chose a sampling rate of 30Hz in order to record detailed movement information. The samples from the accelerometer are timestamped to an accuracy of 20ppm and stored onto an inbuilt 4Gb NAND flash memory chip. With our chosen sampling rate the AX3 is capable of providing continuous data capture for a period of 14 days, which is sufficient for longer-term field studies. The sensor is housed in a tough polycarbonate casing which is hermetically sealed to an IP68 level of waterproofing (1.5m for a period of 1 : 30 minutes) as well as carrying the CE safety mark certification. For sensor configuration and recharging, as well as for data download the platform contains a microUSB port in the side of the housing. Figure 6.1 illustrates the sensing platform.

### 6.3.2 *Data Analysis*

Working with animals carries specific challenges. For example, subjects are rarely cooperative and adherence to a strict activity protocol is impossible. In the case of activity recognition in dogs the former could result in the animal disturbing the sensor placement (e.g., excessive scratching will result in unwanted collar rotations), whereas the latter can result in unusual, idiosyncratic activities that have not been specified in advance. The main requirement for an automatic analysis system is thus its robust and reliable operation, which supersedes the desire of high accuracy recognition of unusual activities.

The collar-worn sensing platform records a continuous stream of tri-axial accelerometer data. Focusing on aforementioned robustness, our analysis procedure utilises a

FIGURE 6.1: Collar based sensor platform used for activity recognition.

segmentation-free analysis approach based on a sliding window procedure for frame extraction (chapter 2.4.2), PCA-based feature extraction (chapter 2.5.2), and an instance-based learning classification backend (chapter 2.6.2). All analysis is based on separate processing of small analysis windows that consist of 1s of consecutive accelerometry samples. Subsequent frames overlap by 50%, which is in line with the state-of-the-art in sliding window based activity recognition (see chapter 2.4.2). For normalisation of the raw sensor data we estimate their empirical cumulative density function (ECDF) and convolute input data with the inverse of the ECDF as described in chapter 5. Keeping to the procedure described in [Plötz et al., 2011a] (see also chapter 4) we further reduce the dimensionality of the features by projecting them onto the first 30 principal components (retaining more than 95% variance). The label for each frame is estimated by majority vote based on the ground truth annotation. However, frames where the majority label constitutes less than 75% of the frame width are withheld from training. This is done to alleviate some of the inconsistencies in annotation that result from ambiguous dog behaviour.

Finally, feature vectors extracted for every frame are fed into the classification backend, namely a $k$-Nearest Neighbour classifier (with $k = 1$), which is trained in a 10-fold cross validation and effectively discriminates between the 16 dog activities specified in Table 6.1 and one rejection class.

## 6.4 Experiments

The overarching goal of our research is the development of an automatic activity monitor for dogs that provides insights into the temporal distribution of a predefined set of animal's behaviours. Such detailed information is invaluable for assessing a dog's health and well-being.

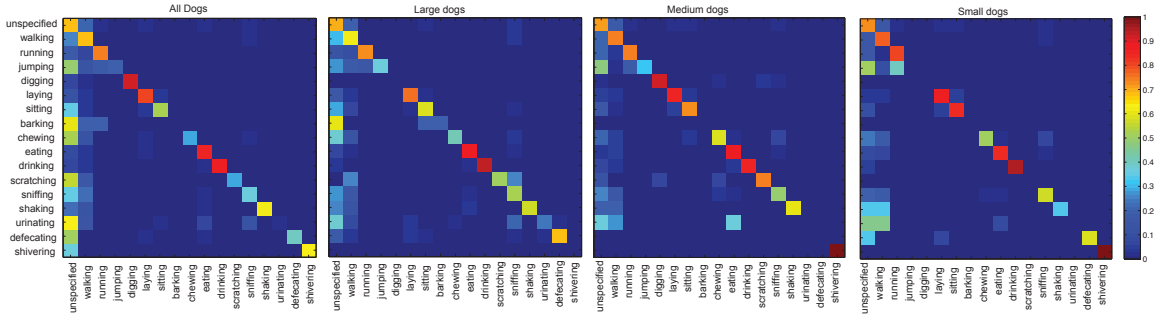



FIGURE 6.2: Confusion matrices illustrating classification performance. From left to right: i) all dogs, ii) large dogs, iii) medium dogs and iv) small dogs (according to Kennel Club sizing criteria).

In order to evaluate the effectiveness of our system we conducted a case study in which we gave the recording platform to a number of dog-owners. We asked the participants to attach the sensor-equipped collar to their dog and record their activities in everyday situations. The owners of the dogs also wore a mobile camera to videotape the recorded activities. The resulting video footage was then used for ground truth annotation, which forms the foundation for our recognition evaluation. All dog owners gave their consent to participate in this study and to use their dog's data for our developments. No animals were harmed while conducting this study, which was conducted in full compliance with the "Animals (Scientific Procedures) Act 1986 (ASPA)" regulations of the UK's home office [ASPA, 2013]. The techniques and protocols used in the study were approved by a University ethics committee.

### 6.4.1 *Dataset*

The dogs used in our case study are listed in Table 6.2. They were chosen so that the dataset covered a wide range of ages, weights, and breeds, as well as both sexes. Each data collection exercise started with the recording platform being configured and attached to the particular dog's existing collar. The dogs were then filmed wearing the collar in both indoor and outdoor settings. For the majority of the recording time the dogs were left to behave how they liked and incidental instances of activities were captured. However, for a few animals it was necessary to make interactions that stimulated them to perform some of the activities in 6.1 (for example ball throwing was used to instigate a bout of running and encouragement to swim was used to instigate shaking). At the end of every data capture session, which lasted between $20-40$ mins for each dog, the recording device was removed and shaken in front of the camera. This action was subsequently used to synchronise the video and accelerometer data.

| Breed | #♀/ #♂ | KC Size Class |
|---|---|---|
| Mongrel | 1/1 | na[1] |
| Miniature Jack Russell | 0/1 | S |
| Dachshund | 0/1 | S |
| Cocker Spaniel | 0/2 | M |
| English Springer Spaniel | 0/1 | M |
| Border Collie | 0/1 | M |
| Bulldog | 1/0 | M |
| Dalmatian | 0/1 | L |
| Labrador | 1/1 | L |
| Great Dane | 1/2 | L |
| Siberian Husky | 1/0 | L |
| Hungarian Vizsla | 1/0 | L |
| Weimaraner | 0/1 | L |
| total: 13 | 6/12 | S:2; M: 4; L: 6 |

TABLE 6.2: Overview of the dogs participating in the experiment.

The data captured was hand annotated (by one expert) against video footage. Based on the definition of dog activities as summarised in Table 6.1 we gave annotators specific and detailed instructions about how to code the activities of the animals. In addition to the specification of the particular activities these guidelines also included precise instructions about coding start and stop points of particular behaviours. Using these instructions we ensured a reliable and objective ground truth annotation of the recorded dataset.

### 6.4.2 *Results*

After frames are extracted from the recorded data streams and labelled according to the ground truth annotation we form 10 stratified folds that are used to both extract features and train the KNN classifier in a cross validation procedure. This results in an overall recognition accuracy of 68.6%. Most of the confusion occurs between the rejection class and *walking*, as can clearly be seen in the confusion matrix (see Figure 6.2). This is largely due to the inaccurate annotation, as the transition between these two activities is not well defined. Other activities can be differentiated at surprising reliability, such as *eating* and *drinking*.

Intuitively, characteristic patterns of the different activities of a dog are heavily influenced by its shoulder height. As an example, consider a very small dog like a Jack Russell, whose modes of transport differ significantly from a Labrador. In order to investigate whether results improve if dogs are grouped according to size, we conducted an experiment with three groups of dogs, classified according to the Kennel Club criteria [Club, 2013]. Confusion matrices are illustrated in Fig. 6.2. The results improve significantly, particularly for small dogs, where modes of transport along with other activities such as digging can be differentiated much more effectively.

It is also anticipated that the results could further be improved through the addition of a multi-variate window size. For example features such as barking and jumping are temporally short in nature and are not well segmented using the same window as sitting or lying.

## 6.5 Summary

Health and wellbeing are of major concern for both domesticated pet dogs and service canines. We have presented a collar-worn activity monitor and a classification system that is capable of recognising 17 dog activities that were expertly identified as being relevant for dog behaviour traits. In a large scale experimental evaluation we have demonstrated that our approach can successfully recognise the aforementioned activities with a reliability of approximately 70%.

The system is the first of its kind that allows for behaviour monitoring of dogs in naturalistic settings. This is important especially for monitoring the welfare of animals that spend a significant time on their own where the owners typically do not have detailed information about their dog's everyday activities and wellbeing. Furthermore, our system could be used for objective assessments of injury recovery and healthiness in service dogs, which has substantial economical impact.

In the future we will explore coupling the core elements demonstrated herein with a fully-automated wireless data transfer and a meaningful graphical visualisation. Such a system has the potential to deliver real time web-based results which opens up the design space for a range of applications such as early warning systems or progress tracking.

[1] No KC classification for mixture breeds.

## 6.6 Implications for activity recognition in naturalistic surroundings

This chapter illustrates the use of state-of-the-art activity recognition techniques for the analysis of dog activities. When studying the impact of naturalistic settings it is beneficial to work with animals, as they are unaware that their behaviour is being monitored and privacy concerns are minimal. Dogs therefore act inherently naturally, particularly in familiar surroundings such as at home or outside in a park. During data collection we did not instruct the dogs to perform specific activities but simply observed their day-to-day interactions with their owners during e.g. natural play. Even though the work in this chapter is an interesting application in its own right it provides some insights for activity recognition in naturalistic environments.

Overall we observe that particularly the unreliability of labels captured in this study poses an issue towards training and evaluation of activity recognition systems. Crucial for the performance of the recognition system presented in this chapter is to avoid including input frames with unreliable labelling in the training set of the classification engine. Even though annotators were provided with high-resolution video recordings and detailed descriptions of each relevant activity it was often unclear when each activity begins and ends. The inherently naturalistic behaviour of dogs simply defies attempts to categorise precisely for small time-scales, and it is likely that this observation also holds for naturalistic behaviour in humans.

Even with very precise and extensive definitions of each activity there will be cases on the boundaries of activities such as walking where reliable annotation is practically impossible. This work relied on explicitly neglecting those periods of time where the annotation is assumed to be unreliable. In many applications it may however be difficult to estimate this for each frame of input data. It is therefore crucial that the recognition approaches employed for naturalistic data are capable of handling unreliable labelling without the need for manual intervention. As discussed in chapter 2.6.5 methods relying on soft-labels appear robust towards this type of noise in the labelling [Thiel, 2008]. Given the promising performance when applied for feature learning in chapter 4 it is deep learning that appears particularly suitable for naturalistic settings.

# Chapter 7. ClimbAX: Automated Skill Assessment for Climbing Enthusiasts

This chapter investigates the use of body-worn sensing for the detection of climbing activity and the automatic extraction of specific skill parameters related to climbing performance. The novel application is aimed primarily at the use within a climbing hall, where people engage in a variety of activities beyond climbing, such as belaying, performing other exercises, warming up or simply having a break. Yet beyond the simple detection of climbing activity we were interested in detailed analysis of climbing activity in extracting insightful skill parameters.

In practice this represents a dilemma for data collection that is typical for naturalistic deployments of HAR systems. On one hand a significant volume of climbing activity and a representative background is required to train and evaluate a climbing detection approach. On the other hand, detailed recordings of climbing performance that include some quantitative measure for climbing skill are necessary to develop and evaluate the skill assessment procedure. Practically, it is very difficult to capture a single data-set that is suitable for both tasks. Climbing halls are a challenging surrounding for video recordings due to lighting conditions, height and other aspects such as dust. In a realistic climbing scenario each participant would climb different routes at their leisure, introducing many (uncontrolled) parameters that complicate the analysis of climbing skill. A data-set suitable for early exploration of climbing skill has to control for such parameters while also supporting high definition video recordings for annotation.

The work in this chapter therefore relies on two separate data-sets: i) a naturalistic data-set captured from a climbing hall from realistic climbing activity for which only low-resolution, diary-based labelling was recorded; and ii) one data-set from a semi-naturalistic climbing competition where we were able to control for some variables such as the exact layout of the route that participants climbed. Overall it illustrates how the use of such a study-setup allows the development and evaluation of a skill assessment


(a) Illustration of the *ClimbAX* assessment system including visualisation of analysis results produced.


(b) Examples of climbing sub-disciplines: Indoor bouldering; Sport climbing; Deep water solo; Ice climbing; Traditional climbing; Aid climbing (*i* to *vi*)

FIGURE 7.1: Overview of the *ClimbAX* system for automatic climbing skill assessment and its potential application cases. See text for description.

system that is likely to be robust towards real-life deployments. The technical approach is based on feature learning described in chapter 4.

## 7.1 Introduction

The sport of climbing has become increasingly popular and is now widely enjoyed as a recreation activity as well as a competitive sport. For example, in the UK the sport "has been on a upward trend since 2005" with a continuous increase in participation [Coldwell, 2012]. The Italian Alpine Club, which is the world's largest, reports the sport in general has had a population growth of 10% since 2009 [CAI, 2013]. As a recreational activity climbing holistically improves both physical and mental fitness, provides a basis for social interactions, and is a way to enjoy the outdoors. Climbing is also being recognised as a competitive activity, and was considered for inclusion in the 2020 Olympics [IOC, 2011].

Similar to other sports, professional climbing requires physical conditioning, applied sports science and training. Elite climbers follow strict training programmes defined with the assistance of and monitored by a coach. In a typical session, a coach will assess the climber through observation and then provide feedback by commenting on their technique, or suggest training routes that will assist in addressing weaknesses. At amateur level, coaching is also desirable and is a service offered by indoor climbing centres. However, the sheer number of climbing enthusiasts render detailed and frequent

feedback from a coach, as it is received by elite athletes, impractical for the amateur. Consequently, amateur coaching is often a group exercise with a typical 1:8 coach to student ratio. The heterogeneity of such groups in terms of climbing skills and experience results in only general feedback rather than in-depth, personalised recommendations and advice.

A wealth of related work exists on self assessment of physical activities using mobile sensing platforms (e.g., [Möller et al., 2012] and references therein). Commercially available devices, such as Nike fuel band [Fuel, 2013] and Fitbit [FitBit, 2013], are effective for improving levels of activity simply through providing and visualising statistics to the user that are related to the frequency and —to some extent— fatigue [Barry et al., 1992]. Some sports self assessment tools are available that focus on the technical skills of the athlete by providing detailed information, not just about the frequency, but also about the quality of the particular activities. Examples include the automatic analysis of golf swings [Grober, 2009] or automatic assistance for swimmers [Bächlin et al., 2009].

In line with the aforementioned analysis tools, we have identified the assessment of climbing skill as a case for ubiquitous computing. We have embarked on developing *ClimbAX* – a sensing and analysis system that replicates professional climbing assessment as it is conducted by human coaches.

*ClimbAX* utilises wrist-worn accelerometers to capture a climber's movements in naturalistic settings. Climbing episodes and individual hold transitions are detected automatically, forming the basis for performance analysis. A variety of performance attributes are developed in this work, which, while being meaningful to climbers, resemble traditional, subjective assessment performed by a professional coach. This climbing skill assessment aims to support future automatic coaching systems that incorporate this objective performance information to devise training plans tailored to the individual.

*ClimbAX* records the climber's movements using a wrist-worn sensing platform that logs high-resolution, tri-axial accelerometer data. This platform is small and sturdy, and does not hinder the climber in their activities. The aggregated data is then processed using an unsupervised analysis procedure, which automatically:

*i)* filters out climbing from background activities;

*ii)* segments climbing sessions with respect to transitions between *holds*, i.e., those moments where the climber remains stationary (fixating themselves on the face they are scaling); and

*iii)* performs climbing skill assessment based on an objective quality scoring scheme.

We evaluated our assessment system in a large field study in a premiere indoor climbing centre assessing the performance of 47 participants of an open bouldering competition event and 6 climbers practicing sport climbing.

The sensing and analysis system presented in this paper allows amateur climbers to track a set of physical performance skills, which can be used either for self-directed training or as a basis for external coaching, and thus improve their performance whilst maintaining health and safety. Figure 7.1 illustrates the developed system and its potential application cases. Objectively measuring climbing relevant parameters represents an important building block for an automatic coaching system as we are aiming for with *ClimbAX*.

## 7.2 Climbing as a Sport

The term *Climbing* is used to collectively group many sub-disciplines each having their own distinctions relating to terrain type, accepted ethics regarding protection and tactics used to ascend [Fyffe and Peter, 1997]. Figures 7.1(b) shows examples of the most widely performed sub-disciplines of climbing.

Popular types of climbing are: *i) Bouldering*, which involves the ascent of relatively low level routes on free standing boulders with just a crash pad to protect the climber in the case of a fall; and *ii) Sport climbing*, where the climber clips their rope into bolts that are pre-placed into the rock, and in case of a fall, a second person ("belayer") will hold fast the rope (with assistance of a friction device) to prevent the climber hitting the ground. Further outdoor climbing sub-disciplines include: *iii) Deep Water Solo* (also known as *Psicobloc*), where the climber uses water below to break a fall; *iv) Ice climbing*, where the climber uses the assistance of crampons and ice tools to ascend; *v) Traditional*, a discipline that employs a strict ethic that all protection placed in the rock must be placed by hand and be removable without damaging the rock; *vi) Aid climbing*, where the climber is permitted to use placed protection as hand and foot holds. *Alpinism* is another discipline that combines aid- and ice climbing at high altitudes. Bouldering and Sport climbing are also frequenty practiced indoors on man-made walls, often constructed from plywood, using shaped resin holds.

### 7.2.1 *Dangers and Difficulties*

Climbing carries risks both in the form of objective danger (for example, a rock falling) as well as an injury through poor judgement of the condition of the terrain, or through poor climbing performance. Little can be done regarding the former other than carefully assessing the general conditions (e.g., weather, composition of the targeted face to be scaled), whereas the main influencing factor for the latter is lack of experience and misperception of one's own skills [Twight and Martin, 1999]. Unrealistic judgments can lead to wrong decisions regarding the individual appropriateness of particular climbing routes, which can have fatal consequences.

The decision whether or not to embark on a particular route is heavily influenced by knowledge of the climber's abilities, which is typically gained through comparison to others who have already completed the particular route. Making objective comparisons between climbers' abilities can lead to both more informed and confident decisions regarding whether a particular route is appropriate for an individual.

Climbing routes are typically ranked according to their difficulty using established grading schemes, such as the internationally recognised French grading system for sport climbs or the Hueco "V" grading system for boulder problems [Fyffe and Peter, 1997]. Gradings typically do not transfer well between sub-disciplines. However, they share the underlying principle of judging how difficult climbs are technically. In the case where there is an apparent objective danger (often judged by the outcome of a fall) a second grade is often given that can be used to interpret the "seriousness" of the route. In the British Traditional System, this grading is descriptive rather than numeric. For example, a route may be classified as "Difficult" or "Very Severe" [Fyffe and Peter, 1997].

### 7.2.2 *What it takes to get high*

Across its sub-disciplines climbing requires a range of physical abilities. For example, climbing large mountain routes requires very good all round stamina, endurance and tolerance to high altitudes, whereas challenges linked to bouldering are often gymnastic in nature and require physical strength, good general coordination, and muscular flexibility. Furthermore, within each sub-discipline there is also scope to specialise for a particular type of terrain. Some climbers for example prefer scaling steep overhanging rocks, which requires very good upper body strength. Others focus on routes that consist

of large numbers of hard individual moves, which necessitates power endurance. Despite this diversity all climbers need to possess a core skill set, which subsumes at least four main physically trainable competencies: *i)* Power used to transition between holds [Quaine et al., 1997]; *ii)* Control over limb movement [Testa et al., 1999]; *iii)* Speed of ascent; and *iv)* Stability whilst on a hold [Horst, 2008, White and Olsen, 2010].

Investigating the reasons for good or bad climbing performance, some studies have gone as far as measuring plasma cortisol (stress hormone) in climbers during and after high stress activities [Draper et al., 2012]. Positive correlations to confidence as well as to somatic and cognitive anxiety in climbing were found. Other studies have measured heart rates as both fatigue and stress indicators. These however, did not unveil any insight due to muscles operating in anaerobic state during climbing [Mermier et al., 1997]. In contrast to such biochemical parameters the climber's experience is difficult to assess. Experience helps a climber to identify the most efficient way to climb through a challenging sequence of moves, and it can help identify the most likely weather conditions that will result in a successful ascent (climbing is highly dependant on rock friction which increases as temperature decreases). In either case it is difficult to reason about the mental state or experience of a climber other than through observing how they perform physically.

It has been demonstrated that parameters relating to the physical performance of a climber can be measured at the interface of the hand and the hold. These parameters vary from core body strength to balance and contact strength [Fuss and Niegl, 2009]. Related studies have exclusively used holds instrumented with strain gauges or vision based systems where climbers were instrumented with markers. While these methodologies demonstrate the validity of the parameters, they are not suitable for deployments in realistic settings. Only very few and rather explorative attempts to automatically asses climbing skills in a real-world context have been undertaken thus far. For example, Pansiot *et al.* attached an accelerometer to a climber's head for recording their movements [Pansiot et al., 2008]. In a small study with 4 participants they derived climbing skill parameters, which although interesting, did not map to any recognised parameters from the sports science literature.

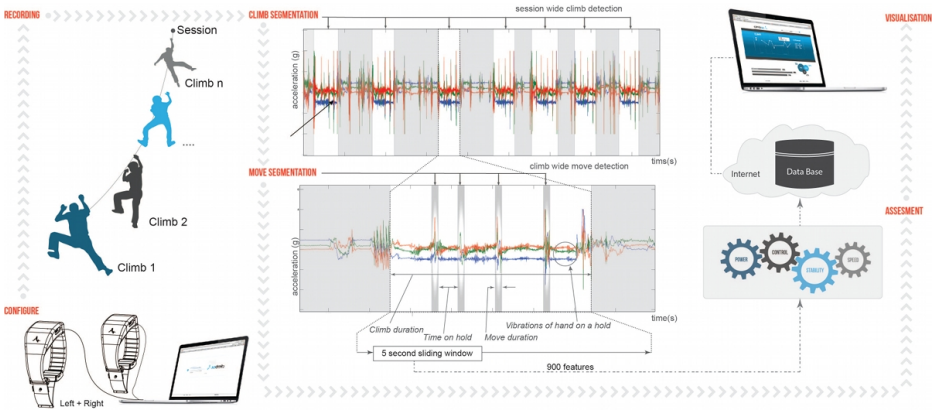



FIGURE 7.2: *ClimbAX*: System overview (see text for description).

## 7.3 Automatic Climbing Performance Assessment

The key to performance improvement in climbing is both increased frequency of exercise [Cordier et al., 1994] and training specific weaknesses and elements of technique [Horst, 2008]. In the elite class these training goals are typically managed with the assistance of a coach. Although a direct transfer of such manual coaching programs to the population of amateur climbers is desirable, resource limitations render expert coaching impracticable. Alternatively, automatic assessments have the potential to make coaching more widely accessible.

Structured and guided self-monitoring and self-assessment represent a reasonable alternative to costly professional coaching. A few technical systems have been developed that support amateur climbers in keeping track of their exercises. For example, smart phone applications are available that walk climbers through sets of fixed routines and record the date they were completed; essentially corresponding to an electronic climbing diary for retrospective (manual) analysis [Beastmaker, 2013, ClimbCoach, 2013]. Such technology supported climbing diaries (and variants thereof) can effectively support climbers in keeping up regular exercising or even increasing participation frequency, which in general has positive effects on their health [Fentem, 1994].

Automatic coaching aids for climbing are required to not only report the frequency and duration of exercise but also a performance breakdown that is presented using terminology that is familiar to the sport. *ClimbAX* has been designed to comply with these requirements. Figure 7.2 gives an overview of our system. Movements are captured using small, wrist-worn sensing devices, which are *configured* to *record* tri-axial acceleration data with high temporal resolution. After a session (which can contain multiple

climbs) the sensor data is uploaded to an analysis platform where climbing orientated data is automatically filtered out (*climb segmentation*) and the moves within each climb are automatically detected (*move segmentation*). Based on the extracted moves, the actual *assessment* is then performed, which is informed by standard climbing grading schemes. Finally, the results are *visualised* both on a session summary basis and at the more fine-grained level of detail corresponding to particular skill criteria from the assessment.

### 7.3.1 *Recording*

With a view on practical deployments in realistic, i.e., non-laboratory, climbing scenarios we adopted a body-worn sensing approach for capturing climbing activities. Apart from the advantage of universal applicability due to minimal requirements on existing infrastructure (such as independence on calibrated camera setups [Sibella et al., 2007]), a wearable, and thus mobile, sensing platform has the benefit of providing detailed and high-resolution data through direct measurements of the climber's movements. Accelerometry in general has proven very effective for assessments of human movements in a variety of application domains [Chen et al., 2012]. In line with previous, explorative studies [Pansiot et al., 2008, Schmid et al., 2007] we employ tri-axial accelerometers for our automatic climbing assessment framework.

Transmissions of rotational and vibrational forces in the range of $0.2 - 20$Hz (human movement range) that are exerted through the fingers have been shown to be measurable using an accelerometer placed on the wrist [Murgia et al., 2004]. Consequently, and coupled with the high level of user compliance the wrist affords, *ClimbAX* sensor system was designed around a watch embodiment. Since climbing requires good symmetry and balance we instrument both wrists of the climber in order to capture the movements of the hand that is transitioning as well as the hand supporting the body during transitioning.

Actual applicability for realistic climbing scenarios requires the movement capturing subsystem of *ClimbAX* to record for a minimum of one day, to be light-weight, scratch proof and hypo-allergenic, and to be sturdy enough for operating in chalky/dusty environments. Accordingly we designed a watch-like sensing platform as shown in Figure 7.3. At its core is a 16-bit, 16 MIPS PIC24 processor, and a 14-bit tri-axial accelerometer (MMA8451Q by Freescale). Sensor readings are sampled at a rate of 100Hz, which provides sufficiently detailed movement information. Samples are stored onto a 4Gb

sized NAND flash memory chip along with associated timestamps (accurate to 20ppm and generated from the PIC24). Communication with the device, e.g., for configuration and data download, is based on a micro-USB connector. The internals of the sensing platform are potted into a poly-carbonate injection moulded case, which is housed by a silicone wrist band. The band was designed to be thin enough to see the screen through yet still provide a scratch-proof and replaceable fixing method. The design of the band, firmware and software tools were released as *Open Hardware* under the Openmovement platform [Openmovement, 2013].

### 7.3.2 *Climb Segmentation*

Our vision of a climbing analysis system comprises an accessory for assessing climbing activities in a naturalistic setting, i.e., not imposing any additional constraints or requirements that would hinder the core exercise. In line with this, *ClimbAX* detects climbing activities, which alleviates its user from the necessity of interacting with the device, e.g., clicking a button before, during or after each climb.

During every-day activities, arm based movements are subject to what is commonly referred to as *symmetry-bias* [Treffner and Turvey, 1996]. Motions by, e.g., one arm automatically initiate a counter movement by the other arm to keep balance. This symmetry is often used to characterise gait, particularly for neuro-degenerative conditions [Yogev et al., 2007]. During climbing this symmetry between the upper extremities is broken as it is crucial for one limb to stay attached to the hold, minimising its movement. Along with tremors related to high intensity activities (vibrations of the hands when on holds caused by fatigue or extreme exertion) this gives rise to specific climbing patterns as they are recorded on the wrists. Our automatic climb detection is based on the analysis of these characteristic movement patterns, which we found are more discriminative than simple assessments of simultaneous upwards wrist orientation with respect to gravity.

Detecting episodes of climbing within continuous streams of accelerometry data corresponds to segmentation of time series data, for which two general processing paradigms exist: *i)* explicit identification of start- and end-points of semantically contiguous bouts (segments); and *ii)* implicit segmentation through extraction of analysis frames and subsequent, isolated classification regarding the patterns of interest [Keogh et al., 2004]. Ambiguity in transitions between non-climbing and climbing activities effectively renders explicit segmentation techniques impractical for climb detection. However, the

aforementioned break of symmetry-bias during climbing results in substantially different sensor data distributions for climbing and non-climbing episodes. Exploiting this, we employ an implicit segmentation approach for climb detection using a sliding window procedure that extracts analysis frames thereby integrating sensor data from both wrists (see chapter 2.4.2).

Our sliding window procedure extracts frames of 5s length with an overlap of 1s, which captures climbing activities very effectively. For analysing symmetry-biases (and breaks therein) we concatenate the tri-axial sensor readings of both wrists into a unified representation. For these frames we then calculate feature vectors that represent the characteristics of the performed activities in a compact way. We employ a feature learning approach based on Restricted Bolzman Machines (RBM) , which has been demonstrated as being very effective for activity recognition tasks in chapter 4.3.3. Following the original approach, we employ 900 hidden units to match the input dimensionality (see below). For our climb detection procedure we down-sample the accelerometer data to 30Hz. Cross-validation experiments suggest that this has no adverse effect on the overall effectiveness while at the same time greatly alleviating requirements on the sample sets required for robust RBM training. Preliminary experiments indicated that training a single RBM suffices to reliable estimate a feature representation that allows robust detection of climbing activity.

Feature vectors are then fed into a statistical classification system that discriminates climbing from non-climbing on a per-frame basis. We have evaluated a number of classification approaches and found that logistic regression works best for climb detection. Finally, the sequences of predicted activity labels undergo temporal smoothing for outlier elimination, resulting in effective segmentation. Figure 7.4 summarises the climb detection procedure.

### 7.3.3 *Move Segmentation*

Even the most complex climbing activities essentially consist of sequences of atomic movement units. These *moves* are defined as:

> *Continuous limb movements that are temporally surrounded by pauses, i.e., static episodes with no significant displacement of the particular limb of interest.*

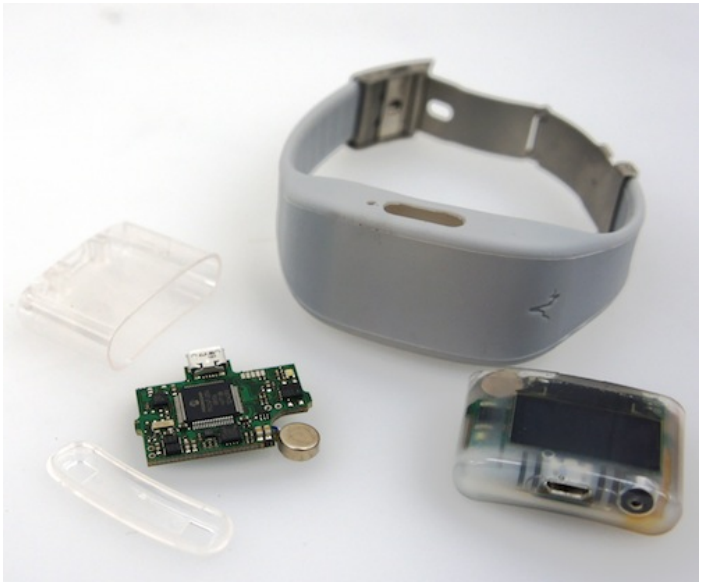


FIGURE 7.3: Wearable sensing platform for recording climbing activities that consists of a high-resolution tri-axial accelerometer, OLED screen (not used), on-board processing unit (PIC), battery, and flash memory.

Consequently, quality analysis of climbing performance is typically based on assessments of individual moves.

*ClimbAX* follows the general approach of move-based analysis. After climbing sessions have been detected (cf. previous section), we segment moves on a per-limb basis, which is important for the generation of detailed assessment information. Although the aforementioned definition of moves suggests a straightforward implementation through detecting smooth sensor displacement trajectories, there are two things to consider when assessing real-world climbing activities: *i)* Typically moves between holds require hand adjustments to reach a stable and comfortable position. Such adjustments add "jitter" to the beginning and end of the actual reaching movements; *ii)* Moves can also correspond to the turning of the hand on a single hold without any reaching movement involved, e.g., for repositioning to rest more comfortably or to prepare for the next (reaching) move.

Taking these considerations into account our move detection focuses on segmenting hands being on holds, which is characterised by low energy values of the acceleration signals, interrupted by temporally short high energy episodes. The latter involves a hand moving to a hold, it's adjustment and other climbing activities such as clipping the rope (e.g., for sport climbing; see ii in Figure7.1(b)). We calculate short-term energies on raw acceleration data using the same sliding window procedure that has been applied

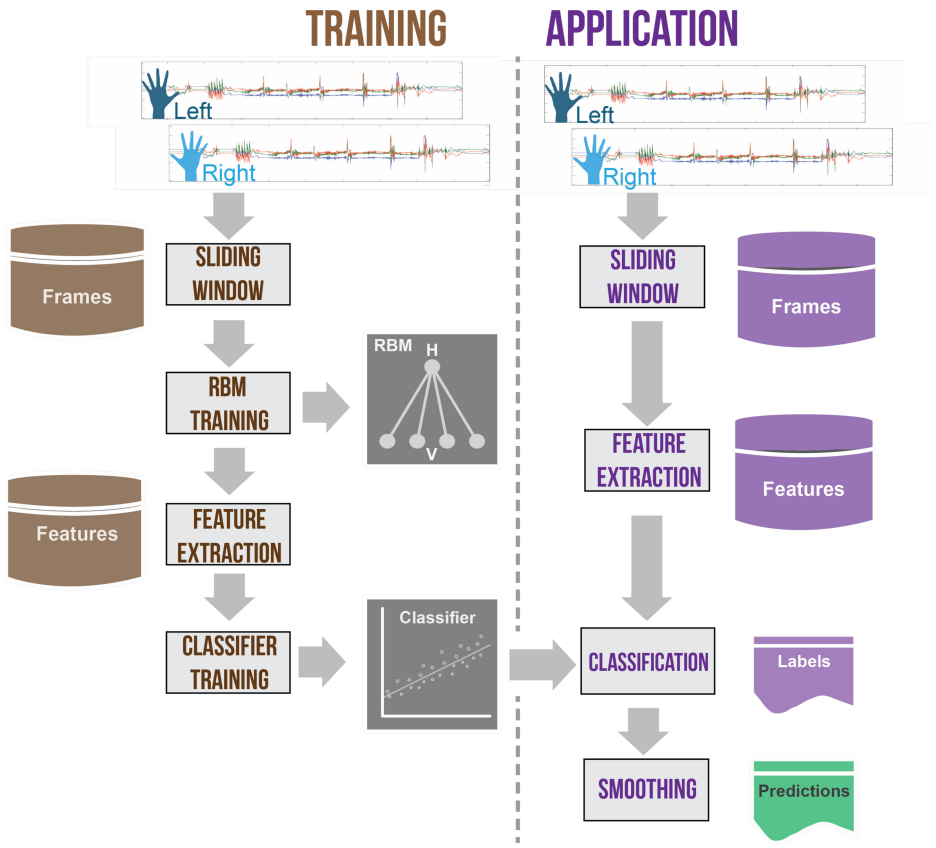



FIGURE 7.4: Overview of climb detection procedure: Sliding window frame extraction and feature learning (using Restricted Bolzman Machine – RBM) for capturing characteristics of movement patterns, which are classified using statistical classification backend.

for climb detection (previous section). Algorithm 1 summarises the move detection procedure.

### 7.3.4 *Assessment*

Quality assessment of climbing as it is performed by professional coaches is —across its sub-disciplines (Figure 7.1(b))— based on a move-specific analysis of certain key criteria that characterise a set of commonly accepted core skills every climber needs to possess and develop [Pansiot et al., 2008]:

**Power** – the ability to transfer isometric strength into a move. Holds that are further apart will require a climber to be more powerful to transition between them.

**Control** – the ability to transition smoothly between holds. Often a climber is required to shift their centre of mass to enable a hold transition to be made, which requires

**Algorithm 2** Automatic Move Detection in *ClimbAX*

**Input:** limb index $l$ (left or right); energy threshold $t$; climb segment $\mathbf{s}$, frame length $F$

**Output:** moves $\mathscr{M} = \{\mathbf{m}_i(\mathbf{s})\}$ for given climb segment $\mathbf{s}$

**procedure** DETECTMOVES($\mathbf{s}$, $l$, $t$)

$\mathscr{M} = \emptyset$ ▷ Initialise moves set

**for all** Frames $\{\mathbf{f}\}$ **do** ▷ Sliding window procedure

Calculate short term energy:

$$E_{\mathbf{f}} = \left(\sum_{i=1}^{F} \mathbf{f}_x(i)^2 + \mathbf{f}_y(i)^2 + \mathbf{f}_z(i)^2\right)^{-1/2}$$

**if** $E_{\mathbf{f}} > t$ **then** ▷ Energy thresholding

$\mathscr{M} = \mathscr{M} \bigcup \mathbf{f}$

**else**

continue

**end if**

**end for**

Perform median smoothing ▷ Outlier elimination

**end procedure**

both core body strength and balance. Poor control will result in jerky limb movements whereas good control corresponds to smooth transitions between stances.

**Stability** – the ability to remain composed while holding onto holds. Small or sloping holds are difficult to grip typically resulting in poor stability as postural or finger repositions are required to maintain a stance on a hold.

**Speed** – defined as timing observation. In most cases, completing a climb in the shortest possible time is desirable.

We aim for replicating expert assessments by measuring the aforementioned core skills in the climbing episodes extracted from the sensor signals.

Intuitively, the power of a climber corresponds to the peak (physical) work they can perform over time, which has been used to assess a climber's arm power in a well controlled experiment in [Draper et al., 2011]. This immediate measure of the arm's displacement over time, however, fails to capture the context of the move performed, i.e. the perceived quality of the holds and footrests involved in a climbing sequence. This context is nevertheless crucial to gain insights about a climber's abilities. Even a climber with very little power will be able to perform a long reaching and quick move from good

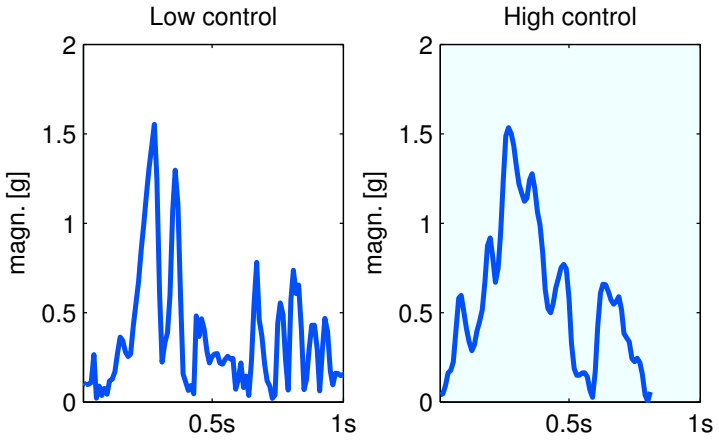



FIGURE 7.5: Two example moves demonstrating low and high control. The left plot shows the motion magnitude (gravity removed) from a climber with a low score for control. The right plot shows the same move by the climbing with the highest estimated control.

quality holds, while the same climber will struggle with small holds that are difficult to grab.

A low quality hold induces high intensity tremors as much strength is required to pull or hang from it. The signal captured from this hand will therefore exhibit a higher signal energy compared to a good quality hold. In order to assess power $P$ for a climb that involves $i$ moves, we measure the relationship between the signal energy $E_m^i$ of the moving hand to the signal energy $E_h^i$ of the hand residing on a hold during a move:

$$p_i = \frac{E_m^i}{E_h^i} \tag{7.1}$$

$$P = \max(\{p_i\}) \tag{7.2}$$

Coaching guides describe *Control* as the smoothness of hand movements during hold transitions [Fyffe and Peter, 1997, Horst, 2008]. Intuitively a climber that shows good control has a great level of coordination, good timing and moves efficiently between holds. A controlled hold transition corresponds to a smooth movement of the hand, without hesitation, that precisely reaches the optimal hand position on the target hold. Poor control often results in *over-shooting* beyond the hold, hitting the wall during the transition, and high impact forces on the target hold due to imprecision (see Figure 7.5).

Control $C$ can thus be characterised as the ratio of energy in short bursts (impacts) against energy in the long run (smooth motion) captured from the moving arm.

$$c_i = \max\left(\left\{\frac{e_t^s}{e_t^l}\right\}_{t\in T_i}\right) \tag{7.3}$$

$$C = \text{mean}\,(c_i)\,, \tag{7.4}$$

where $c_i$ is the control of move $i$ (over time $T_i$), while $e_t^s$, and $e_t^l$ are short-term signal energies calculated using a sliding window with length $t_s$ and $t_l$ respectively ($t_l \gg t_s$).

*Stability* in climbing is a measure for how well attached the hands remain to the hold while not engaged in a hold transition. Poor stability, i.e., unnecessary movements of the hand while on a hold, is most commonly caused by a combination of poor flexibility and core body strength. These unnecessary movements usually correspond to sharp changes in acceleration when, e.g., the hand position on the hold is adjusted. Stability $S$ for a climb is therefore inversely proportional to the variance of the first derivative of motion magnitude (*jerk*) while the hand is not moving:

$$S = \text{std}\left(\frac{\partial\, m_h}{\partial\, t}\right)^{-1}, \tag{7.5}$$

where $m_h$ is the motion magnitude of each hand on hold.

Coaches use the *Speed* of a climber to asses both their route reading ability as well as their fatigue. While there are many ways to define speed (e.g., time taken to ascend a route, or time between limb movements) we chose to measure speed $V$ as the number of moves per second. This methods is thus insensitive to route length and can be directly derived from the climb and move segmentation outputs.

*ClimbAX* calculates estimates of these core skills for every detected move and combines the values into a 4-dimensional skill representation $\mathbf{s} = \{P, C, S, V\} \in \mathbb{R}^4$. In doing so we effectively translate continuous accelerometry data collected by the sensing platform worn on both hands of the climber into sequences of core skill values, which is the basis for both individual and comparative assessment, as well as for progress tracking – all at a great level of detail, which replicates and translates to best practice in existing manual assessment.

## 7.4 Experimental evaluation

### 7.4.1 *Datasets*

Two different datasets were collected in order to evaluate: *i)* the climb segmentation; *ii)* the move segmentation; and *iii)* the automated skill assessment.

The first dataset (*sport climbing*) consists of a total of 42 climbs recorded from 6 participants at two different indoor climbing walls (i, and iii in Figure 7.7). Participants were asked to wear a set of sensors for the duration of their visit to the climbing wall and to go about their regular climbing activities without any specified protocol. After their climbing session participants were asked to produce a diary containing the exact start and end times of each climb. A climb here is defined as the moment the subject starts climbing until they are back on the ground, i.e., it may contain resting and falls. Crucially the data recorded is not limited to climbing activities but contains other activities such as belaying, walking around, resting, etc.

The second dataset (*competition*) was collected during a local bouldering competition, where a total of 47 subjects performed a single climbing problem, which was part of the official competition set (purple holds in Figure 7.6). The route was set up with the particular needs of a performance evaluation in mind. Care was taken so that it contains moves that require both control and power, without favouring one particular skill set or side of the body. Participants were recruited among all competitors with no particular preference, resulting in a representative sample of the audience for such competitions. Based on video recordings the recorded data was annotated for climbs and the exact sequence of moves performed by each participant. In addition to the recordings, the competition results for the majority of the participants were also collected. Both datasets are summarised in Table 7.1.

| Dataset | Participants | Climbs | Moves | Scores |
|---|---|---|---|---|
| Sport climbing | 6 | 42 | – | – |
| Competition | 47 | 47 | 770 | 40 |
| Total | 53 | 89 | 770 | 40 |

TABLE 7.1: Summary of (annotated) data collected in 2 different studies.


FIGURE 7.6: A subject on the route climbed by all participants in the *competition* dataset (purple holds). Increasing numbers indicate the intended sequence of holds (*h* for hands and *f* for feet) although some variation in solutions was observed.

### 7.4.2 *Segmentation of climbing episodes*

In order to evaluate the performance of the climb detection described in this work a 10-fold cross validation was performed on the combination of both the *sport climbing* and the *competition* dataset. For each dataset, frames of 5-second length are extracted with a shift of 1 second. The identity of each frame is decided based on a majority vote based on the ground truth annotations. This set of frames is then split into 10 partitions, each containing a continuous segment of the data (with respect to time), which is retained throughout all experiments. An RBM with Gaussian visible units and binary hidden units is trained for 250 epochs for each fold (see chapter 4.3.3). For each frame, the activation probabilities of the hidden units are retained as feature representation.

Three different classifiers were trained based on the features extracted by the RBM:


FIGURE 7.7: *ClimbAX*: The locations used for data collection: i) Indoor sports climbing wall. ii) Indoor bouldering wall under competition settings. iii) Indoor sports climbing wall with large overhang.

*i)* $k$-nearest neighbour ($k = 1$); *ii)* decision trees (c4.5); and *iii)* standard logistic regression. Results are reported in Table 7.2. After obtaining the results for each frame independently it is straight-forward to apply temporal smoothing based on a window of $n$ samples and a hamming window. Using a simple threshold to detect a climbing episode heavily improves the recognition results. Figure 7.8 illustrates ROC curves for the different classifiers after temporal smoothing is applied (based on a 50-sample window). Logistic regression on the raw, 900-dimensional feature representation clearly outperforms all other classifiers investigated.

Table 7.3 illustrates the best segmentation results for the different datasets using logistic regression. Overall the results improve dramatically if temporal smoothing is employed with a precision of 0.87 and a recall of 0.87. The results for the *Sport Climbing* dataset are particularly interesting as they include plenty of activities unrelated to climbing. This dataset was captured during typical visits to a climbing centre and includes activities such was warming up, stretching, drinking coffee, and walking among others. Some activities that are similar to climbing activity are included as well, such as rope handling and belaying. Overall climbing constitutes just 17% of this set, yet it can still be detected very reliably with a specificity of 0.96.

| Method | Precision | Recall | Specificity |
|---|---|---|---|
| c45* | 0.43 | 0.64 | 0.81 |
| knn* | 0.66 | 0.78 | 0.91 |
| logR | 0.79 | 0.71 | 0.96 |
| PCA+logR* | 0.80 | 0.66 | 0.96 |

TABLE 7.2: Performance of climb detection using different classifiers on raw prediction results (no temporal smoothing).

| Dataset | Precision | Recall | Specificity |
|---|---|---|---|
| Sport climbing | 0.85 | 0.88 | 0.96 |
| Competition | 0.88 | 0.86 | 0.98 |
| Overall | 0.87 | 0.87 | 0.97 |

TABLE 7.3: Performance of climb detection using 'logR' after temporal smoothing. The *Sport Climbing* dataset contains approx. 17% climbing activity along with different activities typical for a visit to a climbing centre.

### 7.4.3 *Segmentation of moves*

Based on the extracted climbing episodes from the *competition* dataset we apply the process described in this work to extract moves, separately for each limb. Each move is treated as an event, and is deemed detected if it overlaps with an automatically extracted move. Overall this results in a precision of 0.79 and a recall of 0.82. The imprecision of the method is largely due to the boundaries extracted by the climb detection, which may exclude moves at the very start and end of a climb. These boundary conditions have a significant impact on the performance figures since the short climbing sequences in this dataset just contain approx. 10 individual moves per hand (see Figure 7.6). However, our results indicate that the extracted moves still adequately reflect the climbers' overall skill.

### 7.4.4 *Assessment parameter evaluation*

Based on the extracted climbing episodes along with their segmented moves, one set of performance attributes (power, stability, control and speed) is estimated for each climber using the process described above. The competition scores recorded in the *competition*

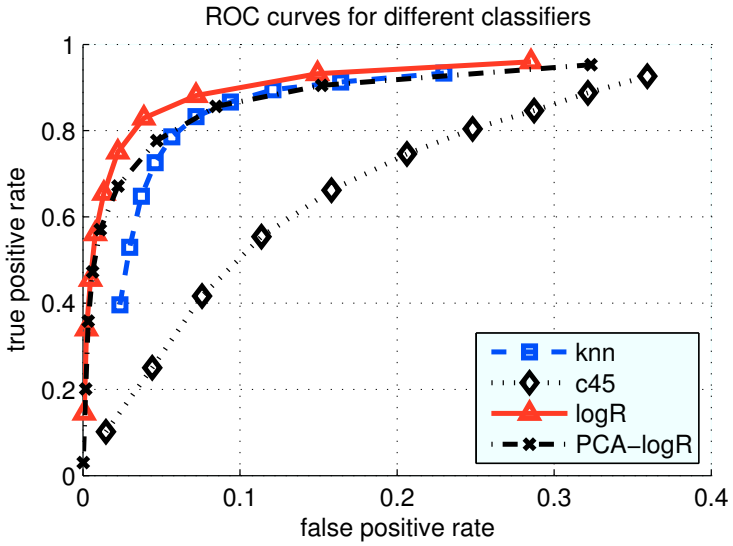



FIGURE 7.8: ROC curves of different classifiers for climb detection after temporal smoothing. Logistic Regression on the raw features ('logR') clearly outperforms other classifiers. Its performance remains comparable to KNN if the dimensionality of the features is reduced using PCA to 100 dimensions.

dataset effectively correspond to an objective, unbiased estimate of a participants climbing ability. Out of the 47 participants, 40 handed in a scoring sheet, which provide the basis for the evaluation of a simple linear model. In this experiment, a linear regression is fitted in a leave-one-climber-out cross-validation and used to predict competition scores based on the performance attributes.

The scatter plot in Figure 7.9 illustrates the results. The predicted scores show an overall positive correlation to the recorded competition scores of 0.74, indicating that our performance attributes are suitable to capture some elements of climbing skill. This is an extremely encouraging result, as the performance of a climber during a competition is influenced by many things a body-worn sensing system is incapable of measuring (such as mood, form, etc.). Furthermore, since just a single climb is observed from each participant, long term characteristics such as (power) endurance and tiredness can not be observed.

Another parameter that has strong implications on climbing style is that of body-weight. Remaining on the wall, even on very difficult and small holds, requires less strength for a very lean climber. We believe that the route we set for this experiment favoured lean climbers with a transition on a difficult hold (hold *h9* in Figure 7.6). Inspired by this insight we performed an additional experiment in which climbers with a body-mass index (bmi) of less than 20 are removed from the set. Following the same approach as for the last experiment, the performance improves significantly with an overall correlation

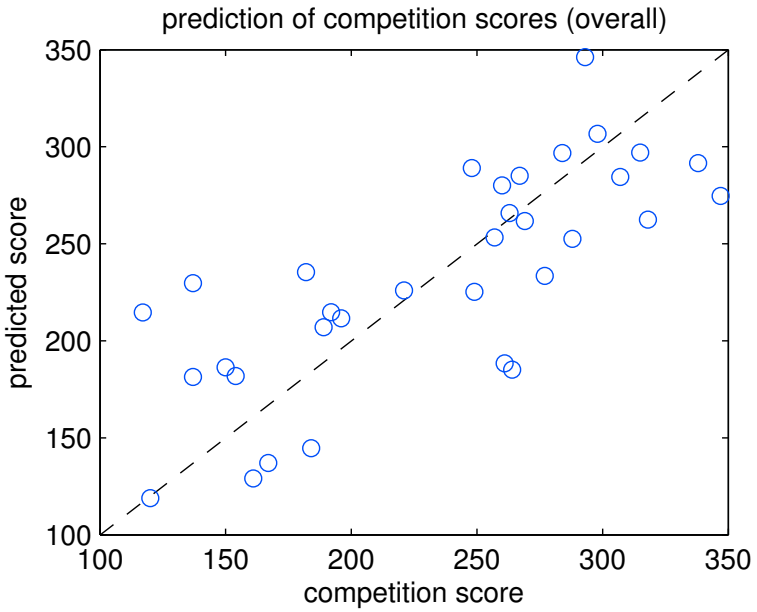



FIGURE 7.9: Scatter plot of climbers' performance in the competition, illustrating the correlation between predicted scores and the ground truth (0.76). The estimated performance parameters of each climb $s \in \mathbb{R}^4$ are used to train a linear model in a leave-one-out cross-validation.

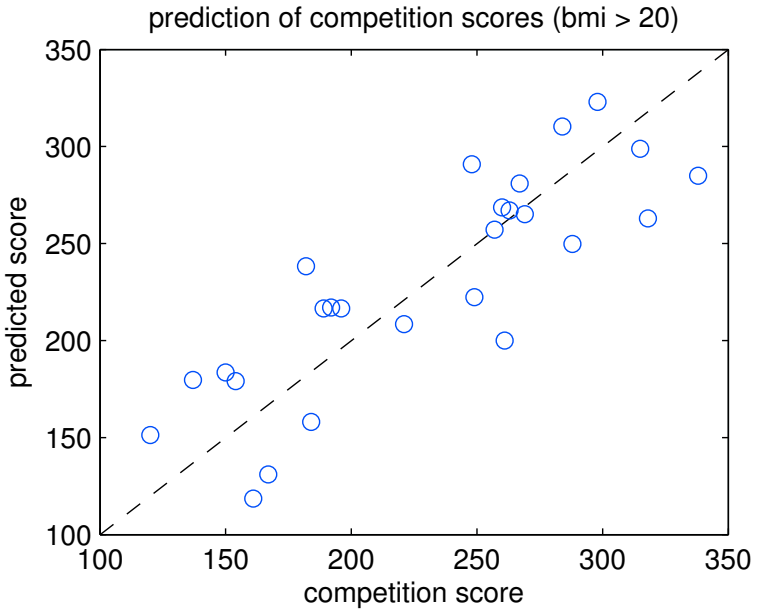



FIGURE 7.10: Prediction performance when climbers with a *bmi* of less than 20 are removed from the set. The prediction shows a correlation to ground truth of 0.84.

of 0.84. A scatter plot illustrating the performance of this reduced set is illustrated in Figure 7.10.

## 7.5 Related Work

Current best practice for the assessment of climbing activities corresponds to manual observation and judgment, typically performed by an experienced coach. While such expert assessments work well for elite climbers, practical resource limitations prevent generalisation to the large number of amateur climbers. The desire for automated climbing assessment served as the motivation for the development of the *ClimbAX* system presented in this paper.

Monitoring general sports activities using ubiquitous computing technology has become very popular in the recent past, as discussed in chapter 2.3.1. The proliferation of inexpensive, miniaturised sensing hardware together with the availability of sufficient computational power in mobile devices has lead to a wealth of applications [Andre and Wolf, 2007]. Apart from logging sports activities a few systems have also focused on assessments of their qualities. To name but a few examples, Fothergill *et al.* developed an automatic coaching system for rowers [Fothergill et al., 2008], Ahmadi and colleagues explored the use of wearable computing for skill assessment in tennis [Ahmadi et al., 2010], Möller *et al.* described skill assessment in fitness exercises using a mobile phone [Möller et al., 2012], and Grober instrumented a golf club with accelerometers to analyse the quality of golf swings [Grober, 2009].

Hardly any approaches have so far been published that are related to the automatic analysis of climbing activities. Notable exceptions are the exploration of body-attached sensors as a means for movement analysis in rock climbers [Schmid et al., 2007], and the use of ear-mounted accelerometers for climber performance monitoring [Pansiot et al., 2008]. However, both studies have either focused on the exploration of the general feasibility of wearable climbing assessment, or targeted very specific aspects of climbing activities. In contrast, our work goes much further by developing a complete framework for generic skill assessment in climbing activities.

Activity recognition underlying the presented climbing assessment is closely related to gesture recognition using wearable computing techniques, which is one of the major research fields within the ubiquitous and wearable computing community [Preece et al., 2009] (see chapter 2.3.2). A large variety of applications has been explored, ranging from analysing activities of daily living, health-related aspects, or work-related activities [Atallah and Yang, 2009a, Ward et al., 2006]. A wealth of analysis techniques have been

employed, whereas the majority of them focus on discriminating the activities of interest rather than assessing their quality.

## 7.6 Discussion

Climbing has become very popular and is now being enjoyed by a large population who value it as a sociable leisure activity that combines physical activities with outdoor experiences in a unique way. Similar to other sports, climbing requires physical fitness and coordination, and progression can only be achieved through repetitive and dedicated practicing. Elite climbers reach (and maintain) their expertise with the support of individualised coaching. Such coaching specifically targets the improvement of individual weaknesses that are identified by experts who continuously analyse their performance. Unfortunately, such expert coaching and performance assessment is not available for most climbers at the amateur level. As a consequence and especially in the light of the complexity of climbing, many amateurs lose motivation by not making enough progress in developing their skills or even put their health on jeopardy through inappropriate or dangerous climbing.

We have embarked on developing an automatic assessment system that analyses the quality of climbing – *ClimbAX*. Ultimately such a system represents an important building block for a digital, personal climbing coach that replicates individualised expert assessment of climbing skills as it is currently conducted by human coaches. In this paper we presented a body-worn sensing system and explored analysis techniques that effectively segment and quantify measures relating to climbing ability. With the assistance of coaches and sport science literature, four core parameters were designed that are relevant for climbing skills: power, control, stability and speed.

We have demonstrated that an automatic analysis approach based on the combined evaluation of aforementioned core climbing skills correlates to scores achieved under competition conditions. This comparison is, however, limited when used for either very good climbers or absolute beginners. In the case of beginners, not enough data was captured as often the climber fell from the route in the first few moves. In the case of very the elite climbers, the route was not significantly hard enough to test their ability. Our results indicate that climbers with lean body-shape were favoured by the route set for our experiments with much improved results upon their removal from the assessment.

While our results are encouraging, they are just based on a single climb per participant. Crucial aspects such as endurance (defined as resilience to fatigue) are inaccessible to the system and a considerable amount of work necessary until am automatic, personal climbing coach becomes reality.

## 7.7 Future Work

This work explores the automatic assessment of climbing ability, with the aim to provide a basis for a (semi-) automated, personalised coaching system. However, the transition from raw performance attributes towards individualised training recommendations is not explored. Of particular interest here is to investigate if automated training recommendations are beneficial for a climber's progression and how this benefit compares to that of a dedicated professional coach, which will be explored in future studies.

## 7.8 Implications for activity recognition in naturalistic surroundings

The system presented in this work is based around automatic feature learning, namely deep belief networks, as described in chapter 4. This allows the automatic inference of features based on large amounts of naturalistic data from realistic climbing activity. The resulting features prove to be effective when applied to a second data-set that is collected in a competition scenario where participants where recorded on video.

### 7.8.1 *Combination of naturalistic and scripted data collection*

The system presented in this chapter effectively utilised data captured from naturalistic surroundings in developing the climbing detection algorithm based on automatic feature learning. The training data for this component of the system was collected in a climbing hall under real-life conditions, including activities unrelated to climbing but typical for such an environment, such as belaying, warming up, or having a break. Crucially that activity was not captured on video, effectively minimising the impact of the act of measurement on the behaviour of the participants.

The automated skill assessment was developed based on data from a real-life climbing competition, even though we were able to influence the design of the climbing route the participants climbed during study setup. The basis of the evaluation of this skill assessment are the results from a realistic competition, and not the subjective impression to an expert watching e.g. video recordings. On one hand this leads to a difficult scenario, as just a short episode of climbing activity is captured which may not be representative of the performance during the rest of the competition (e.g. due to fatigue). On the other hand it provides a much more realistic sense of the system's performance, as such issues would also arise if the system is deployed in a real-life climbing environment.

Effectively this corresponds to a combination of a naturalistic and a scripted or semi-natural data collection to develop and evaluate the activity recognition system. This study setup has two advantages:

**Ease of data collection** Each individual data-set is relatively straight-forward to collect. The naturalistic data requires minimal labelling and no video capture at all, which would be difficult and costly to obtain as climbing halls are a challenging environment for such work (lighting conditions, camera placement, dust, etc.), yet it is an adequate reflection activities in climbing halls. Capturing the second (competition) data-set is more complicated, as video recordings have to be obtained from the participants. Crucially however this data-set is not required to deliver background-activities beyond the actual climbing, as this is already captured in the naturalistic set. This means that no further (scripted) activities following some study protocol are necessary.

**Realistic impression of performance** As large parts of the system are based on naturalistic data it is likely that they will retain much of their performance in realistic, practical deployments. Additionally the skill assessment is evaluated in a realistic scenario that is typical for the envisioned use-case of the system. It would be difficult to obtain both results with a single study.

### 7.8.2 *Summary*

By combining naturalistic with scripted or semi-naturalistic data collection we effectively gain the best of both approaches. This approach allows us to demonstrate the reliability of basic parts of the system while the detailed annotation (or assessments)

from a constrained setting allow detailed development and evaluation of higher level movement analysis components.

The approach is not just suitable for applications in sports and extends to clinical applications. In clinical settings, detailed evaluation, where automated assessments are compared to that of experts, it is crucial to provide sufficient prove for the suitability of a specific technical approach. The same application would benefit e.g. populations affected by degenerative conditions if they work reliably under real-world conditions. The following chapter 8 describes a prototype system that employs this methodology to develop an automated assessment system for the disease state in Parkinson's Disease.

# Chapter 8. Assessing Disease State in Parkinson's Disease in Naturalistic Surroundings

Management of Parkinson's Disease (PD) could be improved significantly if reliable, objective information about fluctuations in disease severity could be obtained in ecologically valid surroundings such as the private home [Maetzler et al., 2013]. Although automatic assessment in PD has been studied extensively, so far no approach has been devised that is useful for clinical practice. Analysis approaches common for the field lack the capability of exploiting data from realistic environments, which represents a major barrier to practical assessment systems. The very unreliable and infrequent labelling of ambiguous, low resolution movement data collected in such environments represents a very challenging analysis setting, where advances would have significant societal impact in our ageing population. In this chapter we propose an assessment system that abides practical usability constraints (see chapter 2.2) and applies deep learning to differentiate disease state in data collected in naturalistic settings (see chapter 4).

This chapter combines the technical insights presented throughout this thesis and follows the recommendations for study design highlighted in chapter 7. Data collection in this chapter is split into two phases: i) a naturalistic setting (at home) where annotations are recorded in collaboration with the participants through the use of diaries; and ii) a semi-naturalistic setting where participants spend time at a laboratory setting where their disease state is assessed by an expert in regular intervals. We illustrate how deep learning based on the ECDF feature representation outperforms other, more traditional methods in this setting, an approach which could be applied to similar settings for other degenerative conditions.

## 8.1 Introduction

Parkinson's Disease (PD) is a degenerative disorder of the central nervous system that affects around 1% of people over 60 in industrialised countries [de Lau and Breteler, 2006]. People affected by PD show a variety of motor features that gain in severity with the progression of the disease, which include rigidity, slowness of motion, shaking and problems with gait [among others]. The severity and nature of these motor features vary over the course of the day, which has a significant impact on the quality of life of people with PD. Management of the condition relies on tailored treatment plans that provide a specific schedule for the type and dosage of a multitude of medications taken by each individual. Devising such treatment plans is a challenge as clinical consultations may be infrequent and only provide a snapshot of the condition, which may not give an adequate picture of the daily fluctuations beyond recall by the individual. Objective, automated means to assess PD in people's daily lives are therefore much desired.

In order to become a useful clinical tool, such automated assessment systems have to be deployed in naturalistic, ecologically valid surroundings such as the private home. Systems based on, or evaluated in, such naturalistic settings are, however, very rare. The reason for this apparent shortcoming is clear: while capturing data in naturalistic environments is straight-forward using e.g. body-worn movement sensors, obtaining reliable labels useful for system development is practically difficult, if not impossible, as even trained annotators would show only modest agreement with experts [Palmer et al., 2010]. In practice only unreliable and infrequent labels can be obtained in such surroundings, for example using symptom diaries kept by each participant. Instead of addressing this issue, current systems for the assessment of PD rely on data captured in the laboratory, where daily life is just simulated [Hoff et al., 2001], or attempt to recreate the laboratory in the private home using e.g. movement tasks under remote supervision by a clinician [Giuffrida et al., 2009]. Research on PD is missing adequate tools that would allow data from ecologically valid surroundings to be exploited in system development, as this problem with its unique challenges, has received little attention from the machine learning community. This represents a significant barrier for practical assessment systems, overcoming which may dramatically improve the quality of life of people affected by PD.

In this chapter we investigate the problem of predicting the disease state in PD patients in naturalistic surroundings, i.e. the daily life of individuals affected by PD. We illustrate that assessment systems have to overcome significant challenges in analysing

large quantities of ambiguous, low-resolution multi-variate time-series data for which only infrequent and unreliable labels can be obtained due to practical usability constraints. Labels are subject to various sources of noise such as recall bias, class confusion and boundary issues and do not capture the main source of variance in the data, as people engage in many (unknown) physical activities that have significant effect on the recorded sensor signals. Based on a large data-set that contains approx. 5,500 hours of movement data collected from 34 participants in realistic, naturalistic settings, we compare how deep learning and other methods are able to cope with the characteristic label noise. We find that deep learning significantly outperforms other approaches in generalisation performance, despite the unreliability of the labelling in the training set. We show how such systems could improve clinical practice and argue that such a setting could serve as a novel test-bed for unsupervised or semi-supervised learning, where improvements would have significant societal impact.

## 8.2 Assessing disease state in naturalistic surroundings

The quality of life of people with PD is significantly affected by fluctuations in the severity of the disease. Periods where motor symptoms (such as tremor or bradykinesia) are more prominent are typically referred to by clinicians and patients as "*off* time". Conversely, periods where motor symptoms are well controlled are referred to as "*on* time". As the condition progresses, motor fluctuations between these differing *disease states* become more frequent and less predictable. Furthermore, prolonged medication usage is associated with the development of additional involuntary movements known as *dyskinesia*. Tailored treatment plans aim to reduce the severity of these fluctuations. In this chapter we focus on the assessment of disease state in PD, as it represents a crucial component for improved management of the condition in clinical practice.

In order to be useful for clinical practice, assessment systems have to be applicable in naturalistic, ecologically valid surroundings such as the private home. Yet research on automated assessment in PD generally does not address this issue. Systems are based on laboratory environments, where participants engage in a series of movement tasks that are part of the clinical assessment procedure in PD [Goetz et al., 2008, Patel et al., 2009]. With extensive instrumentation of the participants those system achieve very good results in e.g. detecting dyskinesia with more than 90% accuracy [Tsipouras et al., 2010]. However, such scripted movement tasks, even if extended to include activities of

daily living to simulate daily live [Hoff et al., 2001], are only a very poor model of naturalistic behaviour. Some systems aim to re-create the clinical assessment in the daily life of a subject while being supervised by a clinician (remotely), effectively simulating such laboratory conditions in naturalistic surroundings [Mera et al., 2012, Giuffrida et al., 2009]. Whether individual assessment systems generalise to naturalistic environments is rarely explored, where a recent review just found 3 out of 36 studies to include data recorded in naturalistic settings, although the authors specifically focussed on this aspect [Maetzler et al., 2013]. Even where naturalistic data is gathered it is not utilised during system development, but instead being used to gain some insight into the performance of systems based on medical prior knowledge (e.g. [Griffiths et al., 2012, Hoff et al., 2004]). This laboratory-driven research represents a significant barrier towards practical assessment systems.

The reliance on controlled laboratory environments stems from the difficulties encountered when collecting and exploiting data from in naturalistic surroundings. This issue is split into two aspects. The most pressing concern from a machine learning perspective relates to obtaining ground-truth information about disease state in PD in naturalistic environments. Even if e.g. video recordings can be obtained, which is unlikely, it can be difficult for annotators to assess the disease state with high reliability [Palmer et al., 2010]. Instead, labels have to be obtained in cooperation with the patients, where the common best practice are disease state diaries [Reimer et al., 2004]. Such diaries just provide an infrequent (e.g. one sample per hour) and unreliable impression of the disease state. Participants may have trouble identifying their own disease state, or fill out the diary retrospectively (recall bias). Additionally, the disease characteristics evolve gradually and are unlikely to change exactly on the hour, leading to issues at the boundaries of the provided labelling.

The second aspect relates to usability aspects of the sensing system. As discussed in chapter 2.2.4, recording (unlabelled) data in naturalistic settings is straight-forward if sensing solutions require little cooperation by the patient, do not rely on external infrastructure, abide by privacy constraints, and follow a suitable physical design of the devices [McNaney et al., 2011]. Any practical sensing system will necessarily be a compromise between the obtainable sensing resolution (e.g. degrees of freedom, number of sensors, ambiguity of recordings) and abiding usability constraints of the target population to maximise compliance. The most suitable sensing approach for naturalistic deployments are small body-worn movement sensors [Maetzler et al., 2013], which

capture multi-variate time-series data that give an impression of the participant's physical activity and overall behaviour with a large amount of noise and inherent ambiguity. The main sources of variance in this movement data are the physical activities that participants engage in, such as walking, and not the overall disease state. The disease state rather has an effect on how activities are performed (e.g. "slower" while *off*). However, the activities that participants engage in are unknown, as collecting additional activity logs to gain an impression of the physical activities of participants would be too burdensome for longitudinal settings, particularly if participants suffer from cognitive decline. This also renders approaches such as *active learning* (e.g. [Stikic et al., 2008b]) difficult to apply for this population, as they also require significant cooperation by the individual.

We can summarise the challenges for exploiting naturalistic data in this setting as follows: *i)* There is a significant disparity between the frequency at which data is collected (e.g. 100Hz) and the accessible labelling (e.g. one per hour); *ii)* The participant-provided labelling is inherently unreliable, subject to recall bias, class confusion and boundary issues; *iii)* The recorded data mostly reflects unknown activities, across which an assessment system has to generalise to obtain an impression of disease state in PD. Addressing these challenges through methodological advances would have significant impact on clinical practice for PD and other degenerative conditions where assessment faces similar issues.

## 8.3 System overview

In response to these challenges we develop a novel approach to the assessment of disease state in PD. Instead of basing the development of our approach on data collected in a laboratory setting, we exploit large amounts of data gathered in naturalistic surroundings, the daily life of people affected by PD. In many applications of machine learning it is easy to obtain large amounts of unlabelled data, and devising systems capable of exploiting such data to improve recognition performance has become a popular field in machine learning. One approach that has been shown to be effective for e.g. phoneme recognition [Deng et al., 2013] and object recognition [Lee et al., 2009] is *deep learning*, where unlabelled data is used to greedily initialise multiple layers of feature extractors (see chapter 4). In this chapter we apply deep learning to the problem of disease state assessment in PD to explore if these methods can cope with the unreliable labelling that results from naturalistic recording environments.

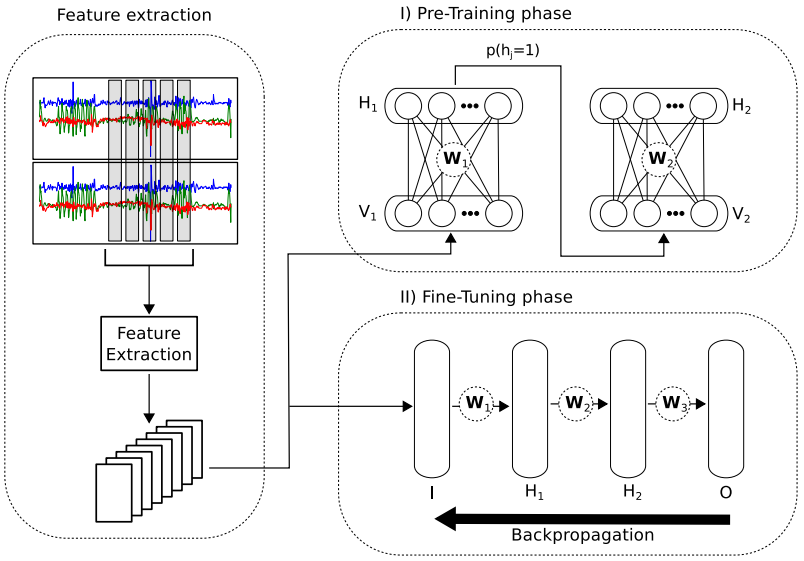



FIGURE 8.1: Overview of proposed feature learning procedure for disease state prediction in Parkinson's Disease. First, the data is split into non-overlapping segments of 1 minute duration. Feature are extracted from each of the segments, and $d$ segments are concatenated to form one sample in the training-set. Subsequently, a series of restricted boltzmann machines (RBMs) is utilised to estimate weight matrices $W_1$ and $W_2$. Those weights are subject to fine-tuning using the (unreliable) labels from the phase 2 data-set using conjugate gradients.

Our system comprises a typical analysis pipeline common for activity recognition in ubiquitous computing (see chapter 2.4). First the captured data is segmented using a sliding window procedure, after which a hand-crafted set of features is extracted from each frame. In cross-validation experiments these features are then used to train a sequence of Restricted Boltzmann Machines (RBMs) (see chapter 4.3.3). A softmax top-layer (see chapter 2.6.1) is added to the trained generative model which is further fine-tuned using conjugate gradients to maximise classification performance (see figure 8.1).

### 8.3.1 *Wearable sensing system*

Our sensing setup consists of two movement sensors, one worn on each wrist of the participant, which have been used in previous applications such as in Autism research (see chapter 3), and sports (see chapter 7). The movement sensors contain a tri-axial accelerometer that measures acceleration along three perpendicular axes with high temporal resolution (100 Hz) (see chapter 2.2.3). These devices are able to capture acceleration data for up to 12 days on a single charge. The sensors are attached using comfortable velcro straps and are waterproof. Colour coding ensured that the sensor

location and orientation remained constant throughout the study. This sensing system represents a compromise between usability and signal quality. The small number of sensors in a convenient location along with their high usability allow data capture in the daily life of the participants with very high compliance. However, for the sake of prolonged battery life no further modalities beyond accelerometers were included (e.g. gyroscopes, magnetometer).

### 8.3.2 *Data collection*

Overall 34 participants were recruited who exhibited mild to severe level Parkinson's Disease (Hoehn and Yahr stages I-IV [Hoehn and Yahr, 1998]), were not significantly cognitively impaired and were taking immediate-release levodopa medication. All participants provided informed consent for involvement and ethical approval was obtained from the relevant authorities. The study design in this chapter follows the insights obtained in chapter 7. We collect data from both a constrained (clinical) and a naturalistic setting in two subsequent phases:

**Phase 1 (LAB)** consists of lab-based recordings. Participants attended a movement research laboratory without having taken their early morning dose of medication (where possible) and spent on average 4 hours in the facility while wearing the sensing system. At regular intervals (e.g. once per hour or more), the current state of the disease was assessed by a clinician. Based on video recordings, a second clinician rated the disease state for each examination. Assessments where the two clinicians disagreed where discarded (overall agreement $> 0.95$). Data is extracted surrounding each of the 141 remaining assessments. The assessment itself is removed as participants engage in a series of movements selected to assist within clinical evaluation but are highly unlikely to be representative of naturalistic behaviour. Data from phase 1 is denoted as LAB throughout the rest of this chapter.

**Phase 2 (HOME)** corresponds to longitudinal recordings in the participant's private homes. After completing phase 1, participants wore the sensing system continuously over the course of a week, including at night. Each participant filled out a disease state diary, a pre-formatted document where ticks indicate disease state for each hour, to the best of their abilities. The diary included: *asleep*, *off*, *on*, and (troublesome) *dyskinesia*. A total of approx. $5,500$ hours of accelerometer data was collected, for which approx.

4,500 hourly labels were provided by the participants ( 80% diary compliance). The labels are inherently unreliable, as symptom characteristics are very unlikely to change exactly on the hour, participants may have trouble classifying their own disease state, and diaries may be filled out retrospectively at the end of the day. Data collected in phase 2 is denoted as HOME throughout the rest of this chapter.[1]

### 8.3.3 *Pre-processing and feature extraction*

Each disease state is characterised by different expressions of the common motor features in PD. During the *off* state, people with PD feel slow, stiff and may show increased tremor. In the *on* state, symptoms are less severe and tremor may disappear completely. Bouts of dyskinesia present as somewhat repetitive involuntary movements that may involve the wrists. Crucially the recorded data does not just contain the expression of the disease states but includes (unknown) naturalistic physical activities that have significant effect on the recorded signal. We extract features from segmented accelerometer data (see chapter 2.4.2), where each segment spans one minute in duration. In order to avoid a possible bias in our experiments and due to the large size of the data-set we extract segments that do not overlap. In these relatively long segments we aim to even out the impact of physical activities and try to capture the underlying characteristics, expressed as differences in the distribution of the acceleration measurements.

From the raw recordings contained in each frame $f^t = \left(f_L^t, f_R^t\right) \in R^{n\times 6}$ the acceleration magnitudes for each sensor $m_L, m_R$ are estimated which are subsequently filtered using a high-pass filter with a cut-off frequency of 0.5Hz to remove the gravitational component. The filtered magnitudes are used to obtain their first derivatives (jerk) $j_L, j_R$. The magnitude of orientation change $c_L, c_R$ is calculated from the raw recordings of each sensor as follows:

$$c_L \;=\; \left\{\cos^{-1}\left(f_{L,i}\cdot f_{L,i+1}\right)\right\}_{i=1\cdots(n-1)}, \tag{8.1}$$

where $(\cdot)$ denotes the vector dot-product, $f_{L,i} \in \mathbf{R}^3$ are the recordings of sensor $L$ at position $i$ (relative time within frame). $c_R$ is calculated accordingly for the sensor on the right wrist. Based on $m_L$ and $m_R$ we estimate the power spectral density $p_L, p_R$ using a periodogram on 10 frequency bands between 1 and 8 Hz to capture repetitive movements typical for motor features in PD.

[1]Data-set is available at **http://di.ncl.ac.uk/naturalisticPD**.

We capture the statistical characteristics of the movement within a frame using the ECDF representation introduced in chapter 5, which corresponds to concatenated quantile functions along with their mean. For each frame we obtain its feature representation $x^t$ by concatenating the ECDF representations of the acceleration magnitudes $m_L, m_R$, jerk $j_L, j_R$, orientation change $c_L, c_R$ and power spectral density $p_L, p_R$. We further include the *time spent not moving* (threshold on $c_L + c_R$) as in [Griffiths et al., 2012], energy, minimum, maximum, standard deviation of $m_L, m_R$ and binary PD phenotype (*tremor-dominant*). Using 10 coefficients in the ECDF representation we extract a total of 91 features from each minute of sensor recordings. In the future, this hand-crafted feature extraction will be substituted with a convolutional architecture alleviating the need for medical prior knowledge.

### 8.3.4 *Training procedure*

The training procedure comprises of two steps. First the real-valued features are normalised to have zero mean and unit variance (per fold in cross validation). We then apply RBMs to learn a generative model of the input features as described in chapter 4.3.3. After training the first RBM, the activation probabilities of its feature detectors are used as input data for the next RBM. This way, RBMs can be used to greedily initialise deep neural networks by adding more and more layers [Hinton and Salakhutdinov, 2006]. We learn at most two consecutive RBMs, where the first one contains gaussian visible units (gaussian-binary) to model the real-valued input features and the next one just contains binary units (binary-binary) (see chapter 4.3.3). Learning rates were set to $10^{-4}$ for the gaussian-binary RBM, and $10^{-3}$ for the binary-binary RBM, with a momentum of 0.9 and a weight-cost of $10^{-5}$. Each RBM is trained for 500 epochs with batches containing 500 samples. Crucially, this first phase of training does not rely on any labels of the input data and is solely driven by the objective to learn a generative model of the training data.

In the subsequent fine-tuning phase we add a top-layer (randomly initialised, $\sigma = 0.01$) to the generative model. This top-layer contains 4 units in a softmax group (see chapter 2.6.1) that correspond to our 4 classes of interest: *asleep*, *off*, *on*, and *dyskinetic*. Using the labels for each input frame we perform 250 epochs of conjugate gradients with batches that gradually increase in size from 256 up to 2,048 (stratified) samples. In the first epoch the weights in all but the top layer remain fixed. Training time averages to around one day per fold on a GPU. Effectively our training procedure first performs

unsupervised learning using RBMs to obtain an initialisation for a discriminative neural network, which is subsequently fine-tuned in a supervised learning procedure using the labelled training examples.

## 8.4 Experimental evaluation

Two scenarios are investigated in this chapter. In the first setting, a variety of approaches and network architectures are trained on the HOME data-set. To minimise the effect of large pairwise similarity of subsequent minutes of recording we follow a leave-one-*day*-out cross validation approach, where e.g. the first day of recording from all patients constitutes a fold. This represents a compromise between realistic assessment of generalisation performance and the required computational effort for training (which is extensive). The second setting simulates best practice for assessment systems in PD, where the smaller but clinician validated LAB data-set is used for training in a stratified 7-fold cross validation which is subsequently applied to the HOME data-set to assess generalisation performance.

In total the HOME data-set contains approx. $270,000$ samples (minutes) and the LAB data-set contains $1,410$ samples extracted from the recordings surrounding 141 individual disease-state assessments. Additional minutes are extracted for networks that span more than one minute in their input, such that the overall number of samples is retained. Since e.g. the HOME data-set is highly skewed towards *asleep* (31%) and *on* (41%) we chose the mean F1 score as primary performance metric:

$$\frac{2}{c}\sum_{i=1}^{c}\frac{\mathtt{prec_i} \times \mathtt{recall_i}}{\mathtt{prec_i} + \mathtt{recall_i}}, \tag{8.2}$$

where $\mathtt{prec_i}$ corresponds to the precision, $\mathtt{recall}_i$ to the recall observed for class $i$ and $c$ to the number of classes (see chapter 2.6.4). The LAB data-set does not contain any instances of *asleep* and the performance is evaluated just using the remaining three classes, even though false positives for *asleep* are included in the calculation of recall and precision.

To illustrate the difficulty of the problem we compare the approach proposed in this chapter with standard classification methods typical for HAR systems as discussed in chapter 2.6. We apply decision trees (C4.5), Naive Bayes (NB), and nearest neighbour classification(1-NN). We further apply support vector machines (SVM) with an

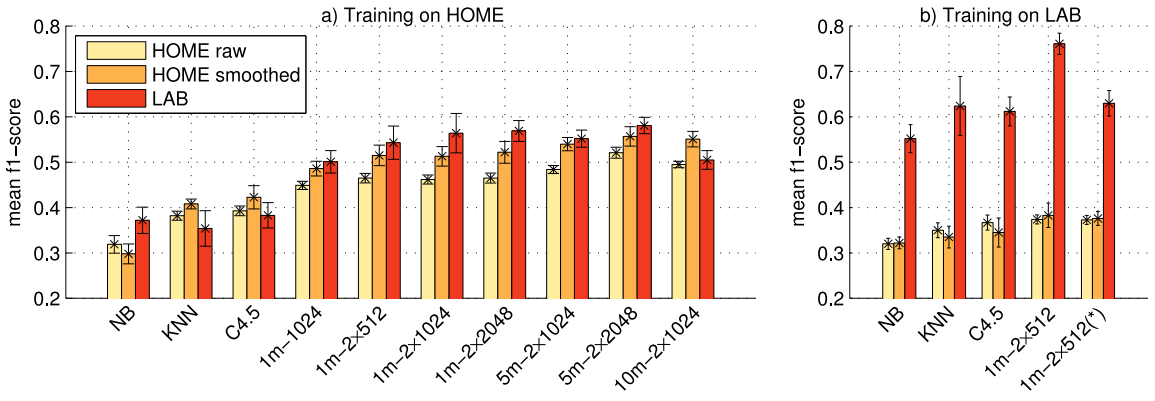



FIGURE 8.2: Recognition results for different models. The left plot shows the performance of models trained on the HOME data-set, the right plot shows the performance for models trained on the LAB data-set. Errorbars indicate one standard deviation, estimated based on the performance of individual folds in the cross-validation or when applied to the validation set. Colour indicates which data-set was used for evaluation. *Smoothed* results are post-processed using the mean over the predictions for one hour. (*) indicates networks pre-trained on HOME and fine-tuned on LAB.

rbf-kernel for training on the LAB data-set. On the HOME data-set we failed to achieve convergence to non-trivial solutions in SVMs. In order to investigate the impact of the layout of the deep ANN proposed in this chapter we evaluate a number of different network topologies the results of which are discussed below.

## 8.5 Results

Recognition results are illustrated in Figure 8.2. The left plot shows the results for approaches trained on the HOME data-set, while the right plot shows results for those trained on the LAB data-set. Labels indicate the method or network topology, e.g. "$5m - 2 \times 1024$" translates to 5 minutes of input and two hidden layers with 1,024 units each. For each approach, three results are reported: i) the performance on individual frames in the HOME data-set ("raw"), ii) smoothed predictions using a sliding window of 60m duration ("smoothed") and a step size of 20 minutes, and iii) the performance on the LAB data-set. The rationale behind smoothing the predictions over time is that in clinical applications the fluctuations would be assessed over longer time-frames, instead of being based on individual minute-by-minute predictions. Furthermore the movements captured within each extracted frame may be unrepresentative for the disease state, where the smoothing process approximates the mean movements over longer periods of time.

We first discuss systems trained on the HOME data-set. Overall the traditional approaches perform rather poorly in this setting. While smoothing slightly improves results, KNN, and C4.5 show a drop in performance on the LAB validation set. However, the various deep network topologies investigated here not only show significantly better performance than e.g. C4.5, but also (mostly) show a performance on the LAB validation set that exceeds the performance on the HOME data-set. We infer that the apparent increase in performance on the validation set illustrates the poor quality of the participant-provided labelling in the training-set, rather than an unexpected generalisation ability. Nevertheless it indicates that deep NNs are able to capture disease characteristics that remain inaccessible to more traditional methods, which may be reasoned in the gradient-based optimisation approach that implicitly placed a degree of weight on each sample. Normalised confusion matrices for the best performing network are illustrated in Figure 8.3. The class with the lowest performance on the HOME data-set is *dyskinesia*. Interestingly that class shows high specificity in the validated LAB data-set, indicating particularly unreliable labels for this disease state in the training-set. Overall adding a second layer and adding more units to the hidden layers improves the results, which is in line with previous results on this type of model. The best results are obtained for networks that span 5 minutes of input. If the input span is increased further to 10 subsequent minutes we see a drop in the performance on the validation set.

The results for systems trained on the LAB data-set differ strongly to those above. While the recognition performance on the LAB data-set (in cross-validation) is very good with peak mean f1-score of 0.76, the generalisation performance when applied to the HOME data-set is very disappointing. We further found no evidence that pre-training a model on the HOME data-set with subsequent fine-tuning on the validated LAB data-set provided significant improvement in generalisation performance (see "(*)" in Figure 8.2). Instead we see a decline in the cross validation performance, which supports our initial assumption that laboratory-based data is just an incredibly poor model for naturalistic behaviour.

### 8.5.1 *Comparison to related approaches*

When trained on the HOME data-set, we see a peak mean f1-score of 0.581 on the LAB data-set, which corresponds to an overall accuracy of 59.4%. On average the classes are differentiated with a sensitivity of 0.57 and a specificity of 0.88. It is difficult to compare

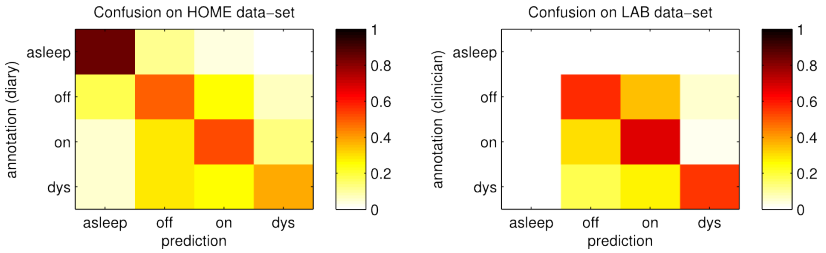



FIGURE 8.3: Normalised confusion matrices on both the (smoothed) HOME data-set (left) and on the Lab data-set (right) for a model with 5 minutes as input and two hidden layers with 2048 units each. While the performance on the class *dys* is relatively low in the HOME data-set there are just very few false positives for this class in the laboratory setting.

these results to prior art, as no systems exist that follow a similar training and evaluation methodology. Hoff et al. [Hoff et al., 2004] report very similar performance figures with sens. and spec. around 0.7 for *on* and *off* states over a 24h period on 15 participants with PD (compared to 4 states in this chapter). However, their sensing approach was based on a network of 7 sensors placed across the body and their prediction relied on thresholds set for each individual to maximise performance, effectively limiting the practicality of their approach. For a gold standard, consider that trained nurses may show relatively low accuracy of 0.65 when assessing the severity of motor complications [Palmer et al., 2010].

Systems trained on the LAB data-set show good performance in cross validation experiments up to a mean f1-score of 0.76. These results are comparable with other systems applied in laboratory settings [Maetzler et al., 2013]. However, our results indicate that the poor generalisation to realistic behaviour of this artificial setting may also affect other systems based on similar laboratory environments, which has so far not been demonstrated.

## 8.6 Discussion

The quality of life of people affected by PD depends on the management of their condition in the form of tailored treatment plans. Devising such plans is a challenge, as objective information about fluctuations in disease state is not accessible in clinical practice beyond recall by the individual. Current best practice in automated assessment of PD is to obtain data in laboratory conditions, where small amounts of clinician-validated behaviour can be observed. While such systems show good performance in this setting, it is unlikely that they generalise to naturalistic behaviour in people's daily lives. In

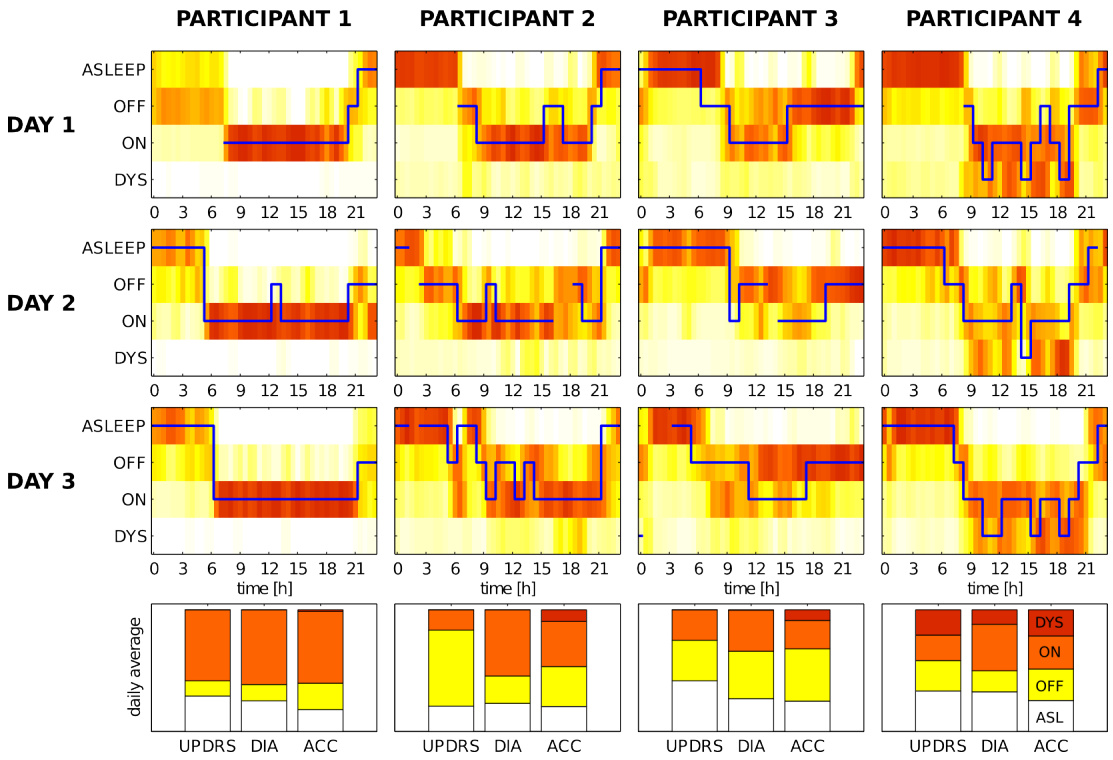



FIGURE 8.4: Predictions of the best performing network for three consecutive days and the mean prediction for all 7 days of four participants. Each subplot shows the colour-coded predictions of the network over time (white=0 to red=1). The blue lines indicate the diary entries recorded by each participant (line omitted for missing entries). The bottom row indicates the distribution of disease state according to patient recall (UPDRS), diary (DIA) and the assessment system (ACC). Best viewed in colour.

order to address this issue, assessment systems have to be based on data collected in naturalistic, ecologically valid surroundings such as the private home.

In this chapter we investigated the problem of the assessment of disease state in PD based on a large data-set of many weeks worth of movement data collected from 34 individuals. We developed a novel methodology for research on this problem, which is based on large quantities of naturalistic behaviour collected in the daily life of people affected by PD. Naturalistic environments pose significant challenges for data acquisition in the form of usability constraints as well as challenges that stem from unreliable labelling obtainable in this setting (further discussed in chapter 2). Labels that are accessible are infrequent with respect to the data sampling rates and inherently unreliable, subject to recall bias, class confusion, and boundary issues.

In our experiments we showed that deep learning seems particularly suitable to discover disease characteristics despite unreliable labelling of training-data, a setting in which other methods such as decision trees provide poor (generalisation) performance. Deep

learning has been applied in similar settings, such as speech [Deng et al., 2013] or object recognition [Lee et al., 2009], where unlabelled data is easily accessible. However, our results indicate that the common approach to pre-train deep architectures on unlabelled data with subsequent fine-tuning based on a (smaller) set of labelled instances does not improve results in this problem setting. The behaviour observed in laboratory conditions just appears to be a very poor model for naturalistic behaviour, as systems trained on that data show disappointing generalisation performance.

The performance of even the best model does not exceed a mean f1-score of 0.6. To an extent such low results are explained by the poor quality labelling. However, even this relatively low performance is useful for clinical practice. Illustrated in Figure 8.4 are the predictions for three consecutive days for four participants of the best performing network, where the predicted disease states clearly show very similar patterns of fluctuation compared to the diary entries for each participant. Beyond the assessment of fluctuations in disease state there are other clinical applications. A common measure for the efficacy of interventions in PD is an overall reduction in e.g. "off time", where the average activation of output-units of our system (ACC) only shows little difference to the current best practice for this assessment (DIA) (see Figure 8.4).

We have not observed any over-fitting to the naturalistic behaviour in the HOME dataset. The unreliable labelling leads to many inconsistencies, which naturally prevent over-fitting. A more pressing concern is under-fitting, where automatically adapting or omitting episodes with low confidence may provide significant improvements over the current results. We found that it is crucial to utilise large mini-batches during training (up to 2,048 samples), which may also stem from the unreliable labelling. Another issue surrounds the feature extraction. Effectively the disease state has little impact on the movement data, whose primary source of variance are the physical activities the participants engage in, such as walking. For systems to generalise across those activities it is crucial to tailor a feature representation towards the underlying movement characteristics. These should be accessible to data-driven approaches that avoid manual feature engineering, such as convolutional architectures or techniques like sparse coding [Bhattacharya et al., 2014].

In summary, the problem of disease state assessment in PD is far from being solved. It appears that current challenges may be overcome if novel methodologies are employed in research on PD, where suitable machine learning methods play a key role. Advances that address the unique challenges of this problem setting will have significant societal impact, as not only individuals with PD but also many other degenerative conditions

would benefit from practical assessment systems. Beyond possible impact, the characteristic challenges of naturalistic settings make for a unique machine learning problem, which could serve as a novel test-bed for the development and evaluation of unsupervised or semi-supervised learning approaches.

# Chapter 9. Summary

## 9.1 Discussion

One of the main aims of ubiquitous computing is the development of automated recognition systems for human activities and behaviour that are sufficiently robust to be deployed in realistic, in-the-wild environments. In most cases, the targeted applications scenario are people's daily lives, where systems have to abide by practical usability and privacy constraints. The development of a recognition approach robust to naturalistic environments requires large amounts of annotated data from representative settings. However, in practice it is difficult to capture reliable ground-truth annotation for this naturalistic data as it is impractical to deploy (and manually label) video recordings in e.g. people's private homes. The majority of systems in ubiquitous computing and rehabilitation therefore rely on data collected from settings that are, to an extent, constrained. Participants may engage in scripted routines or perform simulated activities in an artificial (instrumented) environment. While this setting is perfectly suitable to gain an impression of the complexity of activity recognition in the problem domain it remains doubtful if systems developed in artificial study settings generalise towards real-world behaviour.

We explore the challenges of naturalistic environments towards sensing and analysis of human movements and find that body-worn movement sensors in particular are suitable for real-life applications of ubiquitous computing (see chapter 2). We find that it is crucial that sensing systems cater towards the requirements of the target audience in e.g. limiting the number of sensors placed in convenient locations, which affects the resolution of the sensing system. The different components of typical pipeline-based activity recognition systems are evaluated with respect to their suitability for use with data collected from naturalistic surroundings, highlighting how the reliance on artificial study settings and manual design of the pipeline components give rise to concerns regarding their robustness towards real-life situations.

To an extent, these concerns can be addressed by incorporating additional data-sets that e.g. contain background activities, as illustrated in chapter 3. This can lead to insights that may otherwise remain inaccessible, particularly if the added data is from a naturalistic setting. In practice the additional data can be utilised to demonstrate robustness of a system developed in an artificial setting. It does not, however, assist during the manual design process of the components of a recognition system, which mostly relies on prior knowledge, experience and intuition of the practitioner. Adapting the study design can not fully alleviate these concerns which instead requires novel computational tools that minimise human intervention during the design of activity recognition systems.

The main contribution of this thesis is to show that components of the most common recognition approaches in ubiquitous computing can be substituted with novel computational methods, namely deep and feature learning in the form of RBMs, which show favourable properties for naturalistic environments. These methods follow a data-driven approach that allows effective use of naturalistic data without the need for any annotations. We demonstrate two applications scenarios for these methods: i) deep learning can be applied to extract features from accelerometer data that show a recognition performance superior to other approaches across a variety of application domains (see chapter 4); and ii) deep learning provides a powerful tool to utilise large amounts of unlabelled data to initialise a (discriminative) multi-layer neural network, which shows particularly good performance in naturalistic environments (see chapter 8).

Performance of feature learning can be further improved upon if inertial movement data is represented using the inverse of their empirical cumulative distribution – the ECDF representation developed in this thesis. This representation preserves statistical characteristics of accelerometer data and was demonstrated to provide excellent recognition results on a variety of publicly available data-sets (see chapter 5). It is further very efficient to compute and holds the potential for embedded applications of HAR in resource-constrained environments. It is straight-forward to substitute existing (hand-crafted) statistical feature extraction approaches with this simple analytical process to significantly reduce the risk of over-fitting to artificial study settings. Even though this representation was developed to address challenges typical for data captured with accelerometers it should also be a powerful feature extraction approach for other sensing modalities, where it can similarly substitute or augment (statistical) feature extraction approaches. The insights obtained in chapter 5 also hold for other non-stationary time-series where the underlying distribution may change rapidly over short periods of time,

for which this representation should be particularly suitable. In future work we will investigate the performance of the ECDF representation in these additional settings and how it can be further improved to maximise performance of HAR systems.

The benefit of these methods, when applied in ubiquitous computing, can be maximised by altering standard study design to allow for the collection of unlabelled data. Naturalistic data, particularly in clinical applications, has so far been of limited use as the missing annotation does not allow an evaluation that is sufficiently robust towards scrutiny by domain experts. In this thesis we propose the use of a two-part study protocol for these settings, which combines the use of largely unlabelled or just unreliably labelled data from a real-life environment with an additional smaller set of semi-naturalistic data that is annotated by domain experts. The naturalistic data can be used in conjunction with deep and feature learning to develop a recognition approach robust towards real-life environments, while the second of semi-naturalistic data acts as a validation set annotated to a (clinical) gold standard or can be used to develop more detailed movement analysis in the form of skill assessment. We demonstrate this approach in two applications: i) in sports, where both the detection of climbing activities and the extraction of specific skill parameters benefit from this study arrangement (see chapter 7); and ii) in a clinical settings, where this study setting combined with deep learning effectively presents a novel approach to automated assessment for degenerative conditions such as Parkinson's Disease (see chapter 8).

## 9.2 Limitations and Future Work

**Sequential Feature Learning on Accelerometer Data**

The application of deep and feature learning in this thesis was limited to *static* learning approaches that are suitable for HAR systems relying on sliding window feature extraction. Effectively these methods assume statistical independence of each input dimension, which allows for efficient learning algorithms. This is an assumption that does not hold in the case of raw accelerometer data, as subsequent points are clearly correlated over time. Even though we observed very promising performance of these approaches it is nonetheless likely that this performance can be improved upon significantly if the models incorporate temporal information. At the moment the features extracted are inherently based on the *appearance* of the input data, which may lead to representations that are unstable with respect to specific activities. Subsequent frames of accelerometer data, even if they contain the same activity, are likely to show e.g. characteristic peaks

at different relative positions within the frame. In an appearance-based approach those frames obtain representations that can be very different, which complicates the task of classification as such features do not necessarily fall within one cluster per activity. This reliance on appearance explains – to an extent – why performance is improved if the raw accelerometer data is transformed using the ECDF representation (see chapter 4), as this representation effectively removes temporal dependencies while retaining crucial characteristics. In future work we will explore novel approaches that extend on static feature learning, either by explicitly modelling the covariance structure in input data [Ranzato and Hinton, 2010], through temporal dependencies between model components [Taylor and Hinton, 2009], or through a convolutional approach with shared feature maps down to the sample level [Lee et al., 2009].

**Visible units for accelerometer data**

Another challenge when applying feature learning approaches to accelerometer data surrounds the computational models for the input data, i.e. the visible units in the case of RBMs. In this work we relied on Gaussian units, where input samples are assumed to be drawn from an underlying Gaussian with unit variance and a mean that depends on the input from the layers of the neural network. However, as discussed in chapter 5, accelerometer data is inherently non-gaussian and subject to rapidly changing bias that depends on the orientation of the sensing device. Even if the overall distribution (per axis) is utilised to normalise accelerometer data to zero mean and unit variance, such short-term biases remain for a large fraction of the activities of interest. This is detrimental to the performance of feature learning, as modelling a significant bias (i.e. gravity) in accelerometer data requires configurations of high energy, which are difficult to model. For the successful future application of feature learning to accelerometer data it is crucial to develop a novel type of visible unit that addresses these characteristics explicitly, which is likely to significantly improve the performance.

**Active cooperation by users**

An alternative approach to the pipeline-based activity recognition described in this work is to actively seek cooperation from the user of a recognition system. This would allow for adaptation of the model to idiosyncrasies of the user and their naturalistic setting. These approaches are usually conceived using traditional approaches to e.g. feature extraction and mainly adapt their recognition backend to the activities they observe over time. Similarly to the approaches described in this work, these methods are likely to benefit from the application of deep learning, which may increase their potential for reliable adaptation to the user over time.

# Bibliography

Abowd, D., Dey, A. K., Orr, R., and Brotherton, J. (1998). Context-awareness in wearable and ubiquitous computing. *Virtual Reality*, 3(3):200–211.

Abowd, G. D., Bobick, A. F., Essa, I. A., Mynatt, E. D., and Rogers, W. A. (2002). The aware home: A living laboratory for technologies for successful aging. In *Proceedings of the AAAI-02 Workshop "Automation as Caregiver*, pages 1–7.

Achenbach, T. M. and Rescorla, L. A. (2000). *Manual for the ASEBA Preschool Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Adelsberger, R. and Troster, G. (2013). Experts lift differently: Classification of weight-lifting athletes. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, pages 1–6. IEEE.

Aggarwal, J. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16.

Ahmadi, A., Rowlands, D., and James, D. A. (2010). Towards a wearable device for skill assessment and skill acquisition of a tennis player during the first serve. *Sports Technology*, 2(3-4):129–136.

Ahmadi, A., Rowlands, D. D., and James, D. A. (2006). Investigating the translational and rotational motion of the swing using accelerometers for athlete skill assessment. In *Sensors, 2006. 5th IEEE Conference on*, pages 980–983. IEEE.

Ahmadi, S.-A., Padoy, N., Rybachuk, K., Feussner, H., Heinin, S., and Navab, N. (2009). Motif discovery in or sensor data with application to surgical workflow analysis and activity detection. In *M2CAI workshop, MICCAI, London*.

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Tudor-Locke, C., Greer, J. L., Vezina, J., Whitt-Glover, M. C., and Leon, A. S. (2011). 2011

compendium of physical activities: a second update of codes and met values. *Medicine and science in sports and exercise*, 43(8):1575–1581.

Albinali, F., Goodwin, M. S., and Intille, S. S. (2009). Recognizing stereotypical motor movements in the laboratory and classroom: a case study with children on the autism spectrum. *Proc. Int. Conf. Ubiquitous Computing*.

Allen, F. R., Ambikairajah, E., Lovell, N. H., and Celler, B. G. (2006). Classification of a known sequence of motions and postures from accelerometry data using adapted gaussian mixture models. *Physiological Measurement*, 27(10):935.

Aman, M., Singh, N., Stewart, A., and Field, C. (1985). Psychometric characteristics of the aberrant behavior checklist. *American J. of Mental Deficiency*, 89:492–502.

Andre, D. and Wolf, D. L. (2007). Recent advances in free-living physical activity monitoring: a review. *Journal of Diabetes Science and Technology*, 1(5):760–7.

ASPA (2013). Animals (scientific procedures) act 1986 (aspa). `bit.ly/10rCGf7`. accessed: March 22th, 2013.

Atallah, L., Lo, B., King, R., and Yang, G. (2011). Sensor positioning for activity recognition using wearable accelerometers. *Biomedical Circuits and Systems, IEEE Transactions on*, 5(4):320–329.

Atallah, L. and Yang, G.-Z. (2009a). The use of pervasive sensing for behaviour profiling — a survey. *Pervasive and Mobile Computing*, 5:447–464.

Atallah, L. and Yang, G.-Z. (2009b). The use of pervasive sensing for behaviour profiling—a survey. *Pervasive and Mobile Computing*, 5(5):447–464.

Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., and Havinga, P. (2010). Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *Architecture of computing systems (ARCS), 2010 23rd international conference on*, pages 1–10. VDE.

Axivity (2013). Axivity. `www.axivity.com`. accessed: March 11th, 2013.

Bächlin, M., Förster, K., and Tröster, G. (2009). Swimmaster: A wearable assistant for swimmer. In *Proc. Int. Conf. Ubiquitous Comp. (UbiComp)*.

Bächlin, M., Plotnik, M., Roggen, D., Maidan, I., Hausdorff, J. M., Giladi, N., and Tröster, G. (2010). Wearable assistant for parkinson's disease patients with the freezing of gait

symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446.

Bao, L. and Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. In *Pervasive computing*, pages 1–17. Springer.

Bardram, J. E., Doryab, A., Jensen, R. M., Lange, P. M., Nielsen, K. L., and Petersen, S. T. (2011). Phase recognition during surgical procedures using embedded and body-worn sensors. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, pages 45–53. IEEE.

Barry, D. T., Hill, T., and Im, D. (1992). Muscle fatigue measured with evoked muscle vibrations. *Muscle & Nerve*, 15(3):303–309.

Beastmaker (2013). Beastmaker. `bit.ly/YvAzlR`. accessed: March 11th, 2013.

Bergmann, J. and McGregor, A. (2011). Body-worn sensor design: What do patients and clinicians want? *Annals of biomedical engineering*, 39(9):2299–2312.

Berlin, E. and Van Laerhoven, K. (2012). Detecting leisure activities with dense motif discovery. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 250–259. ACM.

Best, R. and Begg, R. (2006). Overview of movement analysis and gait features. *Computational Intelligence for movement sciences. Neural networks and other emerging techniques.*, pages 1–69.

Bhattacharya, S., Nurmi, P., Hammerla, N., and Plötz, T. (2014). Using unlabeled data in a sparse-coding framework for human activity recognition. *Pervasive and Mobile Computing*.

Biggio, B., Nelson, B., and Laskov, P. (2011). Support vector machines under adversarial label noise. In *ACML*, pages 97–112.

Bilmes, J. A. et al. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

Bishop, C. M. et al. (1995). Neural networks for pattern recognition.

Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

Blanke, U. and Schiele, B. (2009). Daily routine recognition through activity spotting. In *Location and Context Awareness*, pages 192–206. Springer.

Bobick, A. (1997). Movement, activity and action: The role of knowledge in the perception of motion. *Philosophical Trans. of the Royal Society B: Biological Sciences*, 352(1358):1257–1265.

Bulling, A., Blanke, U., and Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):33.

CAI (2013). Club alpino italiano. `www.cai.it`. accessed: March 11th, 2013.

Cawley, G. C. and Talbot, N. L. (2007). Preventing over-fitting during model selection via bayesian regularisation of the hyper-parameters. *The Journal of Machine Learning Research*, 8:841–861.

Chan, M., Estève, D., Escriba, C., and Campo, E. (2008). A review of smart homes—present state and future challenges. *Computer methods and programs in biomedicine*, 91(1):55–81.

Chang, K., Hightower, J., and Kveton, B. (2009). Inferring identity using accelerometers in television remote controls. In *Proceedings of the 7th International Conference on Pervasive Computing*, pages 151–167. Citeseer.

Chang, K.-H., Chen, M. Y., and Canny, J. (2007). Tracking free-weight exercises.

Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S. T., Tröster, G., del R. Millán, J., and Roggen, D. (2013). The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*.

Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., and Yu, Z. (2012). Sensor-Based Activity Recognition. *IEEE Trans on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, 99.

Chou, K. (1995). A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. *Proteins: Structure, Function, and Bioinformatics*, 21(4):319–344.

Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., et al. (2008). The mobile sensing platform: An embedded activity recognition system. *Pervasive Computing, IEEE*, 7(2):32–41.

ClimbCoach (2013). Climbcoach. `www.climbcoach.org`. accessed: March 11th, 2013.

Club, T. K. (2013). The kennel club. `www.thekennelclub.org.uk/`. accessed: March 11th, 2013.

Clutton-Brock, J. (1999). *A Natural History of Domesticated Mammals*. Cambridge University Press.

Coldwell, W. (2012). Indoor climbing: the rise of bouldering-only centres. The Guardian Online Edition, `bit.ly/MWPd1m`. accessed: March 11th, 2013.

Cole, B., Roy, S., De Luca, C., and Nawab, S. (2010). Dynamic neural network detection of tremor and dyskinesia from wearable sensor data. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 6062–6065. IEEE.

Cook, D. J. and Das, S. K. (2007). How smart are our environments? an updated look at the state of the art. *Pervasive and mobile computing*, 3(2):53–73.

Cooper, A. M. and Michels, R. (1994). *Diagnostic and statistical manual of mental disorders*. Number 4. American Psychiatric Association.

Cordier, P., France, M. M., Pailhous, J., and Bolon, P. (1994). Entropy as a global variable of the learning process. *Human Movement Science*, 13(6):745–763.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, volume 21. Chapman & Hall/CRC.

Daniel S. Mills (2010). *The encyclopedia of applied animal behaviour and welfare*. cabi.org.

de Lau, L. and Breteler, M. (2006). Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6):525–535.

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8599–8603. IEEE.

Dosis, A., Aggarwal, R., Bello, F., Moorthy, K., Munz, Y., Gillies, D., and Darzi, A. (2005). Synchronized video and motion analysis for the assessment of procedures in the operating theater. *Archives of Surgery*, 140(3):293–299.

Draper, N., Dickson, T., Blackwell, G., Priestley, S., Fryer, S., Marshall, H., Shearman, J., Hamlin, M., Winter, D., and Ellis, G. (2011). Sport-specific power assessment for rock climbing. *The Journal of Sports Medicine and Physical Fitness*, 51(3):417–425.

Draper, N., Dickson, T., Fryer, S., Blackwell, G., Winter, D., Scarrott, C., and Ellis, G. (2012). Plasma cortisol concentrations and perceived anxiety in response to on-sight rock climbing. *Int. Journal of Sports Medicine*, 33(1):13–7.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification*. Wiley.

Edward M. Jr. Gilbert, Thelma R.Brown (1995). *K9 Structure and Terminology*. Howell Book House.

Edward Price (2008). *Principles and Applications of Domestic Animal Behavior*. cabi.org.

Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 726–733. IEEE.

Eisenhower, A., Baker, B., and Blacher, J. (2005). Preschool children with intellectual disability; syndrome specificity, behaviour problems, and maternal well-being. *J. of Intellectual Disability Research*, 49:657–671.

Fentem, P. H. (1994). Benefits of Exercise in Health and Disease. *British Medical Journal*, 308:1291–1295.

Figo, D., Diniz, P. C., Ferreira, D. R., and Cardoso, J. M. (2010). Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, 14(7):645–662.

Figueiredo, M. A. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396.

Fink, G. A. (2008). *Markov Models for Pattern Recognition – From Theory to Applications*. Springer.

FitBit (2013). Fitbit. `www.fitbit.com`. accessed: March 11th, 2013.

Fitts, P. and Posner, M. (1967). *Human performance.* Brooks/Cole.

Fogarty, J., Hudson, S. E., Atkeson, C. G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. C., and Yang, J. (2005). Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(1):119–146.

Foster, S. L. and Cone, J. D. (1986). Design and use of direct observation. In A.R. Ciminero, Calhoun, K., and Adams, H. E., editors, *Handbook of behavioral assessment*, pages 253–354. Wiley, New York.

Fothergill, S., Harle, R., and Holden, S. (2008). Modeling the model athlete: Automatic coaching of rowing technique. In *Proc. Joint IAPR Int. Workshop on Structural, Syntactic, and Statistical Pattern Recognition (SSPR & SPR)*.

Foxlin, E. (2005). Pedestrian tracking with shoe-mounted inertial sensors. *Computer Graphics and Applications, IEEE,* 25(6):38–46.

Frank, J., Mannor, S., and Precup, D. (2010). Activity and gait recognition with time-delay embeddings. In *Proceedings of AAAI Conference on Artificial Intelligence*.

Fuel, N. (2013). Nike fuel. `www.nike.com/us/en_us/c/nikeplus-fuelband`. accessed: March 11th, 2013.

Fuss, F. K. and Niegl, G. (2009). Instrumented climbing holds and performance analysis in sport climbing. *Sports Technology*, 1(6):301–313.

Fyffe, A. and Peter, I. (1997). *The Handbook of Climbing*. Pelham Books.

Garcia, S., Derrac, J., Cano, J. R., and Herrera, F. (2012). Prototype selection for nearest neighbor classification: taxonomy and empirical study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):417–435.

Giuffrida, J. P., Riley, D. E., Maddux, B. N., and Heldman, D. A. (2009). Clinically deployable kinesia™ technology for automated tremor assessment. *Movement Disorders*, 24(5):723–730.

Godfrey, A., Conway, R., Meagher, D., and OLaighin, G. (2008). Direct measurement of human movement by accelerometry. *Medical Engineering & Physics*, 30(10):1364–1386.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., et al. (2008). Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results. *Movement disorders*, 23(15):2129–2170.

Goodwin, C. J. (2009). *Research in psychology: Methods and design*. John Wiley & Sons.

Goodwin, M. S., Intille, S. S., Albinali, F., and Velicer, W. F. (2010). Automated Detection of Stereotypical Motor Movements. *J. of Autism and Developmental Disorders*, 41(6):770–782.

Gresham, F. M., Watson, T., and Skinner, C. (2001). Functional Behavioral Assessment: Principles, procedures and future directions. *School of Psychology Review*, 30:156–172.

Griffiths, R. I., Kotschet, K., Arfon, S., Xu, Z. M., Johnson, W., Drago, J., Evans, A., Kempster, P., Raghav, S., and Horne, M. K. (2012). Automated assessment of bradykinesia and dyskinesia in parkinson's disease. *Journal of Parkinson's disease*, 2(1):47–55.

Grobel, K. and Assan, M. (1997). Isolated sign language recognition using hidden markov models. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, volume 1, pages 162–167. IEEE.

Grober, R. (2010). An Accelerometer Based Instrumentation of the Golf Club: Measurement and Signal Analysis. *Arxiv preprint arXiv:1001.0956*.

Grober, R. D. (2009). An Accelerometer Based Instrumentation of the Golf Club: Measurement and Signal Analysis. Arxiv preprint arXiv:1001.0956.

Győrbíró, N., Fábián, Á., and Hományi, G. (2009). An activity recognition system for mobile phones. *Mobile Networks and Applications*, 14(1):82–91.

Hammerla, N. Y., Andras, P., Plötz, T., and Olivier, P. (2011). Assessing Motor Performance with PCA. In *Proc. Int. Workshop Frontiers in Activity Recognition*.

Hanley, G. P., Iwata, B. A., and McCord, B. E. (2003). Functional analysis of problem behavior: A review. *J. of Applied Behavior Analysis*, 36(2):147–185.

Hartley, S., Sikora, D., and McCoy, R. (2008). Prevalence and risk factors of maladaptive behaviour in young children with autistic disorder. *J. of Intellectual Disability Research*, 52(10):819–829.

Hausdorff, J. M. (2009). Gait dynamics in parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(2):026113.

He, Z. and Jin, L. (2009). Activity recognition from acceleration data based on discrete consine transform and svm. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pages 5041–5044. IEEE.

He, Z.-Y. and Jin, L.-W. (2008). Activity recognition from acceleration data using ar model representation and svm. In *Machine Learning and Cybernetics, 2008 International Conference on*, volume 4, pages 2245–2250. IEEE.

Helal, S., Mann, W., El-Zabadani, H., King, J., Kaddoura, Y., and Jansen, E. (2005). The gator tech smart house: A programmable pervasive space. *Computer*, 38(3):50–60.

Herring, S., Gray, L., Taffe, J., Tonge, G., Sweeney, D., and Einfield, S. (2006). Behaviour and emotional problems in toddlers with pervasive developmental disorders and developmental delay: association with parental mental health and family functioning. *J. of Intellectual Disability Research*, 50:874–882.

Hinton, G. (2007). To recognize shapes, first learn to generate images. *Progress in brain research*, 165:535–547.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., et al. (2012a). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hoehn, M. M. and Yahr, M. D. (1998). Parkinsonism: onset, progression, and mortality. *Neurology*, 50(2):318–318.

Hoff, J., Van Der Meer, V., and Van Hilten, J. (2004). Accuracy of objective ambulatory accelerometry in detecting motor complications in patients with parkinson disease. *Clinical neuropharmacology*, 27(2):53–57.

Hoff, J., Wagemans, E., and Van Hilten, J. (2001). Accelerometric assessment of levodopa-induced dyskinesias in parkinson's disease. *Movement disorders*, 16(1):58–61.

Holleczek, T., Schoch, J., Arnrich, B., and Tröster, G. (2010). Recognizing turns and other snowboarding activities with a gyroscope.

Hooper, C., Preston, A., Balaam, M., Seedhouse, P., Pham, C., Jackson, D. G., Plötz, T., and Olivier, P. (2012). The French Kitchen: Task-Based Learning in an Instrumented Kitchen. In *Proc. Int. Conf. Ubiquitous Comp. (UbiComp)*.

Horner, R., Carr, E., Strain, P., Todd, A., and Reed, H. (2002). Problem behavior interventions for young children with autism: A research synthesis. *J. of Autism and Developmental Disorders*, 32(5):423–446.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Horst, E. J. (2008). *Training for Climbing: The Definitive Guide to Improving Your Performance*. Falcon, 2nd edition.

Howlin, P., Goode, S., Hutton, J., and Rutter, M. (2004). Adult outcome for children with autism. *J. of Child Psychology and Psychiatry*, (45):212—229.

Huang, X., Acero, A., Hon, H.-W., and Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR.

Huang, X. D., Ariki, Y., and Jack, M. A. (1990). *Hidden Markov models for speech recognition*, volume 2004. Edinburgh university press Edinburgh.

Huo, X., Ni, X., and Smith, A. K. (2004). A survey of manifold-based learning methods. In *Recent Advances in Datamining of Enterprise Data Algorithms and Applications*, pages 691 – 745.

Huynh, T., Fritz, M., and Schiele, B. (2008). Discovery of activity patterns using topic models.

Huynh, T. and Schiele, B. (2005). Analyzing features for activity recognition. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pages 159–163. ACM.

Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., and Bao, L. (2003a). A context-aware experience sampling tool. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 972–973. ACM.

Intille, S. S., Tapia, E. M., Rondoni, J., Beaudin, J., Kukla, C., Agarwal, S., Bao, L., and Larson, K. (2003b). Tools for studying behavior and technology in natural settings. In *UbiComp 2003: Ubiquitous Computing*, pages 157–174. Springer.

IOC (2011). IOC announces new events for Sochi 2014, shortlisted sports for 2020. Olympic.org, `bit.ly/WPhBGU`. accessed: March 11th, 2013.

Iwata, B. A. and Worsdell, A. S. (2005). Implications of Functional Analysis Methodology for the Design of Intervention Programs. *Exceptionality*, 13(1):25–34.

Janssen, W. G., Bussmann, H. B., and Stam, H. J. (2002). Determinants of the sit-to-stand movement: a review. *Physical Therapy*, 82(9):866–879.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer.

Judy, J. W. (2001). Microelectromechanical systems (mems): fabrication, design and applications. *Smart materials and Structures*, 10(6):1115.

Junker, H., Amft, O., Lukowicz, P., and Tröster, G. (2008). Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition*, 41:2010–2024.

Karjalainen, A. (1992). International statistical classification of diseases and related health problems (icd-10). World Health Organization.

Karmaker, A. and Kwek, S. (2006). A boosting approach to remove class label noise. *International Journal of Hybrid Intelligent Systems*, 3(3):169–177.

Katz, S. (1983). Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*.

Keijsers, N., Horstink, M., Van Hilten, J., Hoff, J., and Gielen, C. (2000). Detection and assessment of the severity of levodopa-induced dyskinesia in patients with parkinson's disease by neural networks. *Movement disorders*, 15(6):1104–1111.

Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2004). Segmenting time series: A survey and novel approach. *Data mining in time series databases*, 57:1–22.

Kidd, C. D., Orr, R., Abowd, G. D., Atkeson, C. G., Essa, I. A., MacIntyre, B., Mynatt, E., Starner, T. E., and Newstetter, W. (1999). The aware home: A living laboratory for ubiquitous computing research. In *Cooperative buildings. Integrating information, organizations, and architecture*, pages 191–198. Springer.

King, R., Atallah, L., Darzi, A., and Yang, G. (2007). An hmm framework for optimal sensor selection with applications to bsn sensor glove design. In *Proceedings of the 4th workshop on Embedded networked sensors*, pages 58–62. ACM.

King, R., Atallah, L., Lo, B., and Yang, G. (2009a). Development of a wireless sensor glove for surgical skills assessment. *Information Technology in Biomedicine, IEEE Transactions on*, 13(5):673–679.

King, R., Mcllwraith, D., Lo, B., Pansiot, J., McGregor, A., and Yang, G. (2009b). Body sensor networks for monitoring rowing technique. In *Proc. Int. Workshop on Wearable and Implantable Body Sensor Networks*, pages 251–255.

Kleinberger, T., Becker, M., Ras, E., Holzinger, A., and Müller, P. (2007). Ambient intelligence in assisted living: enable elderly people to handle future interfaces. In *Universal access in human-computer interaction. Ambient interaction*, pages 103–112. Springer.

Korel, B. T. and Koo, S. G. M. (2010). A Survey on Context-Aware Sensing for Body Sensor Networks. *Wireless Sensor Networks*, (2):571–583.

Kranz, M., Möller, A., Hammerla, N., Diewald, S., Roalter, L., Plötz, T., and Olivier, P. (2012). The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices. *Pervasive and Mobile Computing (PMC)*.

Krause, N. and Singer, Y. (2004). Leveraging the margin more carefully. In *Proceedings of the twenty-first international conference on Machine learning*, page 63. ACM.

Kunze, K., Lukowicz, P., Junker, H., and Tröster, G. (2005). Where am i: Recognizing on-body positions of wearable sensors. In *Location-and context-awareness*, pages 264–275. Springer.

Kurz, M., Holzl, G., Ferscha, A., Sagha, H., del Millán, J., and Chavarriaga, R. (2011). Dynamic quantification of activity recognition capabilities in opportunistic systems. In *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pages 1–5. IEEE.

Kwapisz, J. R., Weiss, G. M., and Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82.

Ladha, C., Hammerla, N. Y., Olivier, P., and Plötz, T. (2013). Climbax: skill assessment for climbing enthusiasts. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 235–244. ACM.

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., et al. (2005). Place lab: Device positioning using radio beacons in the wild. In *Pervasive computing*, pages 116–133. Springer.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. (2010). A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150.

Lane, N. D., Xu, Y., Lu, H., Hu, S., Choudhury, T., Campbell, A. T., and Zhao, F. (2011). Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 355–364. ACM.

Lara, O. D. and Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE*, 15(3):1192–1209.

Le Masurier, G. C., Tudor-Locke, C., et al. (2003). Comparison of pedometer and accelerometer accuracy under controlled conditions. *Medicine and Science in Sports and Exercise*, 35(5):867–871.

Lecavalier, L., Leone, S., and Wiltz, J. (2006). The impact of behaviour problems on caregiver stress in young people with autism spectrum disorders. *J. of Intellectual Disability Research*, (50):172–183.

LeCun, Y., Chopra, S., and Hadsell, R. (2006). A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM.

Lee, L. and Grimson, W. E. L. (2002). Gait analysis for recognition and classification. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 148–155. IEEE.

Lee, S., Le, H. X., Ngo, H. Q., Kim, H. I., Han, M., Lee, Y.-K., et al. (2011). Semi-markov conditional random fields for accelerometer-based activity recognition. *Applied Intelligence*, 35(2):226–241.

Lester, J., Choudhury, T., and Borriello, G. (2006). A practical approach to recognizing physical activities. In *Pervasive Computing*, pages 1–16. Springer.

Lin, J., Keogh, E., Wei, L., and Lonardi, S. (2007). Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144.

Liu, T., Inoue, Y., and Shibata, K. (2009). Development of a wearable sensor system for quantitative gait analysis. *Measurement*, 42(7):978–988.

Lo, B., Atallah, L., Aziz, O., El ElHew, M., Darzi, A., and Yang, G.-Z. (2007). Real-time pervasive monitoring for postoperative care. In *4th international workshop on wearable and implantable body sensor networks (BSN 2007)*, pages 122–127. Springer.

Logan, B., Healey, J., Philipose, M., Tapia, E. M., and Intille, S. (2007). A long-term evaluation of sensing modalities for activity recognition. In *UbiComp 2007: Ubiquitous Computing*, pages 483–500. Springer.

Long, X., Yin, B., and Aarts, R. (2009). Single-accelerometer-based daily physical activity classification. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 6107–6110. IEEE.

Longstaff, B., Reddy, S., and Estrin, D. (2010). Improving activity classification for health applications on mobile devices using active and semi-supervised learning. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on-NO PERMISSIONS*, pages 1–7. IEEE.

Lu, C.-P., Hager, G. D., and Mjolsness, E. (2000). Fast and globally convergent pose estimation from video images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(6):610–622.

Lukowicz, P., Intille, S., and Ward, J. A., editors (2010). *Proc. Int. Workshop on How To Do Good Research In Activity Recognition: Experimental methodology, performance evaluation and reproducibility.*

Lukowicz, P., Ward, J. A., Junker, H., Stäger, M., Tröster, G., Atrash, A., and Starner, T. (2004). Recognizing workshop activity using body worn microphones and accelerometers. In *Pervasive Computing*, pages 18–32. Springer.

Ly, D. L., Paprotski, V., and Yen, D. (2008). Neural networks on gpus: Restricted boltzmann machines. *see http://www. eecg. toronto. edu/˜ moshovos/CUDA08/doku. php*.

Machalicek, W., O'Reilly, M., Beretvas, N., Sigafoos, J., and Lancioni, G. (2007). A review of interventions to reduce challenging behavior in school settings for students with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 1(3):229–246.

Madsen, M. S., Ritter, M. A., Morris, H. H., Meding, J. B., Berend, M. E., Faris, P. M., and Vardaxis, V. G. (2004). The effect of total hip arthroplasty surgical approach on gait. *Journal of Orthopaedic Research*, 22(1):44–50.

Maetzler, W., Domingos, J., Srulijes, K., Ferreira, J. J., and Bloem, B. R. (2013). Quantitative wearable sensors for objective assessment of parkinson's disease. *Movement Disorders*, 28(12):1628–1637.

Mannini, A. and Sabatini, A. M. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*, 10(2):1154–1175.

Mantyjarvi, J., Himberg, J., and Seppanen, T. (2001). Recognizing human motion with multiple acceleration sensors. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 2, pages 747–752. IEEE.

Martin, P. and Bateson, P. (1993). *Measuring behaviour: an introductory guide*. Cambridge University Press.

Matson, J. and Lovullo, S. (2008). A review of behavioral treatments for self-injurious behaviors of persons with autism spectrum disorders. *Behavior Modification*, 32(1):61–76.

Maurer, U., Smailagic, A., Siewiorek, D., and Deisher, M. (2006). Activity recognition and monitoring using multiple sensors on different body positions.

McNaney, R., Lindsay, S., Ladha, K., Ladha, C., Schofield, G., Plötz, T., Hammerla, N., Jackson, D., Walker, R., Miller, N., and Olivier, P. (2011). Cueing Swallowing in Parkinson's Disease. In *Proc. ACM CHI Conference on Human Factors in Computing Systems*, Vancouver, Canada.

Menache, S. (1998). Dogs & Human Beings: A Story of Friendship. *Society & Animals*, 1(6):67–86.

Mera, T. O., Heldman, D. A., Espay, A. J., Payne, M., and Giuffrida, J. P. (2012). Feasibility of home-based automated parkinson's disease motor assessment. *Journal of neuroscience methods*, 203(1):152–156.

Mermier, C. M., Robergs, R. A., McMinn, S. M., and Heyward, V. H. (1997). Energy expenditure and physiological responses during indoor rock climbing. *British journal of sports medicine*, 31(3):224–228.

Michahelles, F. and Schiele, B. (2005). Sensing and monitoring professional skiers. *IEEE Pervasive Computing*, pages 40–46.

Mika, S., Schölkopf, B., Smola, A. J., Müller, K.-R., Scholz, M., and Rätsch, G. (1998). Kernel pca and de-noising in feature spaces. In *NIPS*, volume 11, pages 536–542.

Min, C.-H. and Tewfik, A. H. (2010a). Automatic characterization and detection of behavioral patterns using linear predictive coding of accelerometer sensor data. In *Proc. Int. Conf. Engineering in Medicine and Biology*.

Min, C.-H. and Tewfik, A. H. (2010b). Novel pattern detection in children with Autism Spectrum Disorder using Iterative Subspace Identification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*.

Minnen, D., Starner, T., Essa, M., and Isbell, C. (2007). Discovering characteristic actions from on-body sensor data. In *Wearable Computers, 2006 10th IEEE International Symposium on*, pages 11–18. IEEE.

Minnen, D., Westeyn, T., Starner, T., Ward, J., and Lukowicz, P. (2006). Performance metrics and evaluation issues for continuous activity recognition. *Performance Metrics for Intelligent Systems*, page 4.

Moeslund, T., Hilton, A., and Kruger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126.

Möller, A., Roalter, L., Diewald, S., Kranz, M., Hammerla, N., Olivier, P., and Plötz, T. (2012). GymSkill: A Personal Trainer for Physical Exercises. In *Proc. Int. Conf. Pervasive Computing and Communications (PerCom)*.

Mulroy, S., Gronley, J., Weiss, W., Newsam, C., and Perry, J. (2003). Use of cluster analysis for gait pattern classification of patients in the early and late recovery phases following stroke. *Gait & posture*, 18(1):114–125.

Murgia, A., Kyberd, P. J., Chappell, P. H., and Light, C. M. (2004). Marker placement to describe the wrist movements during activities of daily living in cyclical tasks. *Clinical Biomechanics*, 19(3):248–254.

Murphy-Chutorian, E. and Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626.

Murray, J. K., Browne, W. J., Roberts, M. A., Whitmarsh, A., and Gruffydd-Jones, T. J. (2010). Number and ownership profiles of cats and dogs in the UK. *Veterinary Record*, 166(6):163–168.

Oliver, N., Horvitz, E., and Garg, A. (2002). Layered representations for human activity recognition. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 3–8. IEEE.

Olivier, P., Monk, A., Xu, G., and Hoey, J. (2009). Ambient kitchen: Designing situated services using a high fidelity prototyping environment. In *Workshop on Affect & Behaviour Related Assistance in the Support of the Elderly, PETRA-09*.

Openmovement (2013). Openmovement sensing platform. `www.openmovement.co.uk`. accessed: March 11th, 2013.

Padoy, N., Blum, T., Feussner, H., Berger, M.-O., and Navab, N. (2008). On-line recognition of surgical activity for monitoring in the operating room. In *AAAI*, pages 1718–1724.

Palmer, J., Coats, M., Roe, C., Hanko, S., Xiong, C., and Morris, J. (2010). Unified parkinson's disease rating scale-motor exam: inter-rater reliability of advanced practice nurse and neurologist assessments. *Journal of advanced nursing*, 66(6):1382–1387.

Pansiot, J., King, R. C., McIlwraith, D. G., and Lo, B. P. L. (2008). ClimBSN: Climber performance monitoring with BSN. In *Proc. 5th Int. Summer School and Symposium on Medical Devices and Biosensors*.

Patel, M. R., Carr, J. E., Kim, C., Robles, A., and Eastridge, D. (2000). Functional analysis of aberrant behavior maintained by automatic reinforcement: Assessments of specific sensory reinforcers. *Research in Developmental Disabilities*, 21(5):393–407.

Patel, S., Lorincz, K., Hughes, R., Huggins, N., Growdon, J., Standaert, D., Akay, M., Dy, J., Welsh, M., and Bonato, P. (2009). Monitoring motor fluctuations in patients with parkinson's disease using wearable sensors. *Information Technology in Biomedicine, IEEE Transactions on*, 13(6):864–873.

Paul, D. B. and Baker, J. M. (1992). The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics.

PetTracker (2013). PetTracker. `www.pettracker.com`; accessed: 18th March 2013.

Pham, C. and Olivier, P. (2009). Slice&dice: Recognizing food preparation activities using embedded accelerometers. *Ambient Intelligence*, pages 34–43.

Pham, C., Plötz, T., and Olivier, P. (2010). A dynamic time warping approach to real-time activity recognition for food preparation. *Ambient Intelligence*, pages 21–30.

Pirttikangas, S., Fujinami, K., and Nakajima, T. (2006). Feature selection and activity recognition from wearable sensors. In *Ubiquitous Computing Systems*, pages 516–527. Springer.

Plötz, T., Chen, C., Hammerla, N. Y., and Abowd, G. D. (2012a). Automatic Synchronization of Wearable Sensors and Video-Cameras for Ground Truth Annotation – A Practical Approach. In *Proc. Int. Symp. Wearable Computing (ISWC)*.

Plötz, T. and Fink, G. A. (2009). Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(4):269–298.

Plötz, T., Hammerla, N., Rozga, A., and Reavis, A. (2012b). Automatic Assessment of Problem Behavior in Individuals with Developmental Disabilities. In *Proc. Int. Conf. Ubiquitous Comp. (UbiComp)*.

Plötz, T., Hammerla, N. Y., and Olivier, P. (2011a). Feature learning for activity recognition in ubiquitous computing. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*, pages 1729–1734. AAAI Press.

Plötz, T., Moynihan, P., Pham, C., and Olivier, P. (2011b). Activity recognition and healthier food preparation. In *Activity Recognition in Pervasive Intelligent Environments*, pages 313–329. Springer.

Pons-Moll, G., Baak, A., Helten, T., Muller, M., Seidel, H.-P., and Rosenhahn, B. (2010). Multisensor-fusion for 3d full-body human motion capture. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 663–670. IEEE.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990.

Poppe, R., Rienks, R., and van Dijk, B. (2007). Evaluating the future of hci: challenges for the evaluation of emerging applications. In *Artifical Intelligence for Human Computing*, pages 234–250. Springer.

Prato-Previde, E., Custance, D. M., Spiezio, C., and Sabatini, F. (2003). Is the dog-human relationship an attachment bond? an observational study using ainsworth's strange situation. *Behaviour*, 140(2):pp. 225–254.

Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., Howard, D., Meijer, K., and Crompton, R. (2009). Activity identification using body-mounted sensors – a review of classification techniques. *Physiological Measurement*, 30(4):1 – 33.

Quaine, F., Martin, L., and Blanchi, J. (1997). Effect of a leg movement on the organisation of the forces at the holds in a climbing position 3-D kinetic analysis. *Human Movement Science*, 16(2-3):337–346.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann.

Ranzato, M. A. and Hinton, G. E. (2010). Modeling Pixel Means and Covariances Using Factorized Third-Order Boltzmann Machines.

Ravi, N., Dandekar, N., Mysore, P., and Littman, M. L. (2005). Activity recognition from accelerometer data. In *AAAI*, volume 5, pages 1541–1546.

Rehorn, A. G., Jiang, J., and Orban, P. E. (2005). State-of-the-art methods and results in tool condition monitoring: a review. *The International Journal of Advanced Manufacturing Technology*, 26(7-8):693–710.

Reimer, J., Grabowski, M., Lindvall, O., and Hagell, P. (2004). Use and interpretation of on/off diaries in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(3):396–400.

Reiss, A. and Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 108–109. IEEE.

Riedhammer, K., Bocklet, T., Ghoshal, A., and Povey, D. (2012). Revisiting semi-continuous hidden markov models. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4721–4724. IEEE.

Rob O'Reilly, A. K. and Harney, K. (2014). Analog devices inc. `www.analog.com/library/analogdialogue/archives/43-02/mems_microphones.html`.

Robertson, N. and Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding*, 104(2):232–248.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., Doppler, J., Holzmann, C., Kurz, M., Holl, G., Chavarriaga, R., Creatura, M., and del R. Millan, J. (2010). Collecting complex activity data sets in highly rich networked sensor environments. In *7th Int. Conf. Networked Sensing Sys.*

Rojahn, J., Matson, J., Lott, D., Esbensen, A., and Smalls, Y. (2001). The behavior problems inventory: An instrument for the assessment of self-injury, stereotyped behavior, and aggression / destruction in individuals with developmental disabilities. *J. of Autism and Developmental Disorders*, 31(6):577–588.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Saria, S., Duchi, A., and Koller, D. (2011). Discovering deformable motifs in continuous time series data. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*, pages 1465–1471. AAAI Press.

Schilit, B. N., LaMarca, A., Borriello, G., Griswold, W. G., McDonald, D., Lazowska, E., Balachandran, A., Hong, J., and Iverson, V. (2003). Challenge: Ubiquitous location-aware computing and the place lab initiative. In *Proceedings of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 29–35. ACM.

Schmid, T., Shea, R., Friedman, J., and Srivastava, M. B. (2007). Movement analysis in rock-climbers. In *Proc. Int. Conf. Information Proc. in Sensor Networks (IPSN)*.

Schölkopf, B. and Smola, A. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond.

Senior, A., Pankanti, S., Hampapur, A., Brown, L., Tian, Y.-L., and Ekin, A. (2003). Blinkering surveillance: Enabling video privacy through computer vision. *IBM Technical Paper, RC22886 (W0308-109)*.

Serpell, J. (1995). *The Domestic Dog: its Evolution, Behaviour and Interactions with People*. Cambridge University Press.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Shieh, J. and Keogh, E. (2008). i sax: indexing and mining terabyte sized time series. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM.

Sibella, F., Frosio, I., Schena, F., and Borghese, N. (2007). 3d analysis of the body center of mass in rock climbing. *Human Movement Science*, 26(6):841 – 852.

Slyper, R. and Hodgins, J. K. (2008). Action capture with accelerometers. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 193–199. Eurographics Association.

Smith, C. D., Farrell, T. M., McNatt, S. S., and Metreveli, R. E. (2001). Assessing laparoscopic manipulative skills. *The American journal of surgery*, 181(6):547–550.

Smith, R. and Iwata, B. (1997). Antecedent influences on behavior disorders. *J. of Applied Behavior Analysis*, 30:343–375.

Sofuwa, O., Nieuwboer, A., Desloovere, K., Willems, A.-M., Chavret, F., and Jonkers, I. (2005). Quantitative gait analysis in parkinson's disease: comparison with a healthy control group. *Archives of physical medicine and rehabilitation*, 86(5):1007–1013.

Sprager, S. and Zazula, D. (2009). A cumulant-based method for gait identification using accelerometer data with principal component analysis and support vector machine. *WSEAS Transactions on Signal Processing*, 5(11):369–378.

Starner, T., Weaver, J., and Pentland, A. (1998a). Real-time american sign language recognition using desk and wearable computer based video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1371–1375.

Starner, T., Weaver, J., and Pentland, A. (1998b). A wearable computer based american sign language recognizer. In *Assistive Technology and Artificial Intelligence*, pages 84–96. Springer.

Stiefmeier, T., Roggen, D., Troster, G., Ogris, G., and Lukowicz, P. (2008). Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing*, 7(2):42.

Stikic, M., Huynh, T., Van Laerhoven, K., and Schiele, B. (2008a). ADL recognition based on the combination of RFID and accelerometer sensing. In *Proc. IEEE Pervasive Computing Technologies for Healthcare*.

Stikic, M., Van Laerhoven, K., and Schiele, B. (2008b). Exploring semi-supervised and active learning for activity recognition. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, pages 81–88. IEEE.

Strohrmann, C., Harms, H., Kappeler-Setz, C., and Troster, G. (2012). Monitoring kinematic changes with fatigue in running using body-worn sensors. *Information Technology in Biomedicine, IEEE Transactions on*, 16(5):983–990.

Strohrmann, C., Harms, H., Tröster, G., Hensler, S., and Müller, R. (2011). Out of the lab and into the woods: Kinematic analysis in running using wearable sensors. In *Proceedings of the 13th international conference on ubiquitous computing*, pages 119–122. ACM.

Sun, L., Zhang, D., Li, B., Guo, B., and Li, S. (2010). Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations. In *Ubiquitous intelligence and computing*, pages 548–562. Springer.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Tao, Y., Hu, H., and Zhou, H. (2007). Integration of vision and inertial sensors for 3d arm motion tracking in home-based rehabilitation. *The International Journal of Robotics Research*, 26(6):607–624.

Taylor, G. and Hinton, G. (2009). Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032. ACM.

Testa, M., Martin, L., and Debû, B. (1999). Effects of the type of holds and movement amplitude on postural control associated with a climbing task. *Gait & posture*, 9(1):57–64.

Thiel, C. (2008). Classification on soft labels is robust against label noise. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 65–73. Springer.

Tomkins, L. M., Thomson, P. C., and McGreevy, P. D. (2011). Behavioral and physiological predictors of guide dog success. *Journal of Veterinary Behavior: Clinical Applications and Research*, 6(3):178 – 187.

Treffner, P. and Turvey, M. (1996). Symmetry, broken symmetry, and handedness in bimanual coordination dynamics. *Experimental Brain Research*, 107(3):463–478.

Trejos, A. L., Patel, R. V., Naish, M. D., and Schlachta, C. M. (2008). Design of a sensorized instrument for skills assessment and training in minimally invasive surgery. In *Biomedical Robotics and Biomechatronics, 2008. BioRob 2008. 2nd IEEE RAS & EMBS International Conference on*, pages 965–970. IEEE.

Trost, S. G., McIver, K. L., and Pate, R. R. (2005). Conducting accelerometer-based activity assessments in field-based research. *Medicine and science in sports and exercise*, 37(11 Suppl):S531–43.

Tsipouras, M., Tzallas, A., Rigas, G., Bougia, P., Fotiadis, D., and Konitsiotis, S. (2010). Automated levodopa-induced dyskinesia assessment. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 2411–2414. IEEE.

Twight, M. F. and Martin, J. (1999). *The Extreme Alpinism: Climbing Light, Fast, and High*. The Mountaineers Books.

van Kasteren, T. and Krose, B. (2007). Bayesian activity recognition in residence for elders.

Van Kasteren, T., Noulas, A., Englebienne, G., and Kröse, B. (2008). Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 1–9. ACM.

Van Laerhoven, K. and Berlin, E. (2009). When else did this happen? efficient subsequence representation and matching for wearable activity data. In *Wearable Computers, 2009. ISWC'09. International Symposium on*, pages 101–104. IEEE.

Van Laerhoven, K. and Cakmakci, O. (2000). What shall we teach our pants? In *Wearable Computers, The Fourth International Symposium on*, pages 77–83. IEEE.

Van Laerhoven, K. and Gellersen, H.-W. (2004). Spine versus porcupine: A study in distributed wearable activity recognition. In *Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on*, volume 1, pages 142–149. IEEE.

Velloso, E., Bulling, A., Gellersen, H., Ugulino, W., and Fuks, H. (2013). Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference*, pages 116–123. ACM.

Verbeek, J., Vlassis, N., and Kröse, B. (2003). Efficient greedy learning of gaussian mixture models. *Neural computation*, 15(2):469–485.

Vlasic, D., Adelsberger, R., Vannucci, G., Barnwell, J., Gross, M., Matusik, W., and Popović, J. (2007). Practical motion capture in everyday surroundings. In *ACM Transactions on Graphics (TOG)*, volume 26, page 35. ACM.

Von Schroeder, H. P., Coutts, R. D., Lyden, P. D., Billings, E., and Nickel, V. L. (1995). Gait parameters following stroke: a practical assessment. *Journal of rehabilitation research and development*, 32:25–25.

Wagner, J., Plötz, T., Halteren, A. V., Hoonhout, J., Moynihan, P., Jackson, D., Ladha, C., Ladha, K., and Olivier, P. (2011). Towards a Pervasive Kitchen Infrastructure for Measuring Cooking Competence. In *Proc Int Conf Pervasive Computing Technologies for Healthcare (PervasiveHealth)*.

Wang, S., Yang, J., Chen, N., Chen, X., and Zhang, Q. (2005). Human activity recognition with user-free accelerometers in the sensor networks. In *Neural Networks and Brain, 2005. ICNN&B'05. International Conference on*, volume 2, pages 1212–1217. IEEE.

Ward, J., Bharatula, N., Tröster, G., and Lukowicz, P. (2002). Continuous activity recognition in the kitchen using miniaturised sensor button.

Ward, J., Lukowicz, P., and Tröster, G. (2005). Gesture spotting using wrist worn microphone and 3-axis accelerometer. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pages 99–104. ACM.

Ward, J., Lukowicz, P., Tröster, G., and Starner, T. (2006). Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1553–1567.

Ward, J. A., Lukowicz, P., and Gellersen, H. W. (2011). Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):6.

Westeyn, T., Vadas, K., Starner, T., and Abowd, G. (2005). Recognizing Mimicked Autistic Self-Stimulatory Behaviors Using HMMs. *Proc. Int. Symp. Wearable Computing*.

White, D. and Olsen, P. (2010). A time motion analysis of bouldering style competitive rock climbing. *The Journal of Strength & Conditioning*, 24(5):1356–1360.

Wilson, D. R. and Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. *Machine learning*, 38(3):257–286.

Wojek, C., Nickel, K., and Stiefelhagen, R. (2006). Activity recognition and room-level tracking in an office environment. In *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 25–30. IEEE.

Yamazaki, T. (2005). Ubiquitous home: real-life testbed for home context-aware service. In *Testbeds and Research Infrastructures for the Development of Networks and Communities, 2005. Tridentcom 2005. First International Conference on*, pages 54–59. IEEE.

Yan, Z., Subbaraju, V., Chakraborty, D., Misra, A., and Aberer, K. (2012). Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 17–24. Ieee.

Yazdi, N., Ayazi, F., and Najafi, K. (1998). Micromachined inertial sensors. *Proceedings of the IEEE*, 86(8):1640–1659.

Yogev, G., Plotnik, M., Peretz, C., Giladi, N., and Hausdorff, J. M. (2007). Gait asymmetry in patients with parkinson's disease and elderly fallers: when does the bilateral coordination of gait require attention? *Experimental brain research*, 177(3):336–346.

Zappi, P., Lombriser, C., Stiefmeier, T., Farella, E., Roggen, D., Benini, L., and Tröster, G. (2008). Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection.

Zhang, M. and Sawchuk, A. A. (2012). Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 1036–1043. ACM.

Zhou, H. and Hu, H. (2008). Human motion tracking for rehabilitation—a survey. *Biomedical Signal Processing and Control*, 3(1):1–18.

Zinnen, A., Wojek, C., and Schiele, B. (2009). Multi activity recognition based on bodymodel-derived primitives. In *Location and Context Awareness*, pages 1–18. Springer.