FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

# Behavior Modeling based on Depth Data in the Context of Ambient Assisted Living

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

**Doktor der Technischen Wissenschaften**

by

**Rainer Planinc**
Registration Number 0425163

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Priv. Doz. Dr. Martin Kampel

The dissertation has been reviewed by:

| | |
|---|---|
| (Priv. Doz. Dr. Martin Kampel) | (Prof. Dr. José Bravo Rodriguez) |

Vienna, 14th April, 2015

Rainer Planinc

# Erklärung zur Verfassung der Arbeit

Rainer Planinc
Mühlhäufelgasse 27/5/4, 1220 Vienna, Austria

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 14. April 2015

Rainer Planinc


The research presented in this thesis was partially supported by the Austrian Research Promotion Agency (FFG) and the European Union (AAL-Joint Programme) under grant AAL 2010-3-020.

# Acknowledgements

First and foremost I would like to express my gratitude to my supervisor Martin Kampel for giving me the opportunity to carry out my research on this topic, for helping me to find the right way during the last years and to make this PhD thesis possible. Additionally I thank José Bravo for his support during the writing of my thesis and the possibility to visit him and his group during my research exchange.

The last 5 years definitely had its ups and downs and without the support of Robert Sablatnig, Katharina Pois and my colleagues this thesis would not have been possible. My appreciation for all the chats goes to all my (former) colleagues at CVL - they made the last years enjoyable and supported me when I felt lost. I would especially like to thank Michael Hödlmoser for giving me the initial hint where to go and his motivation during the last years. I am also very thankful to the CVL darts team for the regular screen breaks: Andreas Zweng, Sebastian Zambanini, Michael Hödlmoser, Florian Kleber and Albert Kavelar. A big thanks also to Melanie Gau and Fabian Hollaus for their permit to play darts behind their back. A good working atmosphere is the most important part within each job - I am grateful to Stefan Fiel and Elisabeth Wetzinger for always having a coffee and a chat with me. Moreover, I am really thankful to Florian Kleber and Markus Diem for having a coffee with me during cold and rainy days. During the last years I not only got the opportunity to work on my PhD, but I also learned how to solve the Rubik's Cube - thanks to Andreas Zweng (our new neighbor) for pimping my cube!

Additionally I am very thankfully to Ashley and Jeff, real friends I met during one of my travels. Thank you for sharing your house with me, for all the good times we had and thank you so much for proof-reading my thesis!

I am especially grateful to my parents Gertrude and Otto, for their endless support throughout my whole life and for their encouragement. Without their support I would not have made this all. Moreover I want to thank my whole family for their support.

Most importantly, I am deeply grateful to my beloved Rebecca since she always had an open ear for me and always supported me, even though it was not ever easy. She motivated me during bad times and always enjoyed a glass of wine with me during good times. Although I was traveling many times during my PhD, her support and understanding is endless and whenever possible, she explored the world with me - thank you so much!

# Kurzfassung

Das Modellieren von Verhalten ist eine aufstrebende Thematik im Bereich der Bildverarbeitung. In Kombination mit 3D Sensoren handelt es sich hierbei um eine Methodik, welche im Kontext des Ambient Assisted Living Anwendung findet. Die Hauptziele hierbei sind die Unterstützung von älteren Personen durch den Ensatz von Technik, wobei im Besonderen das Erkennen von gefährlichen Situationen und "abnormalem" Verhalten im Vordergrund stehen. Die Definition von "normalem" und "abnormalem" Verhalten ist jedoch stark kontext abhängig und wird daher in dieser Arbeit diskutiert.

In dieser Dissertation wird ein neues räumlich-zeitliches Modell zur Analyse von Verhalten präsentiert. Dadurch wird es möglich, das Verhalten über die Zeit zu modellieren und Anomalien im Tagesablauf zu detektieren (mittelfristig, d.h. im Laufe des Tages). Durch die Detektion von gefährlichen Situationen, welche innerhalb von Minuten auftreten und den normalen Tagesablauf unterbrechen, wird das Modell erweitert. Zusätzlich erfolgen Langzeit-Analysen, um schleichende Veränderungen der Mobilität detektieren zu können. Durch die Kombination dieser Aspekte wird ein umfassendes Modell entwickelt, wobei die Modellierung räumlicher Aspekte dabei auf einem neuartigen Verfahren, welches das Szenenverständnis auf Basis der Bewegung von Menschen ermöglicht, basiert. Da die hierzu verwendeten Tracking Daten jedoch fehlerbehaftet sind, werden die Daten vor der Analyse gefiltert. Zeitliche Abläufe werden mit Hilfe von Aktivitäts-Histogrammen modelliert, wodurch Veränderungen im Tagesablauf von älteren Menschen detektiert werden. Da es sich bei dem vorgestellten Verfahren um ein selbst-lernendes System handelt, müssen keinerlei Informationen oder annotierte Trainingsdaten zur Verfügung stehen.

Um die Vielseitigkeit der Bildverarbeitung optimal einsetzen zu können, basiert das vorgestellte Verfahren auf Tracking Daten, welche mit Hilfe eines 3D Sensors erfasst werden. Da das vorgestellte Verfahren in den privaten Räumlichkeiten von älteren Personen eingesetzt wird, spielt die Privatsphäre eine große Rolle. Daher werden keinerlei RGB Bilder erfasst, noch wird das Tiefenbild des 3D Sensors abgespeichert - eine Analyse der Daten erfolgt daher in Echtzeit.

Die Leistung der vorgestellten Algorithmen wird auf Basis von drei unterschiedlichen Datensätzen evaluiert, welche im Kontext von Ambient Assisted Living erstellt wurden. Die Ergebnisse zeigen, dass der Einsatz des vorgestellten Modells eine detaillierte Analyse des Verhaltens ermöglicht. Obwohl die Evaluierung primär im Kontext von Ambient Assisted Living durchgeführt wurde, können die vorgestellten Methoden auch in anderen Bereichen (z.B. innerhalb eines Büros) angewandt werden.

# Abstract

Behavior modeling is an upcoming area within the field of computer vision. In combination with the evolving development of 3D sensor technologies, both, software and hardware motivate the application of behavior modeling within the context of Ambient Assisted Living. The main goals are the support of elderly people during their daily routines, the detection of critical events and "abnormal" behavior in order to provide immediate help. However, the definition of "normal" and "abnormal" behavior depends on the context and thus needs to be discussed.

Within this thesis, a novel spatio-temporal behavior model is introduced by incorporating spatial and temporal knowledge into one behavior model. This allows to model the behavior over time and detect abnormal behavior during daily routines on the mid-term range, i.e. during the day. In order to extend the proposed model, short-term information (i.e. time frame of minutes) is integrated by detecting critical events, interrupting daily activities of the elderly. Due to the analysis of mobility changes over the duration of months (long-term), a holistic behavior model is proposed. Spatial modeling is enhanced by the introduction of a human-centered scene understanding approach, focusing on scene functionalities and is solely based on long-term tracking information. Since tracking data is noisy, pre-processing steps to filter outliers are introduced, before the scene is modeled. Temporal aspects are modeled by the use of activity histograms, allowing to detect deviations within the behavior of elderly people. In combination, the proposed model allows to detect abnormal behavior based on the time of the day as well as the location, using an unsupervised learning approach. Hence, no prior knowledge of the scene needs to be specified since the model adapts to the scene automatically.

In order to benefit from the flexibility of computer vision approaches, the behavior model is obtained from tracking data, based on a single 3D sensor providing depth information. Since the proposed approach is applied to homes of elderly people, privacy aspects need to be considered. No RGB video data is used and only depth data is analyzed in real-time, hence the depth stream is not recorded.

The performance of the proposed approaches is evaluated on three datasets within the context of Ambient Assisted Living and results show that the use of the proposed system is feasible and provides detailed analysis of the elderly's behavior. Although the evaluation is mainly based on this context, proposed approaches can be applied to different contexts as well (e.g. within an office).

# Contents

CHAPTER 1

# Introduction

Behavior modeling is an emerging topic within the field of computer vision, enabled by an increasing number of cameras and sensors and the lack of human resources to monitor this huge amount of data. Holistic behavior modeling is only applied in the field of visual surveillance, although other fields can also benefit from these approaches. Ambient Assisted Living (AAL) supports the elderly to live independently in their own homes as long as possible, supporting them by technology. The motivation for using technology is based on the increased life expectancy of humans, whereas the number of births is decreasing - resulting in the demographic change [United Nations, Department of Economic and Social Affairs, 2001]. Thus results in an increased need of caretaker resources and increased costs in the health care system. AAL provides countermeasures in order to reduce costs and caretaker resources by supporting elderly people to stay independently, detecting critical events and critical health conditions as early as possible in order to provide immediate help. Facilitating countermeasures at an early stage results in an earlier treatment of diseases, reduced rehabilitation time and a higher quality of life for the elderly. Moreover, caretaker and the health care system are unburdened.

Advances in the development of 3D sensors motivate their use instead of cameras or wearable sensors, since they provide advantages like privacy protection and improved robustness in comparison to cameras. However, camera systems and wearable sensors are well-established, hence the use of a new 3D sensor technology for behavior modeling within the context of AAL is proposed, and strengths and weaknesses of this technology are discussed. In order to analyze the behavior of elderly people, a novel behavior model within the field of computer vision is introduced and advances within the areas of critical event detection, inactivity modeling and scene understanding are proposed. The combination of the behavior model and the availability of 3D sensors motivates the introduction of new approaches together with new sensors within the field of AAL.

Approaches within the context of AAL focus on specific tasks (e.g. support during the hand-washing process [Hoey et al., 2010]) or events (e.g. fall detection [Xinguo, 2008]). However, a holistic behavior model combining the detection of critical events

Table 1.1: Differentiation between short-term, mind-term and long-term

| Term | Time lapse | Example |
|---|---|---|
| Short-term | Seconds, minutes | Falling on the floor |
| Mid-term | Hours, days | Irregularities within Activities of Daily Living (ADL) |
| Long-term | Weeks, months | Deterioration of the health status |

and detecting deterioration of the health status ensures to detect issues already at an early stage (e.g. the beginning of dementia) and thus allows to provide effective countermeasures. In this thesis, a holistic spatio-temporal behavior model is proposed, combining short-term, mid-term and long-term analysis, where Table 1.1 defines the time lapse and illustrates examples.

The detection of critical events (e.g. falls) integrates short-term information into the behavior model, based on the duration of minutes. Critical events like falls interrupt daily routines of the elderly and immediate help needs to be provided, if the person is not able to recover from the incident on their own. Moreover, fear of falling also changes the behavior of elderly people dramatically [Deshpande et al., 2009, Howland et al., 1998]. Modeling the behavior on the time frame of days (mid-term) allows to detect deviations within the daily routines of the elderly. Daily routines are strongly established in humans [Gallimore and Lopez, 2002], hence deviations from these routines can indicate a change of the overall health status. The proposed model incorporates both, spatial and temporal knowledge in order to analyze behavior depending on the specific context (time and location). Thus allows a fine grained analysis of the daily routines of elderly people, in order to accurately model the behavior. Moreover, slow changes of mobility on the long-term (i.e. over the duration of months) is connected with the state of health of the elderly. Reduced mobility is a consequence of a deteriorated health status and leads to decreased physical strength and, in the worst case, to social isolation. On the other hand, increased mobility, especially during the night, indicates the beginning of dementia. In order to propose a holistic behavior model, spatial and temporal aspects need to be considered while analyzing short-term, mid-term as well as long-term behavior, since deviations within each time frame need to be detected and countermeasures have to be taken immediately.

The aim of the proposed behavior modeling framework is to model "normal" behavior and detect "abnormal" behavior, i.e. situations where help is needed. But what is considered as "normal" behavior, what as "abnormal"? This distinction is based on established social norms within our society, describing the expectations of other people from us in specific situations [Cialdini and Trost, 1998, Aarts and Dijksterhuis, 2003]. Hence, within a given situation, if we behave as it is expected from us, we behave "normally". On the other hand, if we behave differently than what is expected from us, we behave "abnormally". This can be illustrated with an example of people's behavior

visiting a library: being silent is a social norm, established within our society. If we (as visitors) are quiet, we fulfill the expectations and thus act "normally". However, it is not expected to scream and run within a library - hence, this would be considered as "abnormal" behavior. However, the boundaries between "normal" and "abnormal" are fuzzy, thus it is not always possible to clearly classify behavior into either "normal" or "abnormal" - e.g. the volume of the voice in a library may be between "silent" and "loud" and thus, a clear classification is not possible. Moreover, the definition of "being silent" is not made explicitly (e.g. level of decibel), but is established implicitly within our society.

Additionally, the expectation is not only dependent on the specific situation, but also on the role of the person within a given context. During a fire, people are expected to try to escape from the fire and thus fleeing is "normal" behavior. On the other hand, firefighter are expected to fight the fire (instead of fleeing) and thus expectations are different, although the context is the same.

Although this example demonstrates the definition of "normal" and "abnormal" behavior, the definition within the context of AAL is more complicated. The definition of behavior depends on the given context and specific situation. However, for this thesis "normal" behavior is considered to be trained over time, thus modeling daily routines of elderly people. On the other hand, "abnormal" can be defined as significant (depending on the context) deviations from the trained "normal" behavior. Although this definition is well defined from a technical point of view, questions may arise from a sociological point of view. The distinction between "normal" and "abnormal" might be technically easy, but not sound sociologically and thus the proposed definition of "normal" and "abnormal" should be used with caution.

## 1.1 Motivation

The use of computer vision in the context of AAL is feasible since unobtrusive systems can be developed, not requiring the user to wear any kind of device or sensor. Especially when not focusing on a single task or activity, the advantages of computer vision are its flexibility and adaptability in comparison to other sensor types (e.g. accelerometers, pressure sensors). Hence, with a single sensor, different applications can be developed. The modeling of behavior on the short-, mid- and long-term allows to analyze the health state of an elderly person and to detect changes over time. Behavior monitoring systems detecting abnormal behavior are widely used in visual surveillance systems. However, within the context of AAL, either spatial or temporal aspects are modeled - but no unsupervised combination of both is proposed so far. Hence, the aim of this thesis is to propose a novel spatio-temporal behavior model, incorporating spatial and temporal knowledge at the same time, allowing to more accurately model the behavior without the need of prior knowledge. The proposed behavior model is evaluated within the context of AAL, although it can be applied to a different context as well. The thesis is comprised of the following goals in order to provide a holistic behavior model:

- **Spatio-temporal behavior model:** current approaches either focus on spatial or temporal aspects, but do not combine both domains in an unsupervised manner. Hence, proposed behavior models are only modeling the behavior over time, or, on the other hand, model sequences within activities (e.g. zone 1 is visited after zone 2), not considering the time of the day. In both cases, only parts of available information (location or time) are obtained and hence the combination of the knowledge within both domains is promising and results in the behavior monitoring on mid-term duration, i.e. to detect deviations of activities within the day.

- **Short-term deviations:** in order to model and detect deviations also on the short-term (duration within minutes), the detection of critical events is proposed, allowing to detect interruptions within the ADL immediately. By integrating a fall detection approach into the proposed behavior model, abnormal behavior can be detected on both, short-term (time frame of minutes) as well as mid-term (time frame of days).

- **Long-term deviations:** to provide a holistic view of the modeled behavior, not only short- and mid-term information is modeled, but also long-term deviations over the duration of months needs to be detected. Self-adapting systems allow to adapt to changes in the elderly's life, but if slow and continuous changes of the health status over several months occur, systems are not able to detect abnormalities, since only fast changes (on daily basis) within the behavior are modeled. Hence, long-term changes in mobility need to be modeled in order to detect deteriorations of the health condition, e.g. enhanced mobility during the night due to an early stage of dementia.

## 1.2 Summary of Contributions

The main contribution of this thesis is the combination of spatial and temporal knowledge, in order to obtain a holistic behavior model. Moreover, the introduced behavior model is extended by incorporating also short-term information as well as modeling the behavior on the long-term in order to detect critical events and changes in mobility respectively. Since the use of a 3D sensor is proposed, challenges from this sensor type arise and approaches to tackle these issues are introduced in this thesis.

The contributions within different areas of computer vision can be summarized as follows, where results have been previously published:

- **Spatio-temporal behavior model:** literature either deals with the modeling of spatial *or* temporal aspects over time, but no combination of this knowledge is proposed. Analyzing the behavior only considering *spatial* aspects of the scene results in the detection of abnormal behavior during pre-defined activities [Hoey et al., 2010,Nguyen et al., 2005,Chung and Liu, 2008]. However, the number of these activities is highly restricted, since each activity need to be modeled accordingly and hence, an exact definition of each activity is needed. Moreover, even when spatial

information is obtained automatically, only the sequence of visiting different areas is considered, not monitoring the exact temporal behavior over time [McKenna and Charif, 2005]. *Temporal* modeling of the behavior results in models describing the behavior over time, but not incorporating spatial knowledge [Floeck and Litz, 2008, Cuddihy et al., 2007]. These models do not consider spatial singularities but model the behavior independently. Hence, the goal of this thesis is the combination of temporal and spatial knowledge to consider spatial singularities while combining it with temporal knowledge in order to propose a location sensitive behavior modeling, allowing an in depth analysis of the person's behavior without the need of prior knowledge. Although behavior is defined as the analysis of activities on the long-term, containing a high degree of semantics [Chaaraoui et al., 2012], the combination of spatial and temporal aspects allows to detect deviations on the mid-term, i.e. comparisons between activities and a trained behavior model on a time frame of a single day.

The use of a 3D sensor is introduced in order to obtain a holistic behavior model. Using this sensor results in 3D depth data of the scene and humans within the scene. Hence, this additional information is exploited to obtain temporal and spatial knowledge. However, using the 3D sensor also results in limitations, e.g. only indoor environments can be monitored and the sensor range is limited. Moreover, tracking algorithms are optimized for active interaction with the sensor in order to provide robust results.

The combination of spatial and temporal information was published in [Planinc and Kampel, 2014a] and [Planinc and Kampel, 2015b], introducing a novel approach to model temporal behavior while considering spatial knowledge within the scene by detecting Region of Interest (ROI). Moreover, temporal models are not applied on a global level, but locally within the detected ROI.

- **Temporal Modeling:** temporal modeling is based on the detection of inactivity, modeled from sensor data (e.g. motion and door sensors) [Cuddihy et al., 2007, Floeck and Litz, 2008]. However, when using a vision-based approach, the construction of inactivity profiles does not yield in accurate results. Hence, the use of histogram comparisons to obtain robust results is proposed in [Planinc and Kampel, 2014b]. Instead of applying a threshold to the inactivity profile, activity is modeled using histograms and compared to a reference histogram, where deviations indicate abnormal behavior.

- **Spatial Modeling:** in contrast to geometric or object-centered scene understanding, a human-centered scene understanding approach using continuous depth information over long-term is proposed [Planinc and Kampel, 2015a]. In contrast to state-of-the-art, the proposed method does not incorporate geometric aspects, but only long-term tracking information of humans in order to model a scene according to its functionalities offering for humans (e.g. people can sit or walk in an area). Since tracking data is noisy and literature state that it cannot be used for the

robust modeling of area functionalities, filtering mechanisms based on the pose of the person are incorporated in order to effectively filter outlier [Planinc and Kampel, 2014c]. Due to long-term tracking information and the proposed filtering mechanisms, the ROI within a scene are robustly modeled, solely based on human tracking information.

- **Detection of critical events:** critical events interrupt daily activities of elderly people and are thus of high interest, since immediate help need to be provided. Moreover, by the detection of critical events (time frame of minutes), short-term knowledge is integrated into the proposed spatio-temporal behavior model. The fear of falling is a prevalent fear amongst the elderly, more than 50% suffer from the fear of falling [Deshpande et al., 2009, Howland et al., 1998]. In combination with a high mortality rate of fallers [Wild et al., 1981], falls are a major risk for an elderly person. Since the elderly are not able to react to emergency situations properly (e.g. due to fainting), an automatic fall detection system is proposed. In order to robustly detect falls of elderly people, a 3D sensor is used in order to track humans and detect falls. This is achieved by incorporating pose information of the person and modeling the major body orientation as well as the distance to the ground floor. Work within this area is introduced in [Planinc and Kampel, 2012a] and [Planinc and Kampel, 2012b].

- **Detection of long-term changes in mobility:** in order to detect long-term changes in mobility, analysis of mobility changes over time are incorporated into to proposed behavior model to provide a spatio-temporal model, covering short-term, mid-term and long-term aspects, being applied and evaluated in the context of AAL. In order to detect slow changes of mobility, reference profiles of the mobility are obtained and compared on the long-term (i.e. the duration of months), in order to detect trends in mobility [Planinc and Kampel, 2013]. These trends can result from early stages of dementia, being reflected by an increased mobility, especially during the night. Moreover, physical problems result in reduced mobility and thus the proposed approach is able to accurately model both, increased as well as decreased mobility.

## 1.3 Thesis Structure

Within this thesis, the background and related work is described before the proposed behavior model is introduced and evaluated. Finally, a conclusion to summarize the thesis is drawn. In more detail, the rest of this work is structured as follows:

**Chapter 2** introduces the background and describes the context of AAL. The aging population as well as the demographic change is described and motivate the introduction of technology in order to support the elderly and to enable them to stay independently as long as possible. An overview of the research results within the field of AAL is presented, since different technologies and approaches are used to provide support.

Since vision-based approaches are flexible and unobtrusive, the requirements for vision-based sensors are discussed, focusing on privacy aspects and introducing 3D sensors. Related work within the field of the *detection of critical events (fall detection)*, *scene understanding* as well as *behavior monitoring* are introduced and discussed. An overview of *fall detection* approaches using various sensor types is presented and the advantages and disadvantages of different sensor types are discussed. Vision-based fall detection approaches are discussed in more detail in order to provide a holistic overview. *Scene understanding* can be classified into object-centered and human-centered approaches. Related work in both areas are presented, the use of human-centered approaches is motivated and related work within this area is summarized. Finally, the definition for *behavior* being used throughout the thesis is introduced and interpretations of behavior analysis in various contexts are discussed. Moreover, the classification between activity recognition and behavior modeling is described in order to motivate the introduction of a behavior model.

**Chapter 3** describes the proposed spatio-temporal behavior model. Starting from a basic behavior model by combining spatial and temporal knowledge, the model is successively extended. The basic behavior model is introduced at the beginning, modeling the ROI of a scene and introducing local temporal behavior modeling, depending on the detected ROI and focusing on the detection of abnormal behavior on mid-term range. In order to add short-term behavior analysis to the proposed model, the detection of critical events is introduced and the features *body orientation* and *spine distance* are proposed. In combination with the detection of the ground floor, these features are used to calculate the major body orientation and the distance to the ground floor in order to detect falls. Long-term monitoring is incorporated into the proposed behavior model by the monitoring of mobility changes over the duration of months and detecting changes within the adapted temporal model over time. Thus allows to detect slow mobility changes, since related work proposes the adaption to new behavior, without performing long-term analysis. Improved accuracy and robustness is obtained by the introduction of a novel temporal behavior model, proposing activity modeling using histograms instead of calculating inactivity profiles. Deviations are detected by comparing the histogram to a pre-trained reference histogram using histogram comparison metrics. The spatial detection of ROI is enhanced by the introduction of a human-centered scene understanding approach, solely based on noisy tracking information. To be able to accurately model functional areas within the scene, tracking data need to be pre-processed to remove outliers in a first step. Sitting and walking areas are modeled by clustering the Center of Mass (CoM) of the tracking data and applying a Kernel Density Estimation (KDE) in order to detect hotspots within the scene. The combination of all proposed approaches results in the advanced spatio-temporal behavior model, introduced at the end of this chapter.

**Chapter 4** presents the evaluation and results of the proposed approaches. Results of the basic behavior model are indicating that the use of a spatio-temporal behavior model allows to model activity in more detail, when only focusing on specific ROI. Due to modeling behavior within regions, more robust results are obtained. Moreover, the performance of the proposed fall detection approaches are evaluated and compared to

state-of-the-art approaches. The results of the long-term temporal modeling approach illustrates the feasibility of this approach, being able to detect slow changes in the elderly's mobility. The introduction of activity based histogram comparisons result in an increased performance in comparison to the use of inactivity profiles. Analyzes of the human-centered scene understanding approach in order to model functional areas of the scene show promising results, although the performance depends on the scene. Hence, quantitative as well as qualitative results are discussed in detail. Finally, the combination of the proposed approaches is evaluated, indicating that behavior in sitting areas is more structured than in walking areas.

**Chapter 5** concludes the introduced approaches and discusses the results. The application of the proposed approach within the context of AAL is described and advantages as well as drawbacks are summarized. Moreover, limitations and challenges when using the proposed method are discussed and future work is summarized in order to overcome these challenges.

CHAPTER 2

# Background and Related Work

Studies show that the aging population raises challenges for the society [United Nations, Department of Economic and Social Affairs, 2001], caused by the demographic change. The demographic change and its implications on technology are discussed, motivating the context of AAL, providing technical solutions in order to overcome the barriers raised by the aging population. The use of 3D sensors within the context of AAL is motivated and privacy aspects are described. The definition of behavior within this work is discussed and related work within the areas of critical event detection, scene understanding and behavior monitoring are summarized in order to motivate the introduction of a behavior model framework by identifying limitations of current approaches. Although the main focus is within the context of AAL, related work and the proposed approaches are generic approaches - hence, they can be applied within the context of AAL, but also within other contexts as well.

## 2.1 Demographic Change and Ambient Assisted Living (AAL)

Due to the aging population and the demographic change, the age distribution of the world will change dramatically, according to the United Nations [United Nations, Department of Economic and Social Affairs, 2001]. This is caused by an increased life expectancy and a reduction of births, changing the age distribution, illustrated in Figure 2.1. Figure 2.1a depicts the changes in the years from 1950 until 2050, where the percentage of younger people (<15 years old) is reduced, whereas the group of older people with the age of 60+ will increase worldwide. These changes results in a higher number of people being in the need of care and, on the other hand, a reduced number of caregiver. The implications of this discrepancy can be minimized by the use of technology within the field of AAL, supporting both, the elderly and the caregiver.

Although the demographic change is a global phenomenon, its implications depend on the stage of development of each country. In order to provide detailed analysis and predict future trends, the distinction between least, less and more developed regions of the world depending on demographic and socio-economic characteristics is introduced by the United Nations [United Nations, Department of Economic and Social Affairs, 2013]:

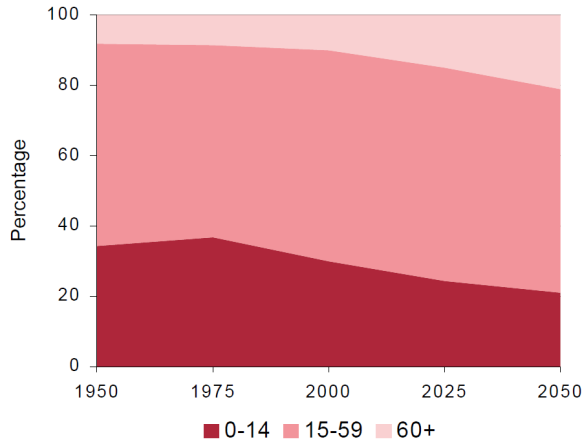
- "The group of **least developed** countries, as defined by the United Nations General Assembly in its resolutions (59/209, 59/210 and 60/33) in 2007, comprises 49 countries." More details can be found in the Annex II of [United Nations, Department of Economic and Social Affairs, 2013].
- "The **less developed regions** include all regions of Africa, Asia (excluding Japan), Latin America and the Caribbean, and Oceania (excluding Australia and New Zealand)."
- "The **more developed regions** include all other regions plus the three countries excluded from the less developed regions."

In the following, differences between countries within different development stages and worldwide trends are provided. The change of age distribution based on the classification into more, less and least developed regions of the world is illustrated in Figure 2.1b: this figure depicts the ratio of people older than 65 and children younger than 15. As can be clearly seen, this ratio changes worldwide, although the shift occurs much faster in more developed countries. However, the changes in age distribution will accelerate in less developed regions within the next 50 years [United Nations, Department of Economic and Social Affairs, 2001].

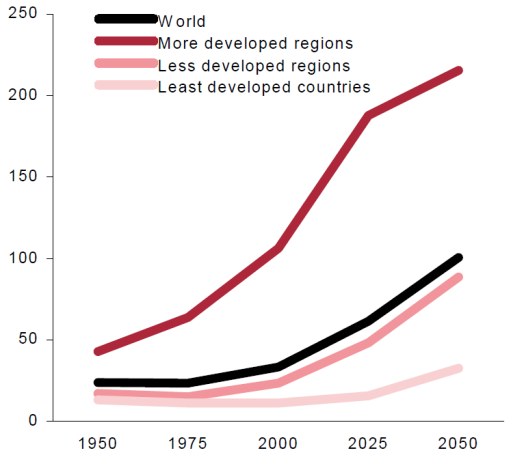
Moreover, the World Population Ageing report from the United Nations in 2013 [United Nations, Department of Economic and Social Affairs, 2013] reveals dramatic numbers according to the aging population:

- Japan is the country with the highest percentage of population being 60 years and older (32%).
- Japan is followed by Italy (26.9%) and Germany, where 26.8% of the population is above 60 years old.
- Austria is placed on rank 17, where 23.5% of the population is aged 60 years and older.
- The United Arab Emirates are placed on the last rank, indicating that only 0.9% of their population is above the age of 60 years.

Aging pyramids illustrates the predicted demographic change from 1970 to 2050 (Figure 2.2 and 2.3). Although the change of age distribution proceeds faster in more developed regions, also less developed regions will experience this change. Figure 2.2 and 2.3 depicts change of age groups from 1970 to 2013 and the prediction of 2050. The

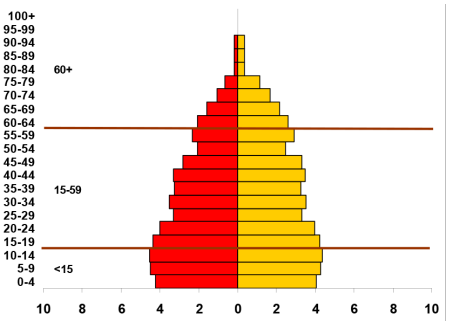



(a) Distribution of age groups worldwide (1950-2050)

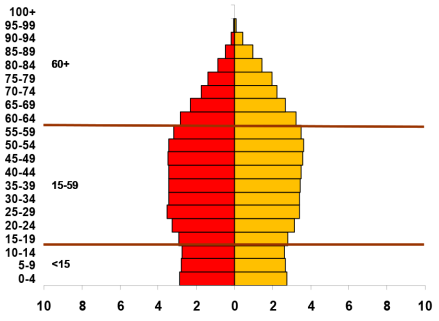



(b) Persons 65 or older per hundred children under 15 (1950-2050)

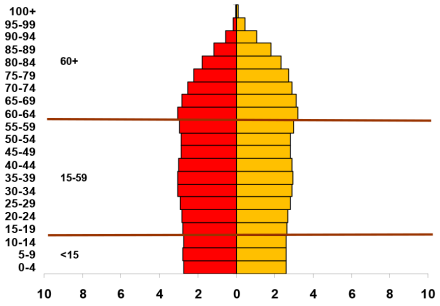
Figure 2.1: Change of age distribution from 1950-2050 [United Nations, Department of Economic and Social Affairs, 2001]




(a) Population distribution 1970

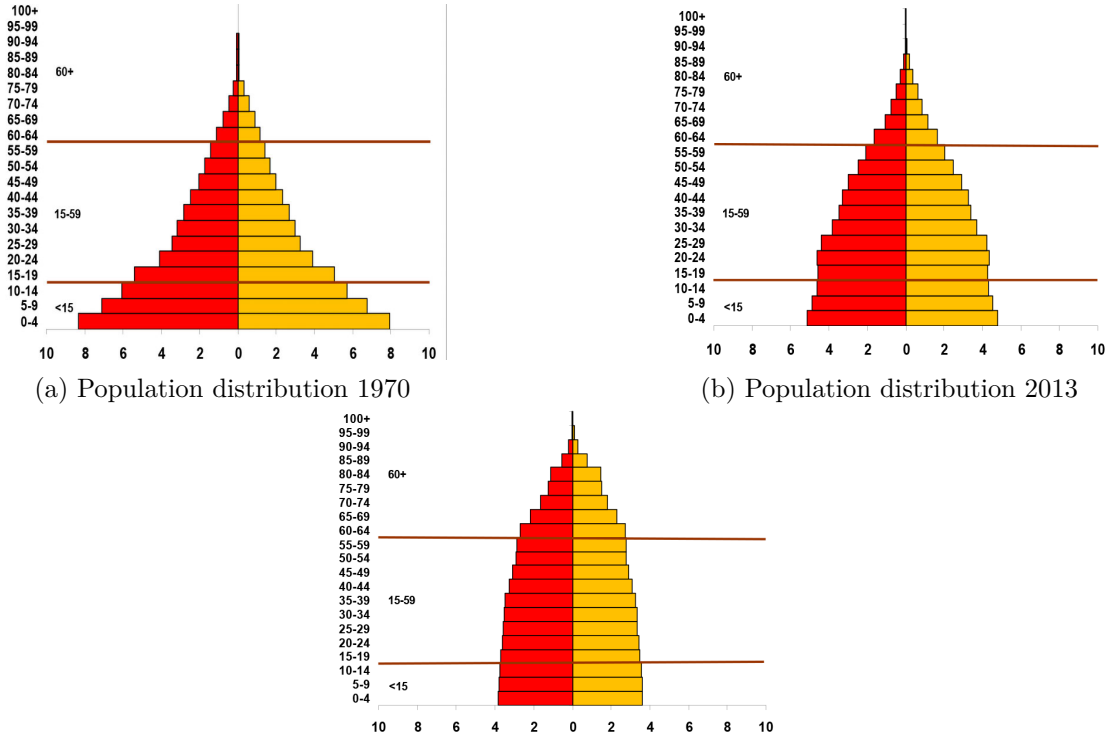



(b) Population distribution 2013




(c) Population distribution 2050

Figure 2.2: Aging pyramids of more developed countries (males shown in red, females shown in yellow) [United Nations, Department of Economic and Social Affairs, 2013]




(a) Population distribution 1970

(b) Population distribution 2013

(c) Population distribution 2050

Figure 2.3: Aging pyramids of less developed countries (males shown in red, females shown in yellow) [United Nations, Department of Economic and Social Affairs, 2013]

number of younger people is constantly decreasing, whereas the number of older adults is increasing, raising challenges for the society.

The main motivation for using technology in order to overcome the challenges posed by the demographic change can be summarized as increasing costs of the health care system, the caregiver burden and the importance of elderly people living independently [Rashidi and Mihailidis, 2013]. However, since the ratio of younger people in comparison to older people changes, current health care and caregiver systems will need technology in order to provide efficient support. This motivates the introduction of AAL, focusing on the support of older adults by providing Information and Communication Technology (ICT). The main goal is the development of technology in order to assist older adults during their ADL and thus enable them to stay independent as long as possible. The main objectives of AAL can be summarized according to the AAL Joint Programme initiative [AAL Joint Programme, 2012]:

- enable people to stay in their own homes as long as possible,
- increase their autonomy, self-confidence and mobility,
- promote a healthier lifestyle and support to maintain their health status,

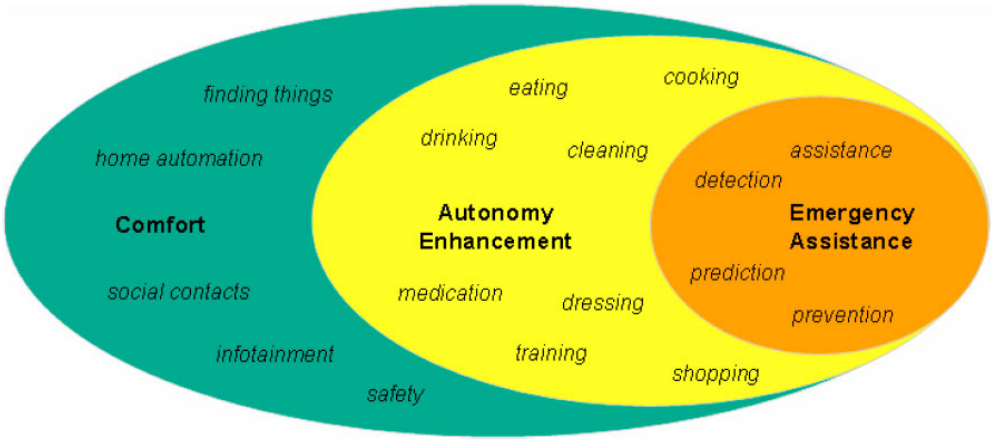



Figure 2.4: Areas within AAL [Kleinberger et al., 2007]

- prevent social isolation and enhance the security,
- also support secondary end-user (i.e. carer, family, care organizations) and
- increase efficiency and productivity.

However, the goal is not to reduce caretaker stuff, but to support them in a way that they can spend more valuable time with older adults. AAL does not only foster the development of new assistive technology for the elderly, but also focus on social inclusion [Sun et al., 2009]. Moreover, developed solutions should be unobtrusive in order to not interfere with the ADL, as long as the system is not needed. According to Kleinberger et al. [Kleinberger et al., 2007], AAL systems need to enhance the quality of life by being unobtrusive, being able to adapt to personal changes over time and provide high usability in order to achieve a high end user acceptance.

In order to assist older adult living on their own, the areas depicted in Figure 2.4 need to be addressed [Kleinberger et al., 2007]: *emergency assistance* addresses the prevention and detection of emergency situations (e.g. falls). If an emergency situation could not have been prevented, it needs to be detected and immediate assistance need to be provided. *Autonomy enhancement* addresses less time critical aspects of an elderly's life. However, in order to enhance the autonomy of older adults, assistance needs to be provided in an unobtrusive way, only when it is needed. A perfect medication reminder for example should only remind to take the medication, if the person forgot to take the medicine. If the elderly took the medicine on their own, no reminder is needed. Solutions in this area are broad, starting from automatic pill dispensers reminding to take the medication to providing assistance for elderly people suffering from dementia during the hand-washing process [Hoey et al., 2010]. The last area within AAL tackles *comfort* aspects in order to ensure the highest comfort for elderly people. This does not only include smart home technology regarding house automation [Scanaill et al., 2006],

but also systems and services to prevent social isolation [Sun et al., 2009]. Tools and technologies within the context of AAL can be categorized into [Rashidi and Mihailidis, 2013]:

- *smart homes,*
- *mobile and wearable sensors* and
- *robotics*

Typical *smart home technologies* consist of infrared motion sensors, Radio-Frequency Identification (RFID), pressure sensors, cameras and microphones. These technologies are also called ambient or ubiquitous sensors, since they are integrated in the flat. *Mobile and wearable sensors* are typically accelerometers or gyroscopes, but also blood pressure sensors as well as glucometers are used. Due to advances in the field of optical sensing, blood glucose, blood pressure and cardiac activity can be measured optically, thus providing a non-invasive measurement. *Robots* for assisting elderly people focus on supporting the elderly while performing their ADL [Hans et al., 2002]. Moreover, robots support older adults by fetching objects, feeding or dressing. But also during meal preparation, shopping and housekeeping, robots can assist elderly people in order to enable them to stay independent.

Rashidi and Mihailidis [Rashidi and Mihailidis, 2013] do not only categorize AAL tools, but also provide a taxonomy of algorithms, depicted in Figure 2.5. According to this taxonomy, research focus on the following areas:

- *activity recognition*,
- *context modeling*,
- *anomaly detection*,
- *location and identity identification* and
- *planning*.

*Activity recognition* is divided into mobile activity recognition, ambient activity recognition and vision-based activity recognition. The first uses accelerometer and gyroscope data, which are obtained from smartphones since accelerometer and gyroscopes are integrated in smartphones and thus provide huge amounts of data. If a network of ambient sensors is available, ambient activity recognition can be performed. However, most of the systems require labeled training data in order to accurately classify different activities. Vision-based activity recognition need to overcome a huge variation within indoor scenes. Since cameras are used, not only the variety of indoor scenes but also privacy concerns are the main challenges of these approaches. *Context modeling* retrieves temporal and spatial information from the surrounding, e.g. the spatial layout of a scene, in combination with ubiquitous ontologies. The area of *anomaly detection*

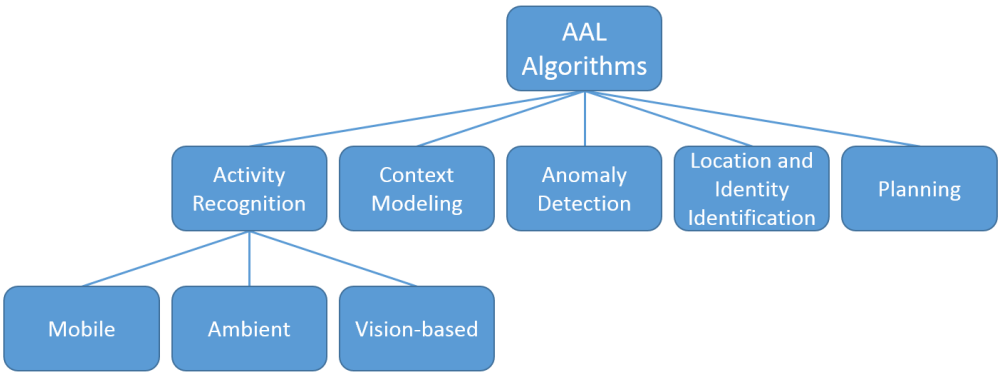



Figure 2.5: Taxonomy of AAL algorithms based on the work of Rashidi and Mihailidis [Rashidi and Mihailidis, 2013]

detects abnormal behavior within the ADL, since deviations from standard behavior indicate health-related problems. The use of Global Positioning System (GPS) in outdoor environments to locate elderly people is feasible, but does not work robustly within indoor environments. Hence, *location and identity identification* algorithms are developed, introducing various indoor positioning systems based on different technologies: smart (pressure sensitive) floors, motion sensors, RFID or ultrasound based tracking systems are just a few of them. Moreover, unobtrusively identifying people is also of interest, in order to personalize the system according to the user's needs. Finally, schedule *planning* is important to allow flexible reminders, support the elderly according to their needs and help them to perform their ADL.

This overview shows that the motivation of this thesis within the context of AAL is feasible, since technologies for elderly people are needed in order to support them properly. Moreover, challenges arise due to the use of technology within private homes. From a technical point of view, the system need to be cheap, easy to install, self-calibrating and adaptable to changing conditions within the scene. Although the use of 3D sensors does not need any sensor to be worn by the elderly, the limited range and occlusions restrict the application of these sensors. However, also societal challenges based on the acceptance of the users and privacy issues need to be considered.

## 2.2 Vision-Based Systems

Within the context of AAL, different sensors types (e.g. accelerometer, motion sensors, ultrasound tracking systems, cameras) are used [Rashidi and Mihailidis, 2013]. However, vision-based systems can be installed with minimal effort and offer the advantage that no sensor need to be worn. Moreover, research show that vision-based systems are flexible and different applications are feasible [Cardinaux et al., 2011]. Vision-based systems are either 2D or 3D systems, where approaches using a single camera are 2D

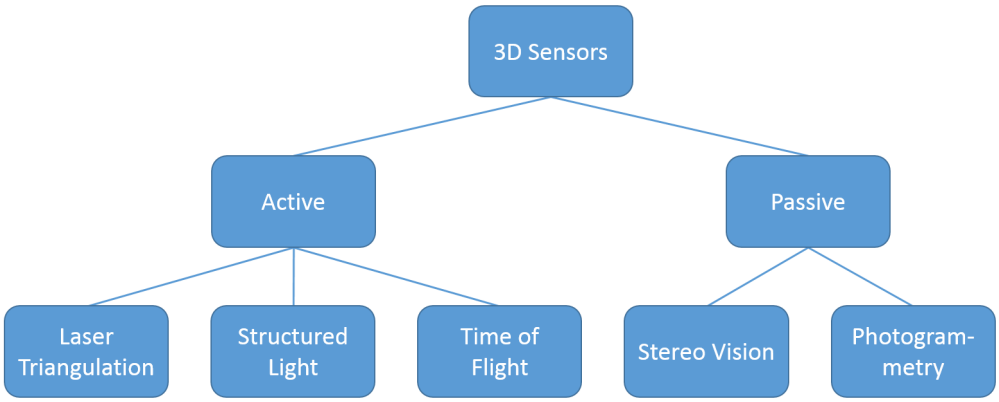



Figure 2.6: Classification of 3D sensors based on the work of Sansoni et al. [Sansoni et al., 2009]

approaches, but when using multiple cameras 3D structures within a calibrated setup are reconstructed [Zambanini et al., 2010].

An overview of selected 3D sensors is shown in Figure 2.6: sensors can be classified into *active* and *passive* sensors, according to the underlying technology [Sansoni et al., 2009]. *Active* sensors project light into the scene and measure the reflection, whereas *passive* sensors acquire information of the scene without the use of additional light sources. In the following, the most prominent and appropriate sensors for the application in the context of AAL are described and their strengths and weaknesses are discussed based on the work of Sansoni et al. [Sansoni et al., 2009].

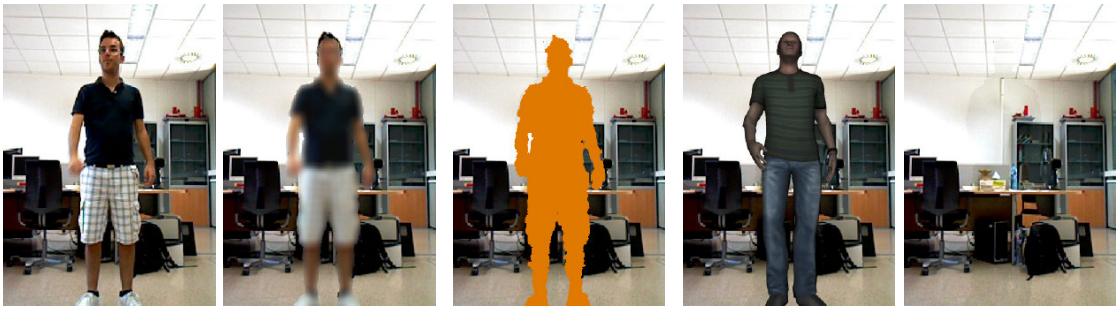
Sensors based on *laser triangulation* project a plane of light using laser stripes into the scene and calculate the distance of all points according to triangulation. The 3D information of the scene is obtained by a variation of the laser stripe within the scene, i.e. the scene is scanned successively and thus requires static scenes during the measurement. In contrast, sensors based on the *structured light* approach obtain the 3D information of a scene by projecting a reference pattern onto the scene and calculating the distance using triangulation. This has the advantage that the 3D information is calculated simultaneously, without the need for scanning the scene successively and thus can be applied to dynamic scenes. *Time-of-Flight (ToF) sensors* send laser pulses into the scene and reconstruct the 3D information of the scene based on time and intensity differences of the reflected pulse. *Stereo vision* records the scene with a calibrated set of two cameras and reconstructs the depth information based on the matching of common points within the recorded images, modeling the functionality of the human eyes. The advantage of this approach is its simplicity and no need for additional (synchronized) illumination. *Photogrammetry* exploits redundancy from different views by identifying and matching common points, recorded in different views. By matching the identified common points, a 3D reconstruction is obtained.

In conclusion, passive systems obtain the scene information based on RGB cameras whereas active systems provide additional lighting in order to measure the depth of the scene. Since especially in the context of AAL privacy need to be considered, active sensor systems are more suitable for the chosen context. Since laser triangulation requires a static scene until the scan is finished, either structured light or ToF sensors are feasible for the use within the context of AAL. Moreover, when applying technology in practice, costs are an important factor. Since Microsoft introduced a low-cost 3D sensor based on the structured light approach in 2010 [Microsoft, 2010], the Microsoft Kinect respectively its reproduction, the Asus Xtion pro [Asus, 2011] is used within this work.

### 2.2.1 Privacy Aspects

Vision-based approaches provide advantages of easy installation and flexibility without wearing a sensor in the field of AAL, but also privacy issues need to be considered carefully. Since the use of cameras in surveillance systems in public places is widely discussed in our society, the use of vision-based systems within private homes of elderly people need to be addressed and privacy as well as dignity of persons using this system need to be ensured. Chaaraoui et al. [Chaaraoui et al., 2012] discuss privacy issues of camera based systems within private homes. The authors [Chaaraoui et al., 2012] propose to ensure privacy by modifying or removing different areas from an image in order to protect the privacy (i.e. anonymization of the subject). This can either be performed by applying image filters (blur, pixelating), modification of the face, or replacement of subjects by abstract visualizations. Moreover, Chaaraoui et al. [Chaaraoui et al., 2012] note that a trade-off between privacy and usefulness need to be obtained. Within the context of AAL, especially the identity, appearance, location and activity of the subject need to be protected. However, only identity and appearance is taken into consideration since the authors [Chaaraoui et al., 2012] state that information about the location and activity is needed in order to provide appropriate help.

Different levels of privacy protection are proposed, depending on the context of the information. Figure 2.7 depicts an overview of the privacy levels proposed by Chaaraoui et al. [Chaaraoui et al., 2012]: camera images do not protect any privacy, since persons can be identified and their appearance is fully shown (Figure 2.7a). Applying a blurring filter, depicted in Figure 2.7b, allows to anonymize the person, but still appearance information is available. Using this visualization does not allow to identify the person any longer, but reveal personal information about the person (e.g. whether the person is dressed). By providing silhouette images, not only features to identify the person but also their appearance are highly protected since only the contour of the person is preserved (Figure 2.7c). Further anonymization lead to the substitution of the person by a 3D avatar, shown in Figure 2.7d. This visualization allows to fully protect the person's identity and appearance, since only pose information is available. Finally, Figure 2.7e ensures the highest level of privacy by fully removing the person from its context. However, this has the drawback that no information about the person itself is available anymore and thus, no conclusions in case of emergency can be drawn.



(a) Camera image (b) Camera image with blurring filter (c) Silhouette of person (d) Avatar representing the person (e) Person removed

Figure 2.7: Visualizations depending on the level of privacy [Chaaraoui et al., 2012]

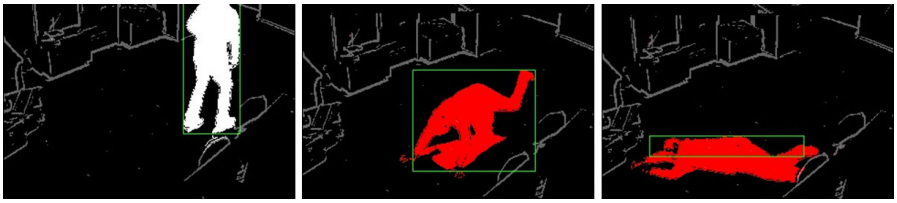


Figure 2.8: Anonymized snapshots using the system proposed by Zambanini et al. [Zambanini et al., 2010]

Although the privacy of the person itself is protected by these visualizations, it is still possible to easily identify the surrounding area of the person. Zambanini et al. [Zambanini et al., 2010] anonymize the camera pictures by applying edge detection algorithms to ensure the dignity of the elderly, depicted in Figure 2.8. Thus also ensures the privacy of the surrounding, while spatial relations are visible. Within this thesis, the use of 3D depth sensors is proposed, since depth information preserves privacy by its design automatically. When using depth information, it is not possible to identify the person, its appearance, or the surrounding of the person and thus, the flat of a person cannot be identified. On the other hand, while preserving privacy, information about the location and activity is available since it is possible to identify e.g. chairs and tables, but not their appearance.

Figure 2.9 shows a RGB camera image together with its corresponding depth image. It is possible to identify the person and its appearance using the information provided of the RGB image, whereas an identification of the person and its surrounding is not possible when using depth images, since only the silhouette of the person is available. Depth images (Figure 2.9b) do not visualize the scene with colors, but the gray level indicates the distance of the objects and surrounding to the sensor. The darker the color, the closer the object is to the sensor. On the other hand, brighter colors are used for objects at a higher distance. Black holes within the depth image indicate reflecting or absorbing areas where no valid depth measurement is available and are caused due to the

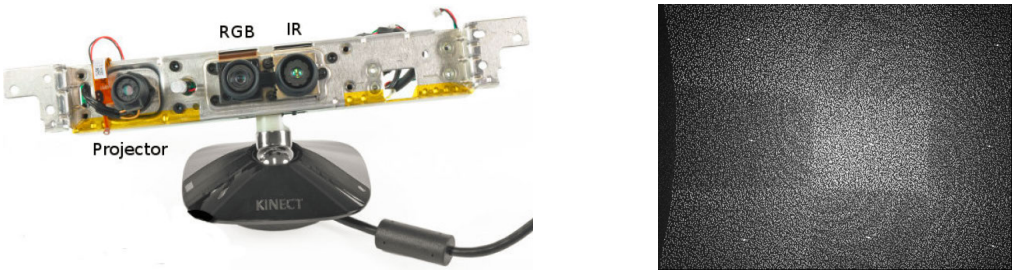
(a) Camera image (b) Depth image

Figure 2.9: RGB camera image and its corresponding depth image

functionality of the sensor. Although the privacy is preserved from a technical point of view, organizational aspects enable to identify the person even within depth images. If the location of the installed sensor is mapped together with the name of the person living in the flat, the elderly can be identified (assuming that only one person lives within the flat). However, although the person is identified, the appearance of the person is fully protected. Only the silhouette is detected and thus no conclusions whether the person is wearing clothes or the emotional state can be obtained since neither the clothes, nor the face is visible. Elderly people are skeptical using new technology and fear, that they are monitored 24/7. Although from a technical point of the privacy is preserved, risks cannot be eliminated but minimized. The use of technology in AAL always comprise risks, but on the other hand provides benefits for the elderly. Hence, elderly people need to balance reasons if the benefits or the risks for a specific system predominate, thus resulting in the decision whether or not to use a system.

### 2.2.2 Functionality of the Kinect

The Kinect sensor was introduced by Microsoft in 2010 [Microsoft, 2010] as an add-on for the Xbox console. It was the first low-cost 3D sensor and thus received a lot of attention from the research community [Smisek et al., 2011]. The Kinect consists of a RGB camera, an infrared projector as well as an infrared camera, depicted in Figure 2.10a. The functionality of the Kinect is based on structured light imaging, where the projector emits a pre-defined infrared light pattern to the scene [Fofi et al., 2004, Han et al., 2013], depicted in Figure 2.10b. Due to the spatial arrangement of the pattern and its varying sizes, as well as distortions depending on the distance to the camera, the depth camera captures the light pattern and an on-board chip calculates a depth map. The RGB camera of the Kinect has a resolution of 640x480 @30 frames per second (fps) or 1280x1024 @10 fps, whereas the depth system (infrared projector and infrared camera) has a resolution of 640x480 @30 fps, offering a practicable range of 0.8 - 3.5 m [Han et al., 2013].




(a) RGB camera, infrared projector and camera of the Kinect [Smisek et al., 2011]

(b) Infrared pattern emitted by the Kinect [McDonald and Levin, 2011]

Figure 2.10: Functionality of the Kinect

Moreover, human pose estimation is proposed by Shotton et al. [Shotton et al., 2011], allowing to extract body parts and skeleton joints based on depth images. Detected skeleton joints of two people provided by the Microsoft Software Development Kit (SDK) are depicted in Figure 2.11. According to Microsoft [Microsoft Developer Network, 2015], the skeleton tracking algorithm is optimized for people facing the sensor, either standing or sitting in front of the sensor. They further mention that the tracking algorithm can fail, if the person is not directly facing the sensor. The number and type of detected joints depends on the SDK to be used - two SDK are mainly used together with the Kinect [Han et al., 2013]: the official SDK provided by Microsoft [Microsoft, 2015], being able to track 20 skeleton joints and the OpenNI [OpenNI, 2011] provided by PrimeSense, allowing to track 15 different skeleton joints [Han et al., 2013]. PrimeSense developed the sensor of the Kinect for Microsoft and introduced, in cooperation with Asus, their own 3D sensor [Asus, 2011]: Asus Xtion pro, offering almost the same hardware specification as the Microsoft Kinect, but built in a smaller case. In contrast to the Microsoft Kinect, the Asus Xtion pro does not contain a RGB camera, but only a depth sensor and thus allows to fully respect the privacy within the context of AAL, since it is technically not possible to obtain a RGB image from this sensor. Hence, for the rest of this work, the Asus Xtion pro is used as 3D sensor in combination with OpenNI [OpenNI, 2011] in order to obtain pose and tracking information.

Although the use of the Kinect has advantages, there are also drawbacks that need to be discussed. The drawbacks of the Kinect are the high system requirements and due to the use of an USB interface, the limited range in comparison to network cameras, since a PC need to be placed close to the Kinect. Moreover, since Microsoft is the only manufacturer producing the Kinect, the use of the Kinect results in a dependency to Microsoft since no similar alternatives are available at the market. However, the market of 3D sensors is evolving and since this hardware is seen as a future trend, new developments will introduce 3D sensors similar to the Kinect.

Figure 2.12 depicts an overview of different versions of the Kinect sensor: Figure 2.12a depicts the Kinect introduced by Microsoft, whereas the Asus Xtion pro is shown in Figure 2.12b. The Kinect for windows v2 is the successor of the Kinect, depicted in

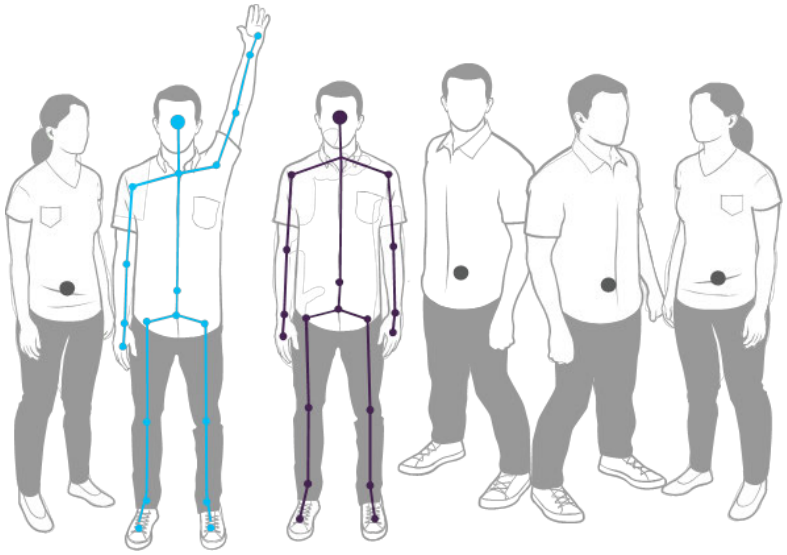


Figure 2.11: Detected skeleton joints [Microsoft Developer Network, 2015]


(a) Microsoft Kinect [Microsoft, 2010]

(b) Asus Xtion pro [Asus, 2011]

(c) Kinect for windows v2 [Microsoft, 2015]

Figure 2.12: Versions of the Kinect sensor

Figure 2.12c. In contrast to the previous versions of the Kinect, the functionality of the Kinect for windows v2 is not based on structured light, but on ToF in order to achieve more accurate results [Microsoft, 2013].

The main advantages of the Asus Xtion pro, especially within the context of AAL, can be summarized as follows:

- *No additional light source needed:* due to the use of infrared light, the Asus Xtion pro also works during the night, when falls of elderly people occur (e.g. when going to the bathroom in the dark).
- *Sensor is robust to changing lighting conditions:* switching the lights on and off does not affect the results of the Asus Xtion pro. However, direct sunlight interferes with the projected infrared pattern and thus, no depth value can be calculated. This restricts the use of the Asus Xtion pro sensor to indoor environments only.
- *No calibrated camera setup is needed:* in contrast to the use of a calibrated multiple camera setup in order to calculate a 3D reconstruction, no calibration is needed.

- *Standard algorithms can be applied to depth information:* standard algorithms from computer vision (e.g. foreground/background segmentation, tracking) can be applied to depth data directly.

- *Protection of privacy:* in contrast to the Microsoft Kinect (version 1 and 2), the Asus Xtion pro only obtains depth data and does not provide a camera. Thus allows to ensure privacy by design, since it is technically not possible to obtain a camera image from the Asus sensor.

### 2.2.3 Accuracy of the Kinect

Smisek et al. [Smisek et al., 2011] evaluated the accuracy of the Kinect sensor, by analyzing its depth resolution. The depth resolution of the Kinect is within the range between 0.65 mm at a distance of 0.5 meters and up to 685 mm at a distance of 15.7 m. These results indicate that the use of the Kinect for indoor environments is feasible, although the accuracy decreases with higher distance. Figure 2.13 depicts a function of the depth resolution of the Kinect, depending on the distance of the sensors and shows that the depth resolution within a range up to ten meters is below 300 mm. Moreover, the performance of the Kinect is compared to the performance of a stereo-camera (two Nikon D60 SLR) as well as a ToF (SwissRanger SR-4000) system. Smisek et al. [Smisek et al., 2011] showed that the Kinect performs similar to a stereo system with medium resolution and outperforms the ToF system in terms of accuracy and costs. Stoyanov et al. [Stoyanov et al., 2011] evaluated the Kinect and two ToF sensors based on ground truth data obtained by a laser sensor. For short distances, the Kinect slightly outperformed both ToF sensors and performed similarly to the accuracy of the laser sensor. However, no sensor achieved the accuracy of the laser sensor on longer distances.

Ditta [Dutta, 2012] compared the skeleton tracking of the Kinect with a marker based system from Vicon and showed, that the errors of the Kinect are approximately 5 mm within a range of 1-3 meters in comparison to the Vicon system. Galna et al. [Galna et al., 2014] evaluated the use of the Kinect for the detection of movement symptoms for people with Parkinson's disease and compared their results with a Vicon system. Normal actions (standing, walking and reaching) are combined with actions from the Unified Parkinson's disease Scale and include, amongst others, hand clasping and finger tapping. The timing of movement is measured accurately as well as extensive movements are detected accurately. Only when monitoring fine movement, the Kinect is not able to obtain accurate results and thus is outperformed by the Vicon system. Overall, Galna et al. [Galna et al., 2014] conclude that the Kinect can accurately detect most movements related to Parkinson's Disease. The accuracy of the Kinect within exergames is evaluated and compared to a Vicon system by van Diest et al. [van Diest et al., 2014]. The outcome of their evaluation shows that the Kinect accurately detects movements of the trunk, but does not detect the movement of hands and feet accurately, resulting in a difference of up to 30% in comparison to the Vicon system. The reasons for the lower accuracy of the Kinect is the reduced resolution (640x480) in comparison to the Vicon system (4704x3456) and the low and irregular sampling frequency [van Diest et al., 2014].

Plantard et al. [Plantard et al., 2015] propose a framework to simulate 500 000 poses at a workplace in combination with different positions of the Kinect, in order to perform an automatic large scale evaluation of the accuracy of the Kinect. The results show that the accuracy depends on the specific pose as well as the position of the Kinect. Although most results are accurate (e.g. error of the shoulder position is 2.5 cm), positions with partial occlusions results in the failure of the skeleton tracking algorithm.

Moreover, the performance of the skeleton tracking system during six different exercises is evaluated by Obdrzalek et al. [Obdrzalek et al., 2012]. Again, a marker based tracking system provides ground truth data and it is shown that the Kinect has a great potential. However, since exercises are performed either sitting or while touching a chair, the skeleton tracking algorithm fails when body parts are occluded or a chair is presented. Although problems with the skeleton tracker are reported, Obdrzalek et al. [Obdrzalek et al., 2012] state that within a more controlled environment, tracking results are better. They conclude that during general postures a variability of about 10 cm can be observed in comparison to the marker based tracking system. Nevertheless, these results show that the use of a Kinect is feasible, but depends on the application and context. Due to its low costs in comparison to other 3D sensors and its accuracy, the Kinect is used in computer vision for achieving different and diverse tasks: approaches using the Kinect for object tracking and recognition, human activity analysis, hand gesture analysis and 3D mapping of indoor environments are just few of them [Han et al., 2013].

## 2.3 Detection of Critical Events

Critical events of getting bad diseases, criminal violence or financial crisis are concerns for the elderly, resulting in fears [Deshpande et al., 2008]. The most common fears in elderly people arise from themselves and their home environment - about 50 percent suffer from their fear of falling [Deshpande et al., 2009, Howland et al., 1998]. In combination with a high mortality rate of fallers [Wild et al., 1981], falls are a critical event to be detected automatically. Moreover, if the elderly are not able to get up on their own again, they may lie on the floor for hours, until help is provided [Wild et al., 1981]. Noury et al. [Noury et al., 2008] have shown that getting help quickly after a fall reduces the risk of death by over 80% and the risk of hospitalization by 26%. Furthermore, elderly people suffering from dementia are not able to react to emergency situations properly [Leikas et al., 1998]. Hence, the aim of assistive systems is not only to assist, but also to reduce the cognitive load on the user [Lubinski, 1991]. This motivates the introduction of a fall detection system, which is able to detect falls and raise alarms automatically. Moreover, these systems boost the confidence of the elderly in living independently [Xinguo, 2008].

Fall detection systems can be divided into three major approaches [Xinguo, 2008]: *wearable devices*, *ambient devices* and *camera-based* (or vision-based) approaches. Figure 2.14 shows an overview of the three major approaches including divisions for each of these approaches into smaller and thus more specific approaches.

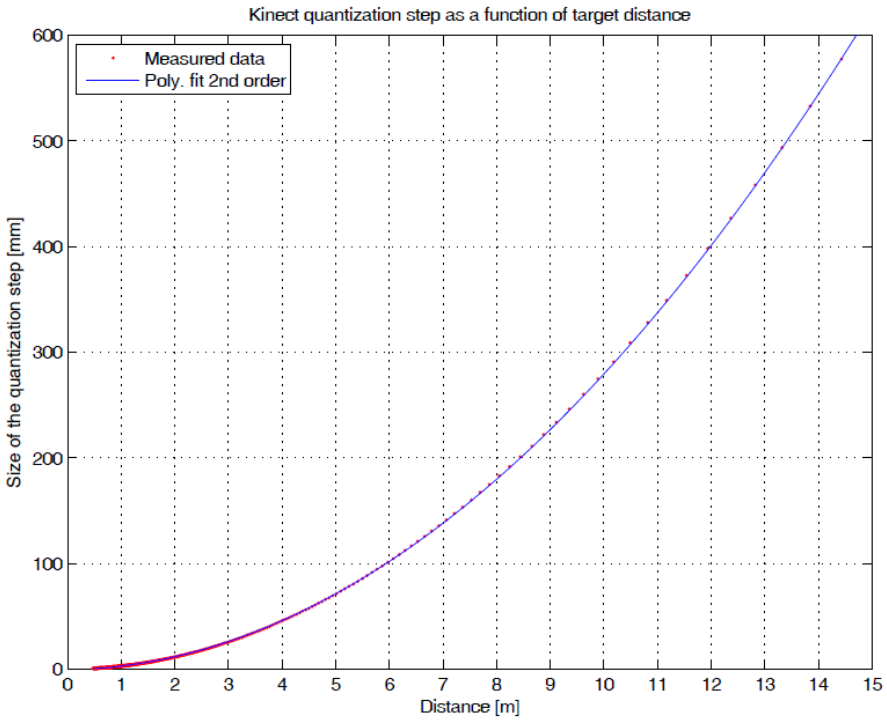



Figure 2.13: Depth resolution of the Kinect [Smisek et al., 2011]

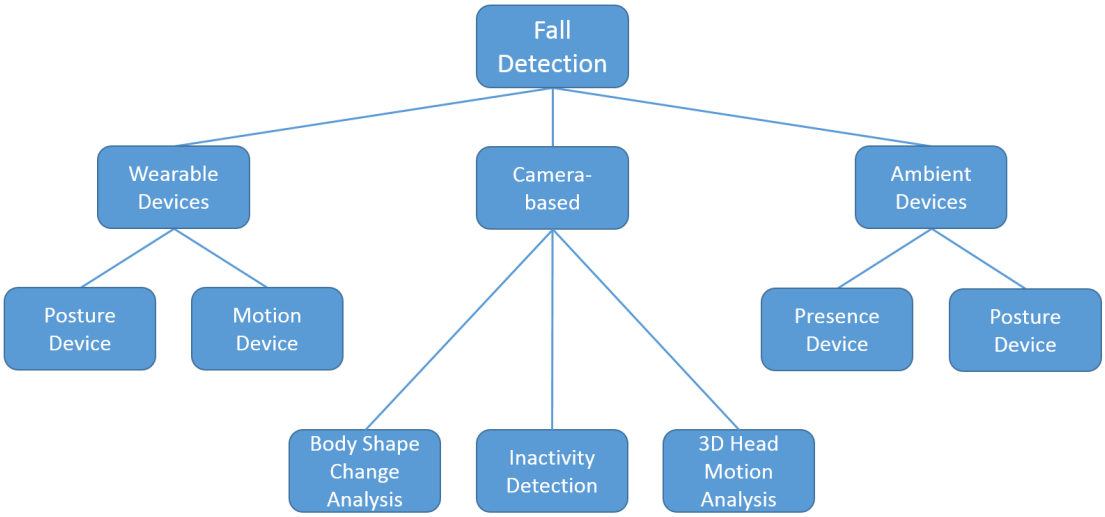



Figure 2.14: Classification of fall detection approaches based on the work of Xinguo [Xinguo, 2008]

### 2.3.1 Wearable Devices

Wearable devices broadly used to assist elderly people are panic buttons, which need to be worn (e.g. on the wrist) by the elderly and pressed if an emergency situation occurs and help is needed [Miskelly, 2001]. These devices have the main drawback that the elderly need to actively push the button - if they are not able to push the button (e.g. due to the loss of consciousness), help cannot be provided. Hence, wearable sensors detecting falls automatically have been developed [Doukas et al., 2007, Bagalà et al., 2012, Fontecha et al., 2013]. These wearable sensors detect the body orientation, the impact of falling (using accelerometers) or the amount of activity/movement. Figure 2.15 illustrates the change of acceleration during a fall: the impact caused by the fall can be recognized by a reduced acceleration followed by a peak in the magnitude before the acceleration returns to a normal level. Särelä et al. [Särelä et al., 2003] combine a panic button (i.e. button on the wrist) together with a movement sensor to detect emergency situations automatically if the user is not able to push the button anymore. Noury et al. [Noury et al., 2003] combine the measurement of the impact together with the measurement of the body orientation and the vibrations on the body surface to build a fall detection device which they called "actimeter". The main advantage of wearable devices are costs, as such systems are cheap - the main disadvantage is that sensors need to be worn, which is very intrusive [Xinguo, 2008]. An evaluation of fall detection systems based on accelerometer data is presented by Bagala et al. [Bagalà et al., 2012]. In contrast to most work being evaluated on simulated falls, evaluation is performed on a dataset of 29 real falls. Results of this evaluation indicate that especially the sensitivity is much lower when testing on real data in comparison to simulated data. Moreover, also the number of false alarms within a 24 hour period using different approaches is analyzed by Bagala et al. [Bagalà et al., 2012]. Figure 2.16 depict the number of false alarms within a time frame of 24 hours, analyzing data from 3 different person. The number of false alarms varies from 3 to 85, depending on the method. It can be noticed that most approaches produce approximately 5-10 false alarms within 24 hours. Please note that furthermore the corresponding sensitivity need to be taken into account, since the sensitivity describes the amount of correctly detected falls. This needs to be considered since an optimal system should have a high sensitivity rate, while not producing false alarms.

More recent research focuses on the detection of falls by not using additional devices, but smartphones since they are equipped with accelerometers and thus can be used to detect falls [He et al., 2012, Tacconi et al., 2011, Abbate et al., 2012]. Abbate et al. [Abbate et al., 2012] propose to use a smartphone for fall detection in combination with a classification engine in order to differentiate between a fall and ADL. The authors state that especially the ADL sitting/lying, jumping/running/walking and hitting the sensor may cause false alarms and are thus analyzed in detail. A neural network is used to classify eight features into one of the ADL classes or the fall class. However, the smartphone need to be worn in a fixed position in order to obtain reliable results. On the other hand, smartphones are not stigmatizing and thus can be used to detect falls without anybody taking notice of this system. Moreover, a combination with ambient or vision-based systems is possible: wearable sensors can be used to detect falls outdoors,

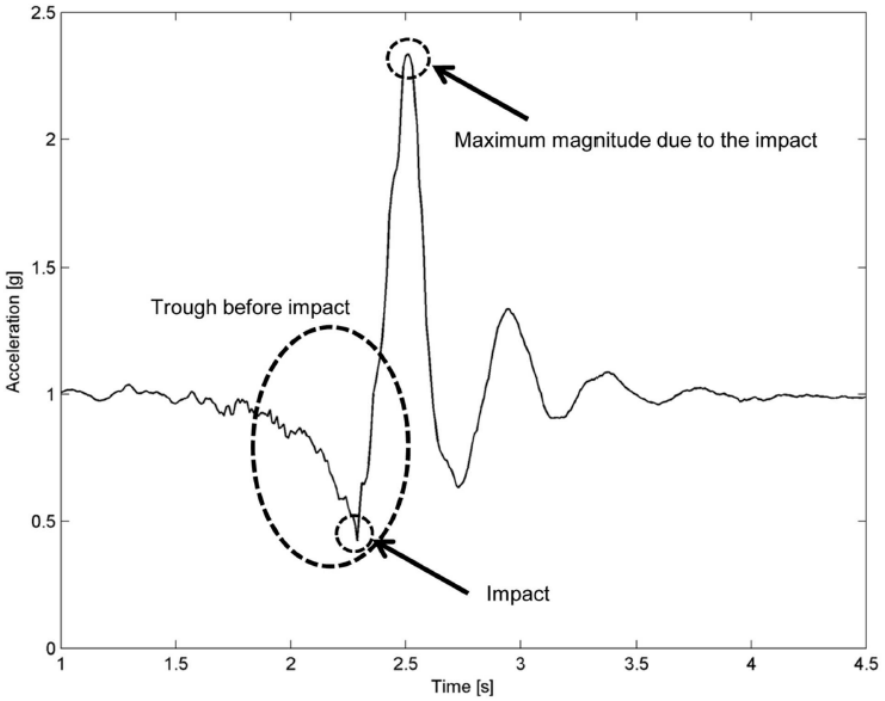



Figure 2.15: Impact of a fall on an accelerometer [Bagalà et al., 2012]

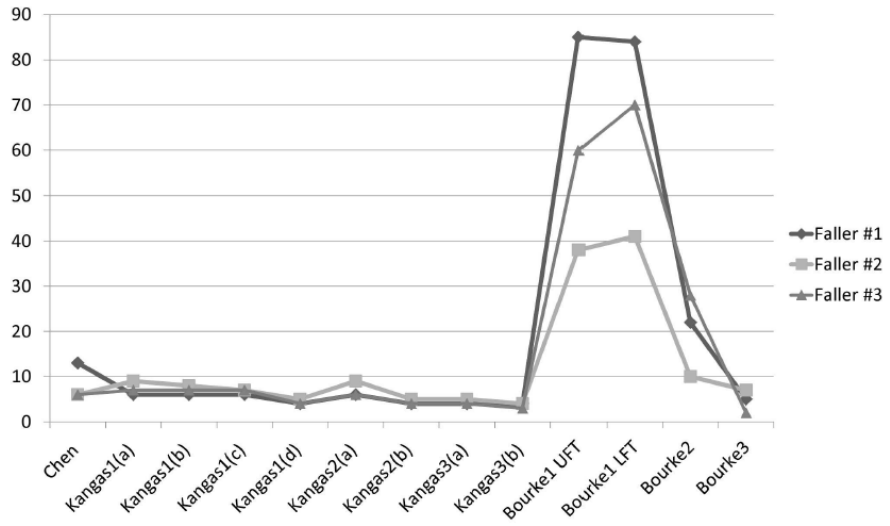



Figure 2.16: False alarms caused within 24 hours (3 user) [Bagalà et al., 2012]

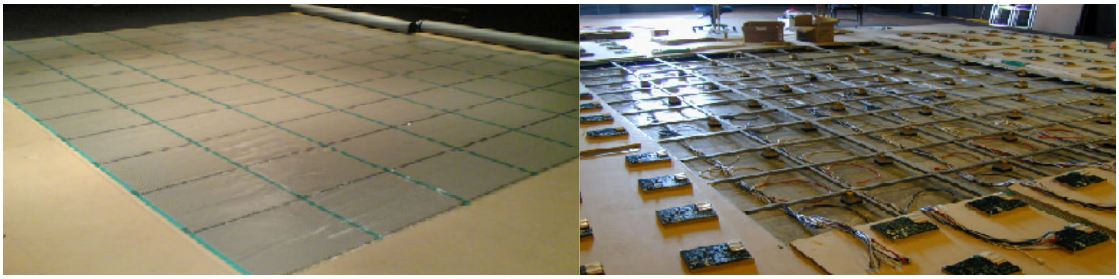


Figure 2.17: Pressure sensitive floor [Rangarajan et al., 2007]

whereas ambient or vision-based sensors detect falls indoors.

Smartphones cannot only be used to detect falls, but also to assess the frailty of a person [Fontecha et al., 2013]. The aim of Fontecha et al. [Fontecha et al., 2013] is the combination of accelerometer data from the smartphone together with clinical indicators (functional, nutritional, cognitive, geriatric and independence in ADL) in order to provide a more objective frailty assessment. During the performance of tests (e.g. get-up and go test), additional data is recorded with the smartphone. Seven different features (dispersion measures) are calculated based on the acceleration data from each test and combined with individual clinical indicators in order to provide a frailty assessment.

### 2.3.2 Ambient Devices

Ambient devices are multiple sensors which are installed within the flat [Xinguo, 2008], turning the flat into a smart home [Scanaill et al., 2006] being able to support elderly people living alone at home [Chan et al., 2009]. Approaches and sensors used in this field are very broad, including measuring the vibration of the floor to detect falls [Alwan et al., 2006, Litvak et al., 2008], detecting falls by using pressure mats [Miskelly, 2001, Zhang et al., 2011], pressure sensitive floors [Rangarajan et al., 2007] (Figure 2.17) or motion sensors [Zhang et al., 2011]. Ambient sensors are not intrusive, as they can be hidden within a smart home, but are difficult to install and have the drawback of a high false alarm rate [Xinguo, 2008].

### 2.3.3 Vision-Based Systems

Vision-based systems are able to overcome limitations of other sensor types [Mihailidis et al., 2002], but raise privacy issues. Hence, in contrast to Xinguo [Xinguo, 2008], within this thesis no video or depth data is recorded in order to respect privacy concerns. Vision-based systems can be distinguished between systems using 2D images and systems using 3D data (e.g. obtained by multiple cameras [Zambanini et al., 2010] or 3D sensors [Jansen et al., 2007]). To overcome limitations of multiple cameras (e.g. calibration is needed) and traditional 3D sensors (e.g. availability and costs) the use of the Asus Xtion pro as a vision-based 3D sensor for fall detection is proposed within this thesis. In the following sections different types of vision-based systems are summarized.

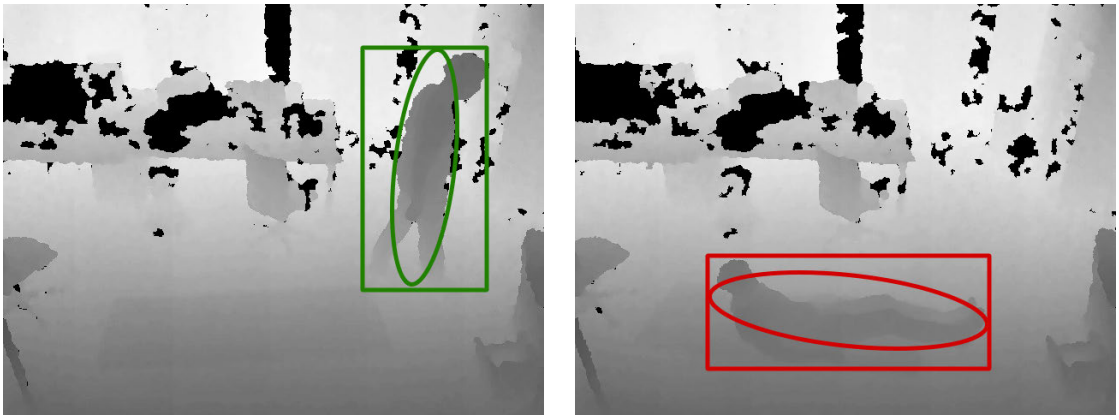


Figure 2.18: Analysis of the bounding box aspect ratio and the orientation of the ellipse to detect falls

### Body Shape Change Analysis

The shape of a person implies the orientation and thus is used to distinguish whether a person is in an upright position or not. The use of the bounding box aspect ratio (width to height ratio) to detect falls is proposed by Anderson et al. [Anderson et al., 2006]. If people are in an upright position, the bounding box aspect ratio is bigger than one (i.e. $height > width$). In case of a fall, the ratio changes to a value smaller than one (i.e. $height < width$). Another approach presented by Rougier et al. [Rougier et al., 2007] uses information of an approximated ellipse instead of a bounding box. Falls are detected by analyzing the orientation of the ellipse as well as the ratio of the major axis of the ellipse. Figure 2.18 illustrates these two approaches and depicts the shape of a person during a normal activity and during a fall. Furthermore, the corresponding bounding boxes and ellipses to analyze the bounding box aspect ratio and the orientation of the ellipse are illustrated. The use of a bounding box and an approximate ellipse for fall detection is feasible, but depends on the quality of the background segmentation. Assuming that the background segmentation yields in robust results, the fall detection also yields in robust results. A fall into the direction of the camera only using 2D images cannot be recognized by both approaches, as the change of orientation of the person cannot be detected.

Approaches not using 3D sensors reconstruct 3D information for humans from silhouettes gained by different camera views [Anderson et al., 2009]. The human is represented by the use of voxels allowing to identify different states (upright, on–the–ground and in–between), depending on the shape of the person. The quality of this approach also depends on the quality of background segmentation, but it has the main drawback of needing a calibrated camera setup.

Zambanini et al. [Zambanini et al., 2010] propose a method to detect falls by using multiple cameras, and they distinguish between an uncalibrated camera-setup and a calibrated camera-setup. When using an uncalibrated camera-setup, scene analysis is

performed on each camera individually. Afterwards, the individual results are combined to get an overall decision. In contrast, if information from multiple cameras using a calibrated camera-setup is combined to reconstruct the person in 3D space, the combination takes place at an early stage. Feature extraction is performed on the 3D reconstruction of the person and a decision whether a fall occurred or not is made afterwards. Compared to other works [Aghajan et al., 2008], their system is not vulnerable to low-quality images (e.g. high noise and low resolution) as only basic information (i.e. silhouettes) are extracted from the image. Using a calibrated camera-setup results in a higher accuracy than using an uncalibrated camera-setup, but it is practically not possible to calibrate the cameras if they are installed in an elderly person's flat or house.

ToF cameras [Oggier et al., 2004] are generating depth maps and can be used for fall detection [Diraco et al., 2010]. Jansen et al. [Jansen et al., 2007] propose a system for pose recognition discriminating the poses standing, sitting or lying by applying a threshold to the height of the centroid. They state that their approach works in nursing homes reliably, but not in real homes due to false alarms.

#### Inactivity Detection

Abnormal inactivity can be determined by tracking people from an overhead position [McKenna and Charif, 2005, Nait-Charif and McKenna, 2004]. Therefore, zones with low activity (and little motion) are identified automatically and marked as inactivity zones (e.g. sofa). If the amount of motion is below a threshold and occurred outside of the learned inactivity zones, this event is defined to be an abnormal inactivity (e.g. person is lying on the floor). Inactivity detection is only able to detect falls indirectly by the lack of motion. Therefore it is important to ensure that the system is able to handle new situations (e.g. a chair is moved to a new position, thus moving the inactivity zone) properly.

A combination of applying a statistical model of inactivity zones and shape-based fall detection is introduced by Zweng et al. [Zweng et al., 2010]. A so called accumulated hitmap models areas with low and high activities. In combination with their shape-based fall detection, the robustness of their approach is enhanced.

#### 3D Motion Analysis

3D head motion analysis by using stereo vision sensors to detect falls is used by Belbachir et al. [Belbachir et al., 2010]. These biologically-inspired sensors feature a massively parallel pre-processing and reduce the amount of data in comparison to stereo vision cameras as they are not frame-based, but event-based. Hence, the motion of people can be determined and the position of the person can be extracted. A fall is detected by tracking the position and velocity of the head, as they assume the position of the head changes rapidly during a fall. Another approach by Rougier et al. [Rougier et al., 2006] uses 3D information obtained by one single camera to track the head of the person and to obtain its trajectory. Not only the head position but also the motion speed is taken

as an indicator for falls as the motion speed is assumed to be higher during a fall than during ADL.

The approaches of Zambanini et al. [Zambanini et al., 2010], Belbachir et al. [Belbachir et al., 2010] and Rougier et al. [Rougier et al., 2006] consider motion speed to detect falls, as they assume that the velocity is higher during a fall than during ADL. However, this assumption should not be made, as falls can also occur slowly and thus are not detected using these approaches.

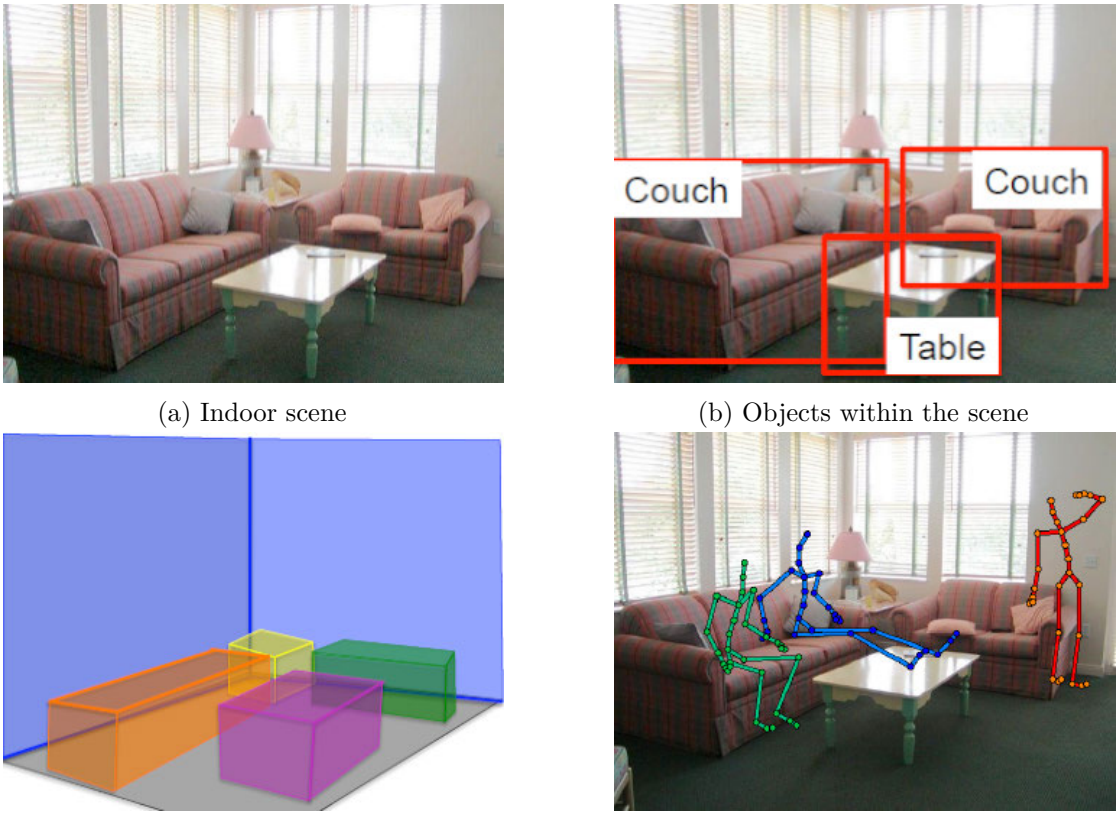
In contrast to the definition introduced by Xinguo [Xinguo, 2008], no restriction to the 3D motion analysis of the head of a person is made, as other body parts (e.g. centroid) are analyzed as well. An approach using ToF cameras detects moving regions within the 3D point cloud in a first step [Diraco et al., 2010]. The person (foreground) is segmented from the background and - in contrast to other works analyzing the head position - the distance of the person's centroid to the ground floor is analyzed. This results in an efficiency of 80% and a reliability of 97.3% when using a centroid-ground floor distance of 0.4 meters as threshold [Diraco et al., 2010].

## 2.4 Scene Understanding

Scene understanding includes the detection and classification of the objects within the scene [Felzenszwalb et al., 2010, Mutch and Lowe, 2006], modeling geometric aspects of the scene (e.g. generating room hypothesis [Hedau et al., 2009, Tsai and Kuipers, 2011], modeling objects within the scene [Hedau et al., 2012]) or understanding the scene from a human-centered perspective, describing how it can be used by humans, focusing on its affordances [Gupta et al., 2011]. The definition of scene understanding can be interpreted in different ways, depending on the specific problem or question. Due to advances in the technology of 3D sensors, indoor scenes are of major interest since 3D sensors yields in more robust results within indoor scenes. Figure 2.19a depicts an indoor scene, where detected objects (couch and table) within this scene are marked with bounding boxes in Figure 2.19b. A more abstract scene representation is based on geometric models, reflecting the scene layout in Figure 2.19c. However, both representations are rather object-centered than human-centered since these representations only represents objects, but do not take the interaction of humans into account. The representation depicted in Figure 2.19d does not depict the objects, but the interaction with humans and thus the functionalities of the objects are described (e.g. it is possible to sit on this object). Within this thesis a novel user-centered scene understanding approach is introduced, thus scene understanding approaches using object classification and geometric approaches are consolidated together and are compared to human-centered approaches.

### 2.4.1 Object-Centered Scene Understanding

Objects within scenes are of high interest since objects provide a meaning to a scene and thus are important for scene understanding. Hence, object detectors and descriptors play an important role in order to recognize pre-trained objects within a scene [Lowe,




(a) Indoor scene

(b) Objects within the scene

(c) Scene geometry

(d) Human-centered scene understanding

Figure 2.19: Interpretations of scene understanding [Gupta et al., 2011]

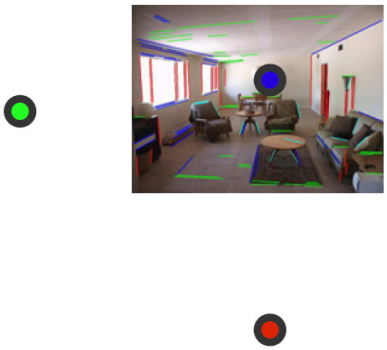
2004]. Object recognition is a challenging task due to high intra-class variability and low inter-class variability. Hence, objects within the same class can have a total different appearance, whereas objects from different classes might look similar. Moreover, changes of the viewpoint and the illumination raise additional challenges [Felzenszwalb et al., 2010].

Since this area is a large field of research, a variety of edge, corner and blob detectors are used for object recognition. Moreover, feature descriptors and machine learning are used to recognize objects, e.g. Scale-invariant feature transform (SIFT) [Lowe, 2004], Speeded-up robust features (SURF) [Herbert Bay et al., 2006] and deep learning [Deng and Yu, 2014]. However, object detection and recognition is far beyond the scope of this work, hence in the following geometric-centered approaches in order to estimate a room layout are discussed. In contrast to object-centered approaches, the goal of geometric-centered approaches is not to detect individual objects (e.g. table), but to model the layout of the scene. However, geometric-centered approaches can be combined together with object-centered approaches in order to detect both, the layout of the scene and the objects within the scene [Hoiem et al., 2008, Bao et al., 2011].

Indoor scenes are highly cluttered and contain objects, making the detection of the 3D room layout difficult, e.g. if wall-floor boundaries are not visible due to occlusions by objects [Hedau et al., 2009]. Hedau et al. [Hedau et al., 2009] propose an approach to detect the vanishing points within a scene, allowing to reconstruct the box orientation in combination with the surface labels in order to model the 3D scene layout. Long line segments within the scene are detected, depicted in Figure 2.20a. Assuming a box layout of the scene and the fact that parallel lines intersect the image plane in vanishing points (under perspective projection), these lines are grouped in order to end in one of three vanishing points - illustrated in Figure 2.20b. Based on these vanishing points, candidate room hypotheses by casting rays from the estimated vanishing points are generated (Figure 2.20c). Candidate room hypotheses model the walls, the ceiling as well as the floor in a box layout. These box layouts are ranked according to a pre-trained model and a score for each candidate hypothesis is calculated. In order to estimate whether the line corresponds to a wall or an object, confidence maps for surfaces are calculated, shown in Figure 2.20d. The top row in Figure 2.20d depicts the confidence maps for the left and the right wall, whereas the bottom row represents the confidence maps for the floor and objects. These confidence maps are generated by using a segmentation algorithm, segmenting the picture into homogenous segments (i.e. floor, walls, ceiling and objects). The segmentation is based on color, texture and edge information in combination with the information of the already obtained vanishing points. In Figure 2.20e the weighted room hypotheses are combined with the confidence maps in order to choose the best ranked room layout. Using this approach, room layouts can be reconstructed although the images are highly cluttered and objects occlude the boundaries of surfaces (e.g. boundary between wall and floor).

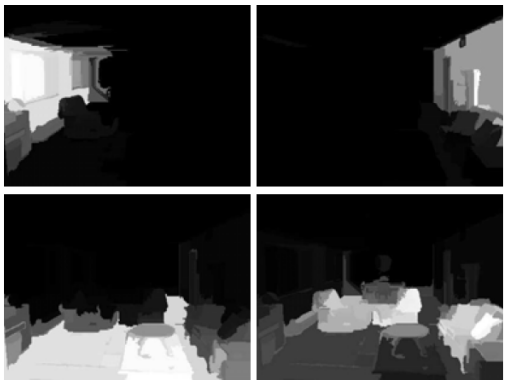
Hedau et al. [Hedau et al., 2012] presents an extension of their approach [Hedau et al., 2009] in order to recover free space in a scene, based on single images. Objects (i.e. furniture) are detected by exploiting the fact that big furniture can be assumed to have box like geometric structures. Starting from a basic room layout detected using the method of Hedau et al. [Hedau et al., 2009] (Figure 2.21a), box like objects are detected within the labeled surface regions - depicted in Figure 2.21b. Assuming that most furniture objects are box like (i.e. bed, sofa, table, chair), the fact that they are aligned parallel to the orientations of the walls and floor is exploited. Figure 2.21c shows detected objects by sliding a cuboid in 3D through the scene [Hedau et al., 2010]. The orientation of the objects is defined by the detected vanishing points and different sizes of cuboids are used in order to generate different object hypotheses. The object hypotheses are scored using an adapted Histogram of oriented gradients (HOG) [Dalal and Triggs, 2005] approach, where the gradients are computed along the directions of the vanishing points. In an additional refinement step, the placement of the cuboid is optimized based on corners and edges within the image (Figure 2.21d). Finally, Figure 2.21e depicts a floor occupancy map, where the height of the cuboids (objects) as well as the confidence is modeled.

(a) Detected line segments within scene



(b) Estimated vanishing points

(c) Different room hypotheses

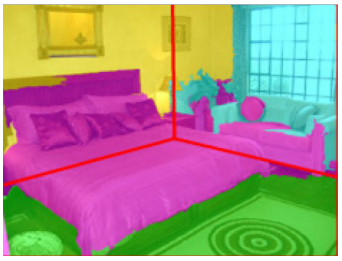


(d) Confidence map for surfaces

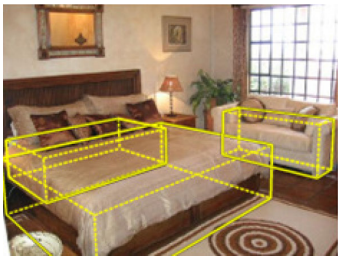
(e) Best ranked room layout

Figure 2.20: Room layout estimation [Hedau et al., 2009]

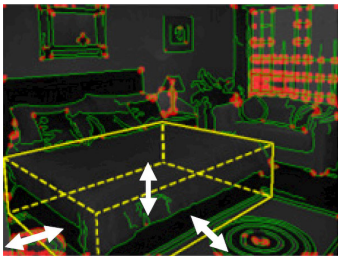
(a) Indoor scene



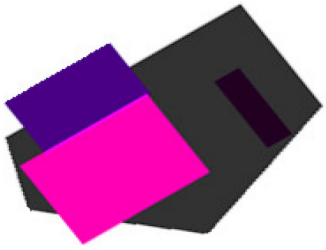
(b) Estimated room layout and surfaces [Hedau et al., 2009]



(c) Estimated cuboids [Hedau et al., 2010]



(d) Refinement of cuboid detection



(e) Floor occupancy map

Figure 2.21: Estimation of free space in indoor rooms [Hedau et al., 2012]

The basic geometry of an indoor scene is modeled by Tsai et al. [Tsai and Kuipers, 2011]. In contrast to previous approaches, a calibrated non-static camera is used and motion information is integrated in order to refine the model. In contrast to Structure-from-Motion based approaches [Dellaert et al., 2000,Hartley and Zisserman, 2004,Pollefeys et al., 2007], no exact 3D point cloud is calculated, but only the geometry is modeled. Hence, it is less computational expensive and thus can run in real-time, to be used on a mobile robot within the context of AAL. Different room hypotheses are generated and represented by the floor and wall surfaces. In contrast to previous work, Tsai et al. [Tsai and Kuipers, 2011] do not assume that the walls are perpendicular to each other, hence, arbitrary room layouts can be modeled - as long as the walls are perpendicular to the ground floor. The first frame of the video is analyzed and room hypotheses are generated using the approach by Delage et al. [Delage et al., 2006], obtaining a color based Bayesian network model by finding the floor-wall boundaries. Due to the analysis of feature points within consecutive video frames, the camera motion can be estimated using the Kanade-Lucas-Tomasi (KLT) tracker [Lucas and Kanade, 1981]. Due to the knowledge of the geometric scene model and the camera motion, the camera motion is predicted within different room layout candidates. The predicted motion is compared to the real motion and thus allows to choose the best fitting room hypotheses.

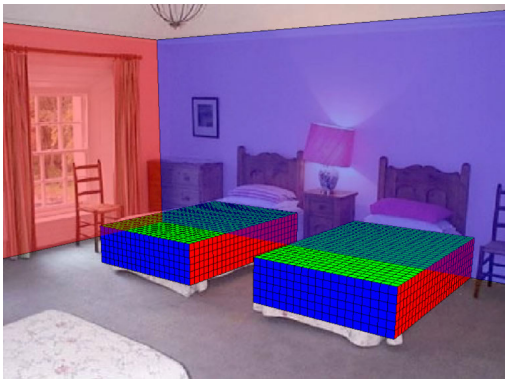
The work of Bao et al. [Bao et al., 2011] combines object detection and geometric layouts of the scene. In contrast to separately modeling object detection and geometric information [Hoiem et al., 2008], Bao et al. [Bao et al., 2011] maximizes the joint

probability for both, detected objects and supporting surfaces in the same step. Due to the idea of combining object detection and the underlying geometry in one single step, the estimation is more accurate than modeling both separately, since additional knowledge is considered. Candidate objects are detected by different object class detectors. Assuming that objects need to be supported by a surface (i.e. an object is always put on top of a table, ground floor, bed, etc.), possible surfaces are modeled. Candidate object detections as well as candidate surfaces are optimized, resulting in object detections, being supported with the highest likelihood by surfaces and the joint probability for both is maximized. Moreover, this approach cannot only be applied to indoor scenes, but to outdoor scenes as well.

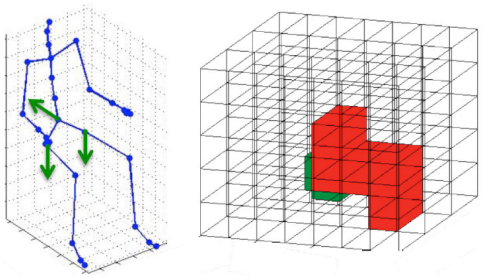
### 2.4.2 Human-Centered Scene Understanding

In contrast to traditional object-centered scene understanding, human-centered scene understanding focus on functional aspects based on information how the person interacts with the scene and which functionality the scene offers for a human [Gupta et al., 2011]. Hence, the interaction between the objects within a scene and persons is modeled in order to obtain its functionalities. Gupta et al. [Gupta et al., 2011] introduce a proof of concept system focusing on the human workspace rather than the object. By combining single-view indoor geometry estimation and human pose analysis of static 3D poses, a combination of 3D scene understanding together with human action modeling is achieved. The scene geometry is based on single images, where the layout of the room is modeled by the floor, walls and the ceiling. Moreover, the objects within the room are modeled using voxels, representing occupied areas of the room. Figure 2.22a illustrates the estimated room layout in combination with the modeling of objects, represented by voxels. Based on this room model, all possible interactions within the room are modeled by fitting different poses into the model, resulting in functional descriptions of the objects (e.g. sitable or touchable objects). Possible poses for a human (e.g. sitting, depicted in Figure 2.22b) are extracted from 3D motion capture data and the pose is again represented by voxels. For each pose not only the occupied voxels are analyzed, but also the required object surfaces needed to support this pose. In the last step, human scene interactions are modeled by analyzing the possible poses within the scene and thus recognizing free areas and surfaces supporting e.g. the sitting pose.

The use of long-term tracking of humans (over the duration from a few minutes to several hours) in order to describe object functions within a room is introduced by Delaitre et al. [Delaitre et al., 2012]. Due to the combination of pose analysis (standing, sitting and reaching) and object appearance, the interaction between human actions and objects are modeled. Poses are detected from time-lapse videos using the approach of Yang and Ramanan [Yang and Ramanan, 2011]. Object appearance is obtained by using dense SIFT, combined with the localization within the image. The combination of the vocabulary of poses together with the object features results in a model, describing the long-term interaction between humans and objects. However, the use of pose estimation does not work robustly when being applied in practice and introduces incorrect pose estimations [Delaitre et al., 2012]. This effect can be minimized by enhancing the amount



(a) Room model with detected walls and objects



(b) Example for the sitting pose, showing occupied voxels (red) and required object surfaces (green)

Figure 2.22: Room and pose modeling [Gupta et al., 2011]

of tracking data and thus enhancing the accuracy of the pose estimation at a specific location - hence, the use of long-term tracking is proposed. In order to obtain long-term tracking information, the authors use time-lapse videos to model different interactions with the same object, thus enhancing the accuracy. However, by using time-lapse videos, only discrete but not continuous information is used and only snapshots are analyzed. In contrast to Gupta et al. [Gupta et al., 2011], Delaitre et al. [Delaitre et al., 2012] does not analyze where people can sit, but where people do sit. Hence, this approach is less theoretical than the approach by Gupta et al. [Gupta et al., 2011], since data from real scenes is used.

While Delaitre et al. [Delaitre et al., 2012] recognize the objects, Fouhey et al. [Fouhey et al., 2012] extends their approach by not only recognizing objects, but modeling the scene in 3D, based on the object functionality. Again, human poses are estimated and human-object relationships are modeled. Due to the use of time-lapse videos, poses at the same object are aggregated over time in order to minimize pose estimation errors. Room hypotheses are generated from appearance information and the correct room hypothesis is chosen due to the information of human-object relationships. Similar to [Delaitre et al., 2012], the authors focus on the poses standing, sitting and reaching in order to classify surfaces into walkable, sitable and reachable surfaces. Based on the pose information, estimates of the functional surfaces are generated and combined with the geometric information obtained by the room hypothesis. Again, Fouhey et al. [Fouhey et al., 2012] state that pose estimation yields in noisy pose estimates and is an open research problem.

The information about moving targets in order to detect the scene geometry from a single static camera is proposed by Taylor and Mai [Taylor and Mai, 2013]. The basic scene geometry, i.e. occluding and occluded objects as well as supporting surfaces (floor) are estimated. Figure 2.23 depicts an example where a shelf is correctly detected and the floor is marked in red, whereas occluding areas are visualized in green and occluded areas are depicted in blue. In contrast to similar work, e.g. by Schodl and Essa [Schodl

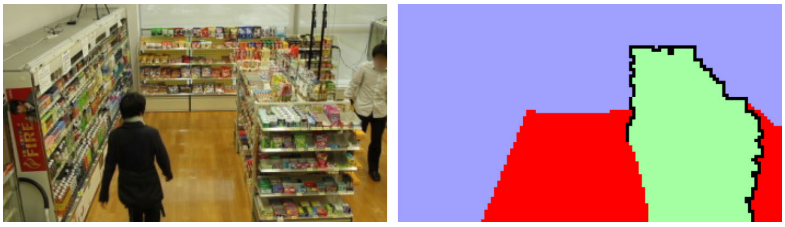


Figure 2.23: Supermarket scene with detected occluding (green) and occluded (blue) regions together with the detected ground floor (red) [Taylor and Mai, 2013]

and Essa, 2001], occlusion boundaries are modeled directly, not estimating depth layers. Taylor and Mai [Taylor and Mai, 2013] do not incorporate pose estimation since pose estimation is seen as open research problem. In order to detect free and occupied areas, moving foreground regions are detected and tracked. Occlusion boundaries are detected by applying the following three premises:

1. occlusion boundaries are always part of the silhouette boundaries,
2. occlusion boundaries cannot occur inside silhouette boundaries, and
3. occlusion boundaries have a persistent location and orientation over multiple detections.

With these premises, an occlusion boundary likelihood map is calculated. The floor is detected by combining footprint detection based on the tracked silhouette boundaries together with a color based appearance model in order to not assume a planar floor surface. The only assumption made by Taylor and Mai [Taylor and Mai, 2013] is that a sufficient number of moving objects is available in order to construct the scene model over time.

Chao et al. [Chao et al., 2013] use information about humans to improve the room layout in highly cluttered scenes, since vanishing points cannot be robustly detected within highly cluttered scenes. Additionally, due to severe occlusions, objects are not detected accurately and thus Chao et al. [Chao et al., 2013] assume that the detection of people is more robust than the detection of objects, resulting in a more robust room hypotheses. A first room hypotheses is generated by estimating the vanishing points of the scene by using the approach of Hedau et al. [Hedau et al., 2009], not only considering straight lines, but human pose information (i.e. sitting and standing) and the relationship of humans is incorporated to detect the vanishing points. This additional information allows to model the vanishing points more accurately and thus yields in improved room hypotheses in highly cluttered scenes.

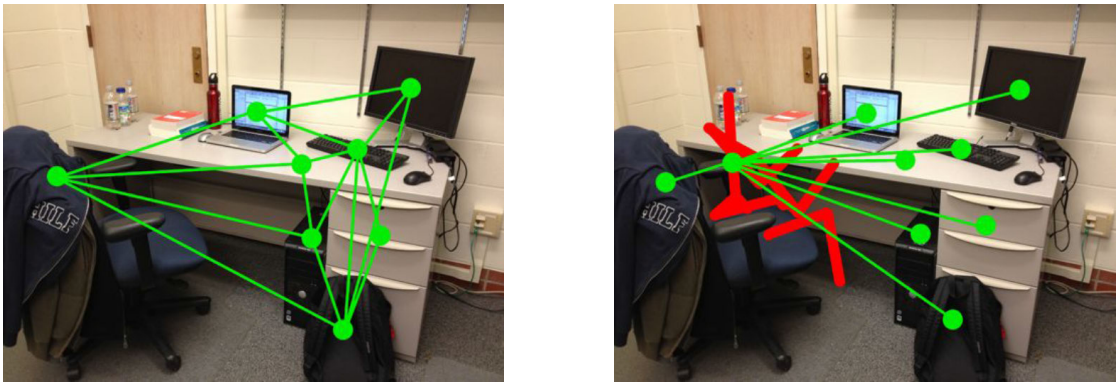
In contrast to all previous 2D camera based approaches, a human-centered scene modeling approach based on depth videos is introduced by Lu and Wang [Lu and Wang, 2012]. For this proof of concept, a background model is learned from the depth data and used to obtain the human silhouette. In combination with pose estimation performed

on the silhouette of the person, objects are modeled as 3D boxes within the scene. Together with geometric knowledge (estimation of vanishing points) of the scene, a room hypotheses including supporting surfaces for human actions is created and walkable areas are estimated. However, analysis and evaluation of the algorithm is performed on only several minutes of data and only deals with a few depth frames, but not on the long-term. Moreover, the authors [Lu and Wang, 2012] state that skeleton data could theoretically be used as well, but is not stable enough to obtain reasonable results, since skeleton data is noisy and defective.

In order to deal with noisy and defective skeleton data, Azimi [Azimi, 2012] summarizes approaches to smooth the skeleton tracking data by applying smoothing filters. Smoothing filters are applied to smooth the skeleton data over time, e.g. to remove jitter during tracking. In contrast to other scenarios (e.g. rehabilitation), a highly accurate position of the skeleton joints is not required for human-centered scene understanding, since small jitter of the joints does not influence the obtained scene functionalities. However, in the case of long-term tracking, tracking errors (i.e. an object is considered to be a person) are more important than jitter, but cannot be filtered by applying smoothing filters since the skeleton itself is correct (i.e. does not contain jitter), but it is fitted to an object instead of a person.

Instead of analyzing human motion and behavior to generate a scene model, Jiang et al. [Jiang et al., 2013] use RGB-D data and assume that objects within a scene are arranged by humans and thus have a meaningful object to object relationship. For example, the relationship between a computer monitor, a mouse and a keyboard is not arbitrarily chosen, but arranged in a way a human can use it. Hence, instead of only modeling the object-object relationships, hidden humans are incorporated in order to obtain a more meaningful model without detecting or tracking a person. Figure 2.24 depicts the object-object relationship (Figure 2.24a) and the proposed model, considering that all objects are meaningful arranged by a human (Figure 2.24b). Jiang et al. [Jiang et al., 2013] work with hidden (hallucinated) humans since they state that human-object interactions rarely occur in comparison to object-object relationships since more scenes contain objects than real human-object interactions. Hence, a human context can be considered even without observing a real person, but only by the arrangement of objects. Six human poses (three sitting and three standing poses) are obtained from the Cornell Activity Dataset [Sung et al., 2011], containing real human skeleton data. This information is used to model the hidden human in the proposed scene model. Objects are detected based on their appearance and in combination with object affordances a scene model considering the hidden human is obtained.

A spatio-temporal approach to model human activities and object affordances within RGB-D video is proposed by Koppula et al. [Koppula et al., 2013]. Activities are detected by exploiting RGB-D information together with object affordances, i.e. associated affordances [Gibson, 1979], in order to reliably detect activities (e.g. making tea) and sub-activities (e.g. pouring water). Activities and object affordances are modeled using a Markov Random Field (MRF). Activities are obtained from the skeleton tracker provided by OpenNI [OpenNI, 2011], although the tracking results are not accurate since the



(a) Object-object relationship model (b) Object-hidden human relationship model

Figure 2.24: Object relationship modeling [Jiang et al., 2013]

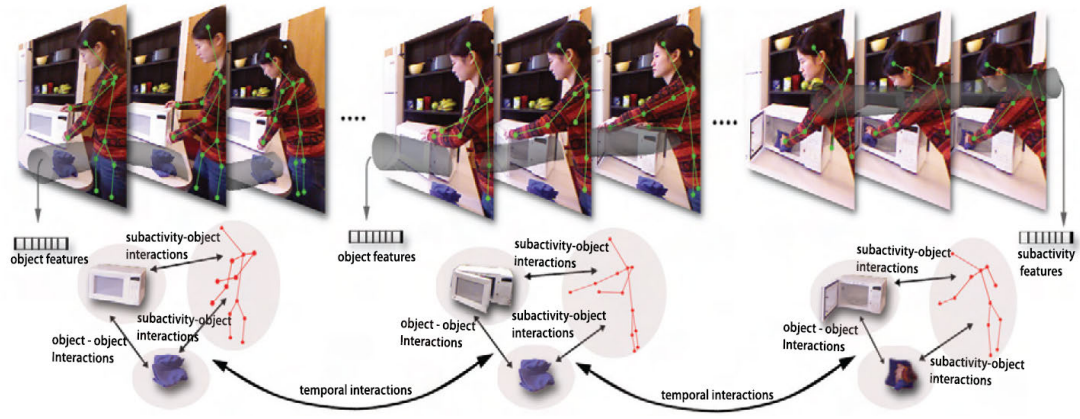



Figure 2.25: Spatio-temporal approach combining object affordances and activities [Koppula et al., 2013]

skeleton tracking is not optimized for occluded body parts [Koppula et al., 2013]. Objects are detected and tracked in close areas to the human by using a set of 2D object detectors in combination with particle filters. Figure 2.25 illustrates the approach of Koppula et al. [Koppula et al., 2013]: the activity is segmented into smaller sub-activities in short temporal chunks. For each sub-activity, the objects are detected using object features, whereas the person is represented by skeleton joints. The object-object and sub-activity-object interactions are modeled for each sub-activity individually. By adding temporal interactions between all sub-activities, the overall activity can be described.

## 2.5 Behavior Monitoring

Behavior monitoring is based on the establishment of behavior models, trying to optimize the prediction and explanation of human action [Aarts and Dijksterhuis, 2003]. Human behavior is guided by social norms, influenced by the behavior of other humans and is controlled by what we believe, other people expect from us [Cialdini and Trost, 1998]. Social norms within a given environment can be described by situational norms [Aarts and Dijksterhuis, 2003]. Situational norms are generally accepted beliefs and describe how to behave in a specific situation (e.g. silence in libraries or churches). Aarts and Dijksterhuis [Aarts and Dijksterhuis, 2003] describe two ways, how situational norms are learned: either people behave in a way that they think it is approved by other people, or, on the other hand, people imitate the behavior of other people in a specific situation in order to define social reality - in other words, they follow the majority of people. Heywood [Heywood, 2011] describes norms as a system of shared behavioral patterns that holds society together. Moreover, Heywood [Heywood, 2011] defines social norms as "Informal rules shared by groups or societies that guide behavior and have positive and/or negative consequences that help to make the behavior more or less self-correcting". In other words, norms are set of behavior rules the society agreed upon, defining how to behave in different situations. Depending on whether a person follow this accepted beliefs or not distinguishes between "normal" (expected and agreed behavior) and "abnormal" behavior (disapproval of behavior) in a specific situation.

In the field of computer vision, behavior is defined in multiple ways [Chaaraoui et al., 2012, Popoola, 2012]. Despite different definitions of the term behavior, the main focus of research is the detection of "abnormal" behavior, which is a deviation from "normal" behavior, i.e. the detection of not expected and approved behavior within a given context. The terms "normal" and "abnormal" highly depend on the context, hence the following describes definitions of "normal" and "abnormal" behavior within the context of video surveillance in general and, more specifically, in the field of AAL.

In video surveillance, one definition of abnormal behavior can be the behavior of moving objects, which draws the attention of a human observer due to unexpected behavior [Popoola, 2012]. In order to clarify the term moving objects, Candomo et al. [Candamo et al., 2010] defined four different types of interactions:

1. single person without interaction,
2. multiple person interactions,
3. person-vehicle interactions and
4. person-facility interactions.

Based on this definition of interactions, examples for abnormal behavior are loitering, abnormal crowd movement due to panic, breaking a car window or entering a facility in an incorrect way. In the field of AAL, behavior monitoring is defined in a different way, since behavior is monitored in a different context. Behavior analysis within AAL

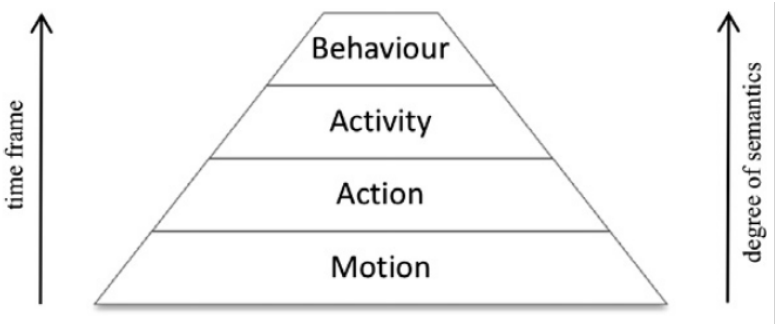



Figure 2.26: Taxonomy of Chaaraoui et al. [Chaaraoui et al., 2012]

focus on the analysis of a person's daily routine and their ADL [Chaaraoui et al., 2012]. Hence, abnormal behavior are deviations from the person's daily life, since deviations (e.g. changed behavior patterns) may be a result from a deterioration of the health condition (e.g. Alzheimer disease) [McKhann et al., 1984]. Although abnormal behavior seems to be well defined, the term behavior itself is not clearly defined, especially it is mixed up with similar terms like action, activity, etc. As a result, in the literature different taxonomies are used, highly depending on the context [Chaaraoui et al., 2012]. Nevertheless, this thesis follows the taxonomy proposed by Chaaraoui et al. [Chaaraoui et al., 2012], where behavior is defined as "the highest level of complexity with the longest time duration". This is motivated by the definition and distinction of motion, action, activity and behavior, depending on the duration and degree of semantics, illustrated in Figure 2.26. An overview is presented in Table 2.1: *motion* is defined to have a very short duration within frames or seconds and thus is used to detect movement and to segment the foreground. *Actions* are simple human primitives as sitting and walking within the duration of seconds and minutes, where motion is not only detected but it is also recognized what the person is doing. At the *activity* level, different ADL are recognized - this means that a sequence of actions in a particular order is detected within minutes or hours, depending on the performed ADL (e.g. cooking, housekeeping). *Behavior* is defined to offer high semantics and can be described as the way people are living and performing their ADL on an everyday basis, within a long time frame.

Behavior analysis approaches within the field of AAL either focus on specific ADL [Hoey et al., 2010, Chung and Liu, 2008, Duong et al., 2005], or they focus on a more abstract analysis of activity and behavior patterns [Monekosso and Remagnino, 2010, Virone and Sixsmith, 2008,Cuddihy et al., 2007]. Although this thesis focus on an abstract analysis without the restriction to pre-defined ADL, approaches analyzing specific ADL are described briefly. Moreover, abstract analysis of behavior patterns also contains the analysis of abnormal inactivity, since inactivity is also a measurement to detect behavior changes on an abstract and global level.

Table 2.1: Differentiation between motion, action, activity and behavior [Chaaraoui et al., 2012]

| Degree of semantics | Time lapse | Description |
|---|---|---|
| Motion | frames, seconds | Movement detection, background subtraction and segmentation; gaze and head-pose estimation |
| Action | seconds, minutes | Establish with which objects the person is interacting. Recognize simple human primitives (sitting, standing, walking, etc.) |
| Activity | minutes, hours | Tasks that consist of a sequence of actions in a particular order. ADL are recognized (e.g. cooking, taking a shower or making the bed) |
| Behavior | hours, days, ... | Highly-semantic comprehension comes into play (ways of living, personal habits, routines of ADL) |

### 2.5.1 Activity Level

Hoey et al. [Hoey et al., 2010] propose a system in order to assist people suffering from dementia during the process of hand-washing. They track both, the hands as well as the towel and observes the current state within the process of hand-washing. The process itself is modeled using a Partially Observable Markov Decision Process (POMDP) in order to determine which step within the hand-washing process is carried out. If a person is not able to carry out the current step of the hand-washing, help is provided in order to complete this step and to move to the next step. Although the health status can be detected by this system by analyzing the performance of the person, the main goal of this work is to provide assistance in order to wash the hands properly.

Activity detection based on tracking information is introduced by Nguyen et al. [Nguyen et al., 2005]. The environment is split into a grid of different areas and landmarks are identified. Activity detection is based on the visit of a person to specific landmarks and thus considers spatial information in order to detect the activity. As an example, the landmarks door, cupboard, fridge and dining chair are visited during the activity "short meal". A set of primitive activities (i.e. transitions between landmarks) are defined and are recognized using a hierarchical Hidden Markov Model (HMM). This approach considers spatial aspects, but temporal aspects (time of the day) are not considered. Hence only activities are detected, but abnormal behavior cannot be detected. In other words, the system is able to detect whether the pre-defined ADL is performed properly, but no conclusion about the overall health state can be obtained.

The activity of the elderly within a nursing center is modeled by Chung and Liu [Chung and Liu, 2008]. The foreground is extracted by using a background model, in order to

detect motion within the video. Based on the segmented foreground, the posture as well as the sequence of motions are extracted and used to recognize the activity. The authors consider spatial and temporal information as well as information about the activity the person is performing, in order to provide contextual information. Hence, location, time and the type of activity is modeled. However, the type of activity (e.g. "person walks to the toilet") need to be defined in advance. Moreover, also the locations need to be defined in advance, without learning the locations automatically. Due to the required pre-definition of all activities and locations in detail, this approach cannot be easily adapted to other scenarios.

Duong et al. [Duong et al., 2005] build a model of ADL using a modified version of a HMM. The HMM consists of two layers, where the bottom layer models actions and the top layer models activities. Please note that in this thesis the taxonomy of Chaaraoui et al. [Chaaraoui et al., 2012] is used, whereas Duong et al. [Duong et al., 2005] introduced their own taxonomy, where actions are named "atomic activities" and activities are named "high-level activities". Within their framework, actions are model using a spatio-temporal approach and thus considering the location and its duration. A typical sequence of actions during an activity is learned during the training phase and deviations from e.g. the duration or order while performing the ADL are detected. Although changes of behavior are detected on an action level, no overall behavior is modeled and thus do not allow to obtain an overview of the current health state of the person since the person might perform one ADL worse, but on the other hand another ADL might be performed better. Hence, only focusing on single ADL does not allow to obtain the health state.

### 2.5.2 Behavior Level

In order to obtain a more holistic view of the person's health state, the following work does not (only) focus on ADL, but also provide an aggregated holistic overview. Monekosso and Remagnino [Monekosso and Remagnino, 2010] equipped a home-lab with different sensors (e.g. motion detector, temperature sensor, lighting status) and developed a HMM for behavioral trends. After a training phase without any abnormal behavior, the model is tested in order to detect abnormal behavior. However, detected deviations during the evaluation not only result from the person, but also from the sensors itself, since noise disturbs the sensor data and influence the results. A similar approach is presented by Jain et al. [Jain et al., 2006], where sensor are installed within a home and lifestyle trends are detected. Moreover, not only sudden changes are detected, but also drifts (i.e. slow changes) can be detected. This is important since health changes do not necessarily happen abruptly, but can occur slowly.

Temporal aspects of activity patterns and trends are analyzed by Virone and Sixsmith [Virone and Sixsmith, 2008]. Again, the motivation is the detection of deviations during the performance of ADL, where the following ADL are defined: sleeping, dressing, eating, bathroom, meal preparation, hygiene, cleaning, phoning, washing, walking and sitting. Deviations are detected on each activity separately by using an unsupervised approach [Virone et al., 2008] in order to estimate the behavior of the person and detect

deviations from a normal behavior. Evaluation is based on motion sensor data combined with a stove-top temperature sensor and a bed-based vital sign monitor, gathered by 22 residents in an assisted living setting and software simulations. Spatial aspects are modeled by the placement of one motion sensor per room, hence only information about the occupancy in a room is retrieved, but not the exact location within the room. Although pre-defined ADL are analyzed, results are aggregated and allow to provide a holistic view of the health status, based on the Circadian Activity Rythmus (CAR).

Nait-Charif & McKenna [Nait-Charif and McKenna, 2004] use tracking information from an overhead camera to summarize activity in home environments. The movement of the person is tracked and the room is divided into entry/exit zones, inactivity zones and transition areas. A typical use of the room is modeled as follows: a person enters the room via an entry zone, moves to one or more inactivity zones and finally leaves the room via an exit zone. Transition areas are defined to be areas where the transition from an entry/exit zone to an inactivity zone or between inactivity zones take place. Inactivity zones are learned automatically using the approach introduced in [McKenna and Nait-Charif, 2004]. The person's speed is analyzed to define whether the person is active or inactive. Depending on the location of the person during the inactivity, the system detects whether the inactivity occurs in an already pre-defined inactivity area or outside such areas. This allows to detect abnormal inactivity which can be caused by a fall. Furthermore, activity patterns (i.e. sequence of visiting different zones) are analyzed and deviations of patterns are detected. Figure 2.27 depict trajectories from an overhead camera together with the learned entry and exit zones, as well as detected inactivity zones. The work of Nait-Charif & McKenna [Nait-Charif and McKenna, 2004] focus on spatial aspects of inactivity, but temporal aspects are not taken into consideration since only the sequence of visiting zones is analyzed but not associated with the time of the day (e.g. the sequence of visiting different zones may change depending on the time).

In contrast, Floeck & Litz [Floeck and Litz, 2008] and Cuddihy et al. [Cuddihy et al., 2007] focus on temporal aspects of inactivity. Activity data is collected using 30 sensors (i.e. motion detectors, door and window sensors) resulting in an activity profile [Floeck and Litz, 2008]. These profiles are introduced due to the sensors used, combining the data from different sensors to one profile. An inactivity profile is constructed by analyzing the duration of inactivity over time, defining inactivity as no activity from any sensor. As long as no activity is detected, the duration of inactivity raises continuously over time, shown in Figure 2.28. If any kind of activity is detected, the inactivity duration is set to zero (e.g. between 7 and 8 AM). Afterwards, the inactivity duration raises since no activity is detected between 8 and 9 AM. Due to the combination of motion and door sensors, the approach proposed in [Floeck and Litz, 2008] is able to differentiate between inactivity due to absence of the person (data obtained by door sensors) and inactivity when the person is present. Figure 2.29 depicts an inactivity diagram, distinguishing whether a person is present or absent when inactivity is detected.

In order to detect abnormal inactivity, the inactivity profile is compared to a pre-trained reference profile (e.g. average inactivity profile of one month). Therefore, the profiles are divided into $n$ different time slots. Floeck & Litz [Floeck and Litz, 2008]

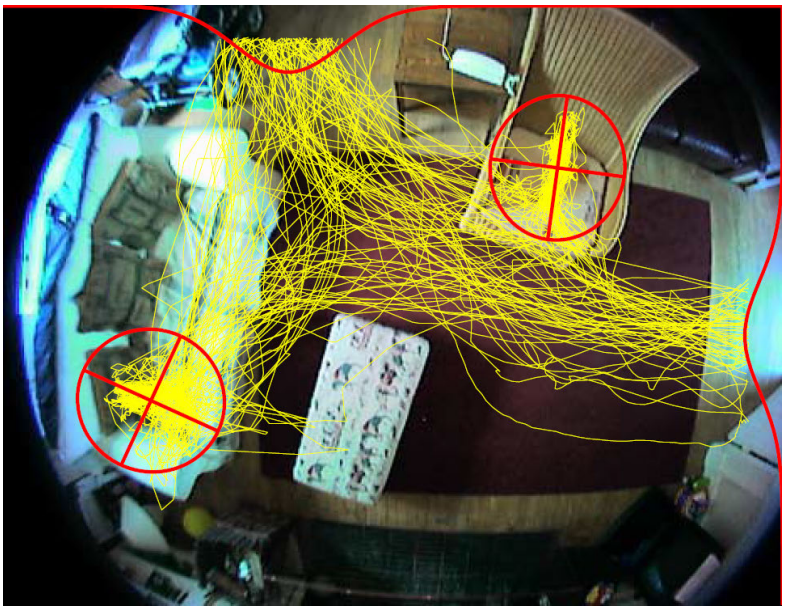


Figure 2.27: Inactivity and entry/exit zones together with trajectory data [Nait-Charif and McKenna, 2004]

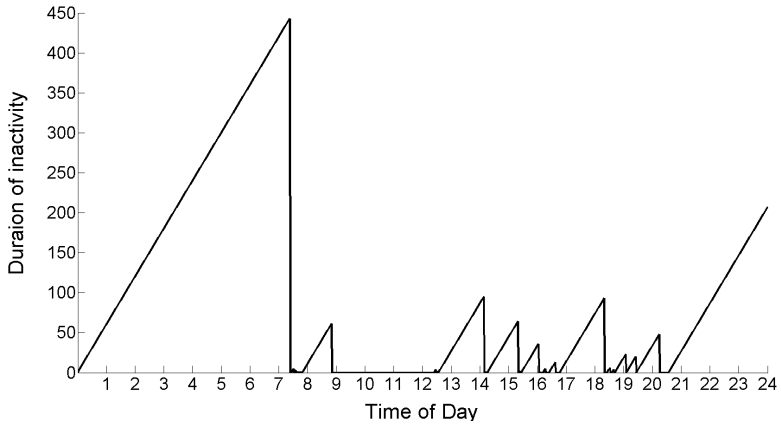



Figure 2.28: Inactivity profile

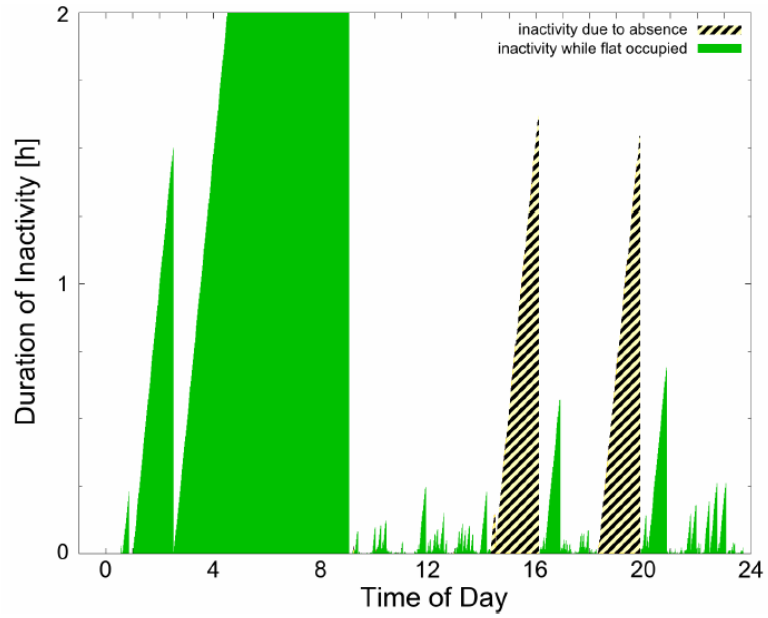



Figure 2.29: Inactivity profile considering data different sensor types [Floeck and Litz, 2008]

calculate the integral of inactivity of each time slot and combine all $n$ time slots to one feature vector per day, which is compared to the reference vector using the Dice coefficient [Dice, 1945]. By introducing a tolerance value and a convolution with a weighting vector, small temporal and numerical deviations are compensated (e.g. sleeping 5 minutes longer than normal). Since the inactivity profiles are compared on a one-day vector basis, deviations are detected at the end of the day. However, extensive evaluation of this approach is missing and thus no performance measures when being applied to real world scenarios can be obtained.

Cuddihy et al. [Cuddihy et al., 2007] use door sensors to detect if a person leaves the flat in order to minimize false alarms when no person is present. Similar to [Floeck and Litz, 2008], the authors use inactivity profiles and each day is divided into $n$ time slots. A reference alert line is learned over the duration of 45 days by analyzing the maximal inactivity duration at each time slot and adding buffers to allow small deviations. The uniform and variable buffer act as vertical tolerance and ensures, that the sensitivity of the algorithm is adapted according to the amount of inactivity (i.e. the algorithm is more sensitive during active times and less sensitive during inactive times). Furthermore, time shifts are compensated by applying a weighting function to the inactivity data and thus considering also adjacent intervals providing a temporal buffer. Each time interval is compared to the corresponding time interval of the alert line immediately, hence alarms are raised at the end of each time interval. The alert line is adapted based on a 45 days rolling window approach, hence the alert line is learned from the last 45 days and

adapts to behavioral changes automatically. An extension of this approach is proposed by Weisenberg et al. [Weisenberg et al., 2008], utilizing not only passive infrared motion and door sensors, but also wearable accelerometer and bed occupancy sensors. These additional sensors improved the performance of the system significantly.

Similar to Cuddihy et al. [Cuddihy et al., 2007], an approach utilizing data from smart home sensors in combination with an alert line is proposed by Moshtaghi and Zukerman [Moshtaghi and Zukerman, 2014]. In contrast to previous work, different regions are manually defined and the alert line is modeled individually for each region. However, the distribution of the sensor types within the homes is suggested to be performed together with the residents [Moshtaghi et al., 2013] in order to properly define interesting regions and to obtain suitable results. Hence, in order to define a new region, new sensors need to be installed accordingly. The distribution of the inactivity is modeled per region, allowing for different alert thresholds in different rooms and thus being more flexible than a global approach.

## 2.6 Summary

Within this thesis, a behavior model focusing on short-term (i.e. critical events within minutes), mid-term (changes within daily activities) and long-term (i.e. mobility changes over the duration of months) is proposed. In order to achieve this goals, advances in the field of critical event detection, scene understanding as well as behavior monitoring are introduced. The following summarizes the motivation for this work, based on the state-of-the-art.

**Critical Event Detection:** studies show that falls are a major risk for the elderly, especially when living in their own flat. In combination with the evolution of new 3D sensors, the introduction of an automatic fall detection system based on 3D information is motivated. Fall detection approaches can be classified into approaches using wearable sensors, vision-based approaches and approaches using ambient sensor information. Elderly people are reluctant to wear sensors and ambient devices require construction work in order to install them. In the field of vision-based fall detection, approaches focus on body shape change analysis, inactivity detection or 3D motion analysis. Since 3D sensors based on structured light or ToF ensures to protect privacy aspects while yielding in feasible results, the use of the Asus Xtion pro instead of single or multiple 2D cameras is proposed. Literature shows that 3D information yields in more robust results than 2D approaches, but requiring a calibrated camera setup - thus needing more effort when installation the system. Since related work is based on either 2D images or focus on low-level vision tasks when using Asus Xtion pro, a fall detection approach exploiting high-level information obtained by the Asus Xtion pro is proposed. In contrast to the focus on the fall event mentioned by Xinguo [Xinguo, 2008], the proposed work neither focus on the fall event nor is the fall constricted to time constraints (i.e. the fall process lasts from x to y seconds). Raising an alarm automatically if a person is detected to be on the ground and is not able to get up anymore is proposed, as this is the situation where help is needed - independently of the reason for being on the floor (falling or

lying down on purpose). Hence, if a person lies down on the floor on purpose and is not able to get up again, an alarm will be raised since help is needed anyway. Furthermore, the approach combines and benefits from all sub categories of vision-based approaches defined by Xinguo [Xinguo, 2008]: body shape change analysis is performed by analyzing the major orientation of the person, whereas a person lying on the ground is part of inactivity analysis, since the person is not moving or getting up. Moreover, 3D motion analysis is done by tracking the person's skeleton position in a 3D environment over time.

The proposed fall detection approach is based on 3D depth respectively skeleton information, providing a robust and self-calibrating fall detection system by estimating the pose of the person in combination with the ground floor. Moreover, due to the use of depth information instead of cameras, the privacy of elderly people is protected.

**Scene Understanding:** traditional scene understanding approaches are geometric and object-centered [Felzenszwalb et al., 2010, Gupta et al., 2013, Bao et al., 2011]. Since objects within a scene are related to humans, humans need to be considered during scene analyzes as they provide additional information about the scene and the structure of the scene. Hence, scene understanding, being part of behavior modeling, should be human-centered instead of object-centered. However, human-centered scene understanding approaches only use discrete information of 2D time-lapse videos [Gupta et al., 2011, Delaitre et al., 2012, Fouhey et al., 2012], discarding temporal and continuous aspects. In the field of human-centered scene understanding, different approaches of pose estimations are used - poses are either estimated from 2D images or from 3D depth data. However, all have in common that pose estimation is an open research problem, since the results are not robust and noisy [Lu and Wang, 2012, Taylor and Mai, 2013, Koppula et al., 2013]. Especially the results of the OpenNI skeleton tracker [OpenNI, 2011] are noisy and inaccurate, so that it cannot be used for pose estimation [Lu and Wang, 2012]. Recent approaches combine human tracking information together with geometric information or detection of objects in order to refine the scene model [Lu and Wang, 2012, Chao et al., 2013, Koppula et al., 2013].

The proposed human-centered scene understanding approach within this thesis is based on continuous and noisy long-term tracking data of humans. Tracking information is obtained by the use of OpenNI [OpenNI, 2011] and thus, pre-processing of tracking data is proposed before being used for scene understanding, since tracking data is not robust and noisy. Moreover, the proposed scene understanding approach does not incorporate geometric information and thus is solely based on long-term tracking data (position and pose) of humans. With the proposed approach, a scene can be modeled according to the functionalities being used by the human and does not need any supervised training, nor prior knowledge of the scene structure.

**Behavior Monitoring:** behavior monitoring within the field of computer vision mainly focus on the analysis of surveillance cameras in order to detect abnormal behavior [Candamo et al., 2010, Popoola, 2012]. From a practical point of view, the detection of changes in behavior within the context of AAL is important in order to provide immediate help - however, different taxonomies are used [Chaaraoui et al., 2012]. Within this work, the taxonomy of Chaaraoui et al. [Chaaraoui et al., 2012] is used, since it classifies

motion, action, activity and behavior according to its duration and degree of semantics. According to this, behavior is defined as personal habits and daily routines over the course of weeks and months. Providing assistance is one of the goals of AAL and thus systems to provide assistance during the ADL are available (e.g. during hand-washing [Hoey et al., 2010]). Since these systems can monitor the performance during ADL over time, they are able to monitor the behavior. However, approaches focus on one or more ADL and analyze the performance of specific and pre-defined ADL [Nguyen et al., 2005, Chung and Liu, 2008, Duong et al., 2005]. Hence, they do not consider the overall health status of the person and thus only draw conclusions about the performance of ADL, but not their holistic behavior. Behavior monitoring providing a holistic view of a person's behavior is either analyzing activity patterns [Monekosso and Remagnino, 2010, Jain et al., 2006, Virone and Sixsmith, 2008] or detecting abnormal inactivity [Cuddihy et al., 2007, Floeck and Litz, 2008, Moshtaghi and Zukerman, 2014]. These approaches are mainly based on the use of sensor data from smart homes (i.e. passive infrared sensors, door sensors) and only little work uses computer vision [Nait-Charif and McKenna, 2004].

This thesis proposes a novel behavior monitoring approach by using computer vision in order to allow a fast and simple sensor placement to avoid construction work. The proposed behavior monitoring approach does not only consider temporal aspects, but also integrates spatial aspects and the context of the scene by introducing a human-centered scene understanding approach. Thus results in the automatic detection of ROIs and the possibility to model the temporal behavior within "interesting" regions. In contrast to other approaches, the proposed approach detects ROI fully automatically, without the need of any intervention by learning and adapting to the person's behavior over time. With the proposed framework, the persons overall health state can be observed since the framework does not focus on pre-defined activities, but learns a holistic behavior model over time. Changes within the global behavior of the person (e.g. reduced mobility) are automatically detected, not focusing on manually pre-defined activities. Moreover, the proposed approach incorporate short-term, mid-term as well as long-term knowledge in order to monitor and distinct deviations within the time frame of minutes, days and months.

CHAPTER 3

# Methodology

Within this thesis, the taxonomy of Chaaraoui et al. [Chaaraoui et al., 2012] is used, defining behavior as long-term analysis including high semantic information. Following this definition, daily routines of people are of interest and thus, the context of AAL is chosen in order to apply and evaluate the proposed behavior model. With the proposed behavior model, deviations from the daily routine of older persons are detected, since deviations are mainly the result of changes within their health state. Hence, the early detection of deviations is important, since then appropriate countermeasures can be taken. However, this implies that not only long-term deviations (duration of months) within behavior need to be detected, but also critical events on short-term (duration of minutes), since these events interrupts daily routines and immediate reaction (e.g. from a caretaker) is needed.

The proposed framework combines temporal modeling of the daily behavior (i.e. modeling the activities throughout a day) together with spatial knowledge of the scene (i.e. where activities are performed) in order to consider both, the time as well as the place where daily routines are performed. However, this work does not classify activities into pre-defined classes, but introduces a generic behavior model, not focusing on a set of activities, but on behavior. This is done in order to propose a self-learning framework, where no annotated data is needed to train the system. Hence, temporal and spatial knowledge is obtained automatically, allowing a high flexibility of the framework. Moreover, no activity classes need to be defined, nor any manually pre-defined routines are needed (e.g. person get up from bed, walks to the bathroom, walks back to the bed), allowing to introduce a generic model, indicating changes of the health status of older persons.

Figure 3.1 summarizes the structure of the proposed model: the aim of the proposed behavior monitoring is the detection of changes within the daily routines. Changes within the daily routines can be classified as short-term (i.e. critical events interfering immediately with daily routines), mid-term (i.e. changes within daily routines on a per day basis, e.g. deviation within meal intake) or long-term deviations (i.e. changes in

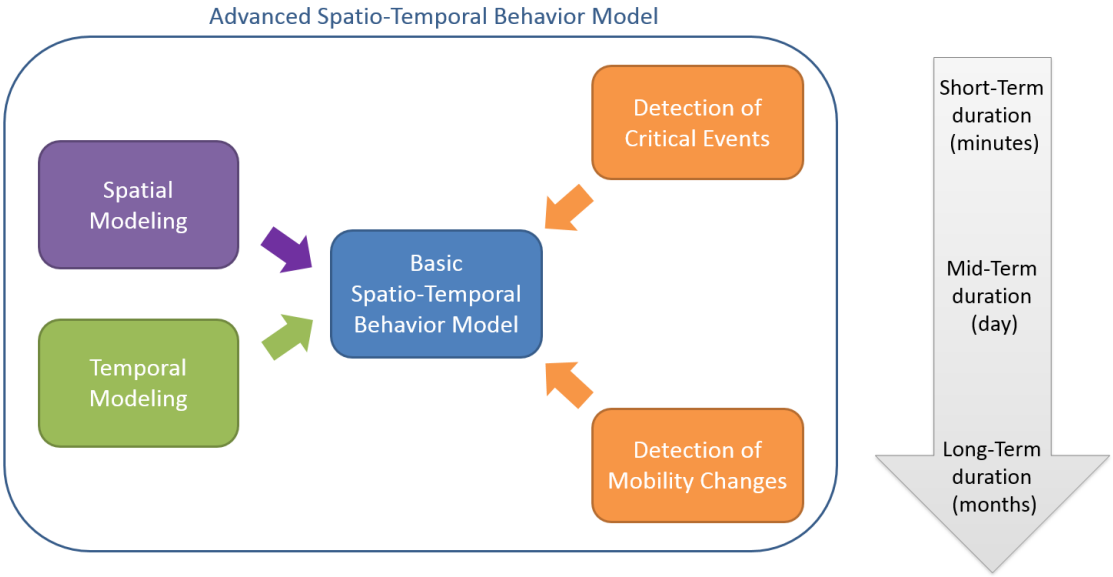



Figure 3.1: Overview of the thesis

mobility over months). Due to the combination of spatial and temporal information, a basic spatio-temporal behavior model, focusing on mid-term duration, is proposed. In order to incorporate also short-term critical events, interrupting daily routines of elderly people, a fall detection approach is incorporated into the behavior model. Adding long-term analysis of mobility allows to extend the proposed model to the long-term range, thus resulting in a spatio-temporal behavior model, focusing on short-term, mid-term and long-term range, allowing to model the behavior of the elderly on different time ranges. Due to the introduction of activity histogram comparisons as well as a human-centered scene understanding approach, spatial and temporal modeling is improved, resulting in the proposed advanced spatio-temporal behavior model.

The Asus Xtion pro sensor is used in this work, since it provides continuous data, allowing to not only monitor data over time (temporal aspects), but also contextual aspects within a scene can be considered (spatial aspects) and thus, only one single sensor is used. Moreover, with a 3D sensor critical events can be detected robustly and allows to use 3D tracking data for all modules of the proposed framework. When applying the behavior model in real-world scenarios, only one single sensor needs to be installed - having the advantage that the proposed behavior model can be applied within the context of AAL easily. However, the use of the Asus Xtion pro is restricted to indoor environments, since sensors based on structured light only works indoors due to interferences with the infrared parts of the sunlight.

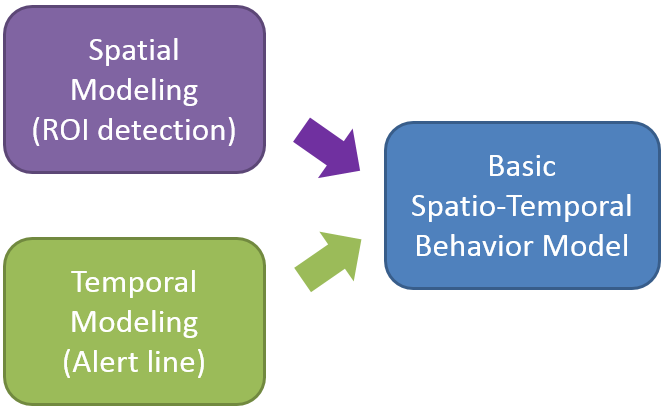



Figure 3.2: Workflow

## 3.1 Basic Spatio-Temporal Behavior Modeling

In order to obtain a first version of a behavior model, spatial and temporal approaches from the literature are combined in order to introduce a novel spatio-temporal behavior modeling approach. The proposed approach introduces the use of spatial information in combination with temporal aspects in order to enhance the accuracy of detecting abnormal inactivity and to reduce the number of false alarms. In contrast to the state-of-the-art, a depth sensor together with a tracking algorithm provided by the OpenNI SDK [OpenNI, 2011][1] is used to collect tracking data. The use of a depth sensor is motivated by the capability to combine the proposed approach together with other approaches (e.g. fall detection systems [Mastorakis and Makris, 2012]) without the need for additional sensors. In contrast to Cuddihy et al. [Cuddihy et al., 2007], no door sensors are used. Thus, no information about the absence of a person due to vacation, shopping, etc. is available. However, the information from door sensors can be integrated into the system in order to improve the accuracy.

The workflow of the proposed approach is illustrated in Figure 3.2: in the first step, ROI $R_i$ are detected by the algorithm. In the second step, alert lines are calculated based on the approach of Cuddihy et al. [Cuddihy et al., 2007], but for each spatial region $R_i$ individually.

### 3.1.1 Region of Interest Detection

Tracking data is captured in world coordinates relative to the sensor and is distorted. In order to obtain accurate results, the tilt of the depth sensor is obtained from OpenNI [OpenNI, 2011] and used to rectify the motion data. Pixel-wise accumulation of motion

[1]Since PrimeSense is not supporting the OpenNI project any longer, please note that the proposed approach is fully independent from third party companies since the work is based on high level 3D tracking data obtained from depth images. Hence, other depth cameras and tracking algorithms can be used in order to obtain the long-term tracking data.

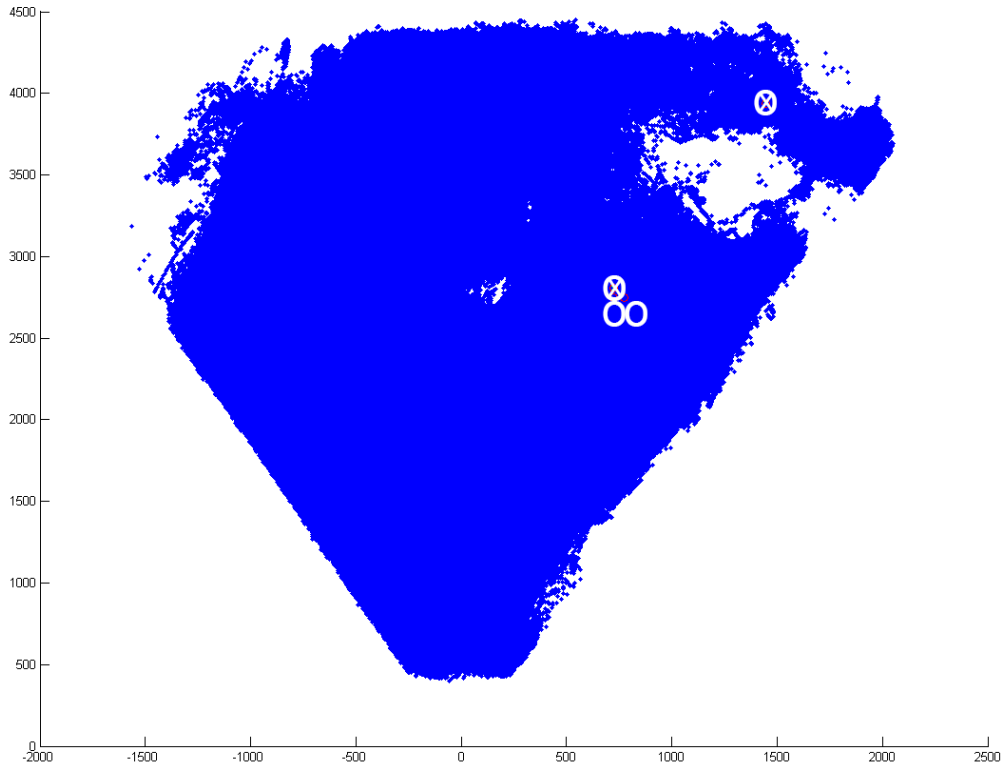



Figure 3.3: Top view of motion data and detected regions of interest (dataset 1): circles represent the initially calculated ROI by applying a threshold whereas crosses mark the final ROI centers after using non-maxima suppression

data indicates ROI and thus the pixels (and their surroundings) with the highest amount of motion data are defined as ROI. This is achieved by using a grid-based approach, calculating a 75x75 histogram of the motion data. In order to retrieve interesting regions of the scene, all bins with motion data higher than 60% of the maximal motion are pre-filtered as ROI. This initial regions are refined by applying non-maxima suppression, where similar region centers are eliminated. As a result, the center of $i$ regions $R_i$ containing a high amount of motion are obtained and are thus used as ROI. Figure 3.3 depicts the top view of a scene including motion data. Moreover, regions with a high amount of motion are detecting by applying a threshold and are marked as circles. After applying the non-maxima suppression, only relevant ROI are extracted, marked as X. These relevant regions are then used in the second step in order to calculate regional alert lines.

An example of the ROI is shown in Figure 3.4: although it seems that the two detected ROI are close together, the 3D representation in Figure 3.5 illustrates that different distances are taken into account and thus, distinct ROI are identified.

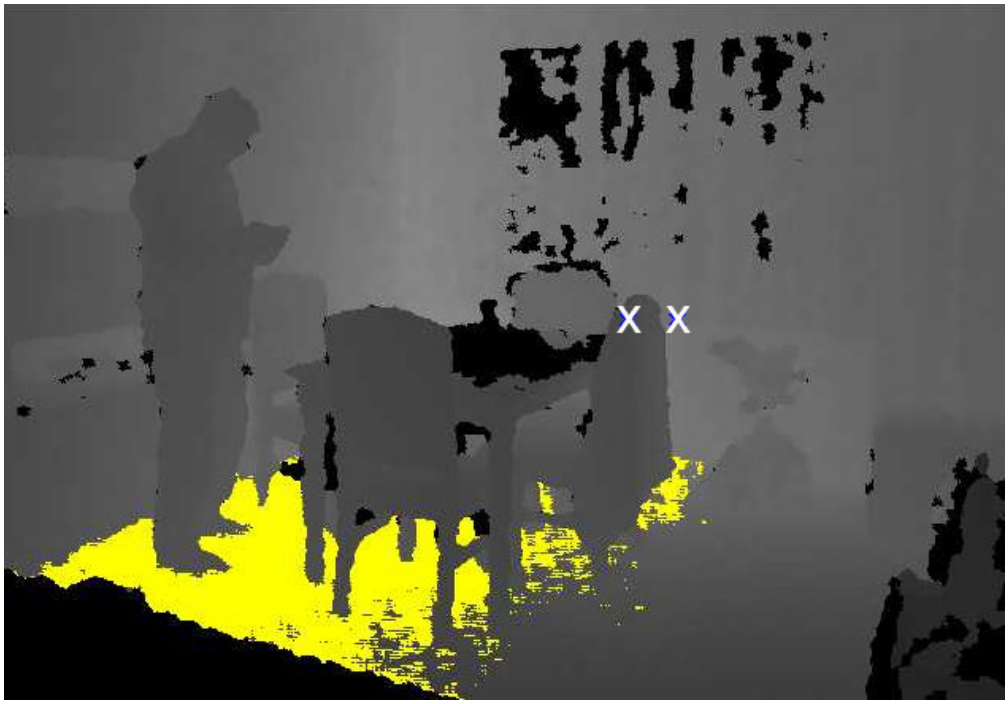


Figure 3.4: ROI (dataset 1): depth image with the ground floor marked yellow, ROI marked as X

Using the proposed region based approach allows to focus on specific regions being of high interest and eliminates a high amount of motion data from other areas. This information about the interesting areas of a scene can be either used to refine the modeling of inactivity and detect abnormal inactivity, or can be used to detect time depending activities. Regions with high amount of movement contain typical, regular activities (e.g. eating) since those activities are performed at the same place (e.g. at a table).

An example is shown in Figure 3.5: one ROI with the highest activity is located in the back right corner, where the window is located. The second significant ROI is detected at the chair at the right side of the table. Another possible, but not significant and thus not detected ROI is located at the opposite of the table where another chair is placed. All ROI can be explained and interpreted very easily: ROI 1 at the window indicates regular ventilation of the flat, whereas ROI 2 at the table reflects regular meal consumption. Since data is obtained from a flat where an elderly couple live in, the possible ROI at the opposite of the table indicate the area of the second person during food consumption. This indicates that the proposed approach is able - in combination with further knowledge of the scene - to not only monitor activity in general, but to model specific activities (e.g. food consumption) and the behavior of different persons individually (i.e. it is assumed that person A sits on the right side of the table whereas person B usually sits on the left side of the table).

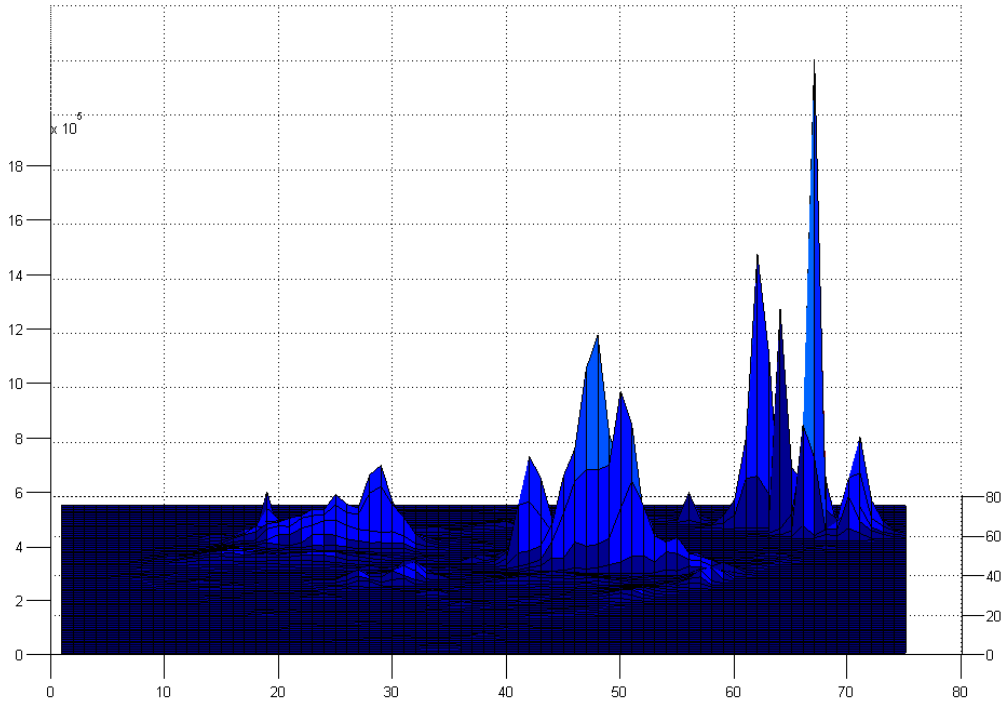



Figure 3.5: ROI (dataset 1): 3D visualization of the 75x75 histogram

### 3.1.2 Alert Line Calculation

The inactivity profile (i.e. alert line) is calculated for each region $R_i$ during the training phase, using the method of Cuddihy et al. [Cuddihy et al., 2007]. Only activity within this region and their surroundings, defined by a maximum tolerance radius $r$ (e.g. $r$ =25 cm), is taken into consideration. In comparison to the approach introduced by Cuddihy et al. [Cuddihy et al., 2007], regional alert lines contains a higher amount of inactivity since motion information is only analyzed in a small area of the scene, defined by the ROI.

After training the alert line, activity is analyzed in each region $R_i$ individually and compared with the corresponding alert line $A_i$. The comparison is done per time interval and an alarm is triggered if the duration of inactivity is above the threshold of the alert line for this interval. Hence, abnormal inactivity (i.e. deviation from the trained inactivity profile) is detected already at the end of the time interval allowing to provide immediate alarms. Abnormal inactivity is defined as reduced movement at time intervals, where a high amount of movement is present (calculated during the training phase). This abnormal inactivity can be caused due to illness or other physical impairments (not being present during the training phase) or the absence of the person.

## 3.2 Detection of Critical Events

The detection of critical events allows to add short-term aspects to the proposed behavior model since the behavior model focus on daily routines of the elderly. Critical events within short-term are interrupting daily routines and thus are of high interest since immediate help is needed. Moreover, experienced falls results in a higher fear of falling [Legters, 2002], thus changing the behavior of the older person. Zweng et al. [Zweng et al., 2010] show that the accuracy of their fall detection approach is higher when using a 3D reconstruction of the person, but having the main drawback of needing a calibrated camera setup. Therefore the 3D reconstruction of a person, using the Asus Xtion pro instead of multiple cameras is proposed in this thesis.

When using the Asus Xtion pro as 3D sensor, the low-level vision tasks motion detection, foreground / background segmentation as well as pose estimation are preprocessed in the SDK [2] [Shotton et al., 2011]. As a result, high-level data (i.e. coordinates of specific body junctions) are accessed directly. Since the use of the integrated pre-processing steps offers high-level data, no low-level vision algorithms (e.g. foreground / background segmentation) need to be applied anymore. The proposed algorithms calculate the orientation of the person's major axis and the height of the spine (relative to the ground floor) as features. In contrast to other works [Rougier et al., 2011, Pramerdorfer, 2013], feature analysis is not performed on a low-level using the camera picture or the depth image, but the proposed features are directly applied to the skeleton information.

The proposed fall detection approach combines body shape analysis together with inactivity detection and 3D motion analysis. A fall is detected by analyzing the body orientation and the height of the spine. If a person is detected to be on the floor and is not able to get up on her/his own within a specified time, an alarm is triggered since the person is not performing any ADL (inactivity detection). The workflow is illustrated in Figure 3.6: starting with a depth image obtained by the Asus Xtion pro, skeleton information is extracted and the ground plane is estimated by OpenNI [OpenNI, 2011]. Although the skeleton information provided by OpenNI is optimized for upright poses, it also works with different poses (e.g. lying on the floor). Based on the coordinate data of the skeleton, features to determine the pose of the person (i.e. orientation of the body and distance to the ground) are extracted. A final decision about the pose of the person is made by a fuzzy logic framework. This approach is chosen in order to reduce the computational load and therefore no special hardware requirements are needed.

### 3.2.1 Feature Extraction and Body Orientation

Tracking information from body joints is used in order to detect the body orientation. In order to calculate the body orientation, the major axis of the person is estimated by approximating the coordinates of the shoulder, spine and hip by calculating the mean slope between these three skeleton coordinates, visualized with a line. Figure 3.7

[2] The sensor data can either be accessed with the official Microsoft SDK or with the open source SDK OpenNI.

illustrates the depth image with a person and the corresponding major axis. The ground plane parameters $a$, $b$, $c$ and $d$ of the ground plane $\epsilon$

$$ax + by + cz = d \tag{3.1}$$

are estimated by using the OpenNI [OpenNI, 2011] SDK. Due to the combination of tracking data and the estimated ground plane, the following features are calculated:

- *Similarity between the body orientation and the ground plane:* the pose is estimated by calculating the similarity of the person's major body orientation and the ground floor. The inner product between the vector representing the body orientation $\vec{b} = \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix}$ and the normal vector of the ground plane $\vec{n} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$ is calculated in order to obtain the angle between the vectors, resulting in a pose estimation

$$P(\vec{b}, \vec{n}) = \begin{cases} \text{lying}, & \text{if } |\vec{b} \cdot \vec{n}| < t_{lying} \\ \text{upright}, & \text{else} \end{cases} \tag{3.2}$$

  If the orientation of the person is parallel to the ground floor, and thus the inner product is smaller than the threshold $t_{lying}$, the person is defined to be "lying" (either on the floor or on the bed). Please note that the person may be classified as lying while he or she is bending down, since only the orientation with respect to the ground floor is analyzed in this step. If the major orientation is orthogonal to the ground floor, the person is in the position "upright". Please note that the result of the inner product is zero, if both vectors are orthogonal to each other. This holds true if a person is lying on the floor since in this case the normal vector of the ground plane and the body orientation are orthogonal to each other.

- *Spine distance to the ground floor:* the distance $d$ between the spine and the ground floor is calculated using the Hesse normal form, allowing to determine whether the person is lying on the floor or on the bed. The distance $d$ between the spine skeleton joint $S$ and the ground plane $\epsilon$, represented by its normal vector $\vec{n}$, is defined as

$$d(S, \epsilon) = \left| \frac{\vec{n} \cdot \vec{SP}}{|\vec{n}|} \right| = |\vec{n_0} \cdot \vec{SP}| \tag{3.3}$$

  where $P$ is an arbitrary point on the plane $\epsilon$. The integration of this feature is essential, since otherwise it is not possible to determine if a person is lying on the bed or on the floor, which results in false alarms.

The use of depth and skeleton data allows to generate an abstract visualization in 3D, depicting the ground floor, shoulder (center), spine and hip (center) of a person, depicted in Figure 3.8. The use of the similarity of the body orientation and the ground floor is illustrated in Figure 3.8a and Figure 3.8b: in contrast to a person being in an upright

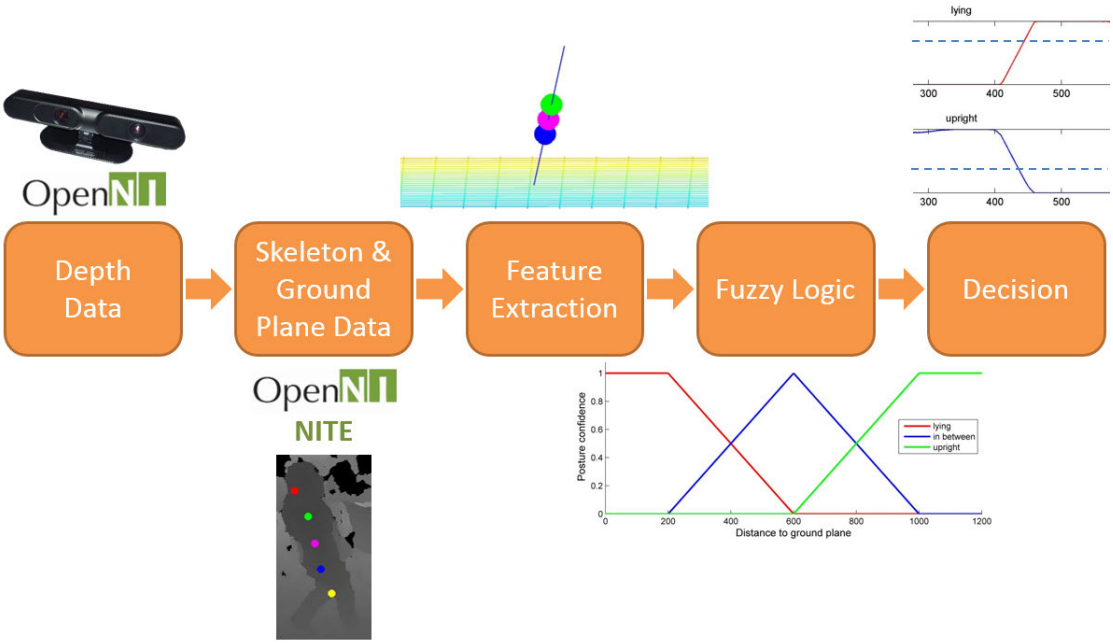



Figure 3.6: Workflow of the proposed fall detection approach

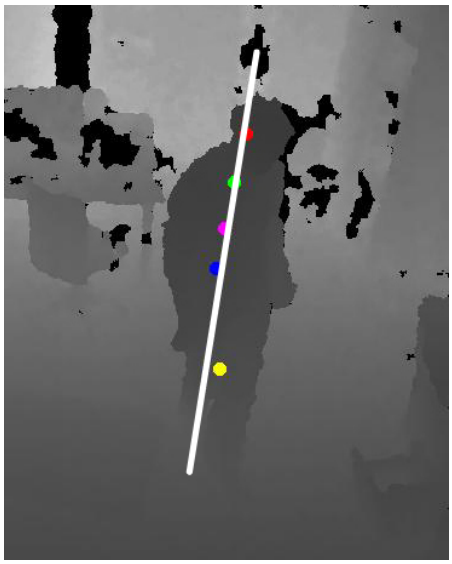
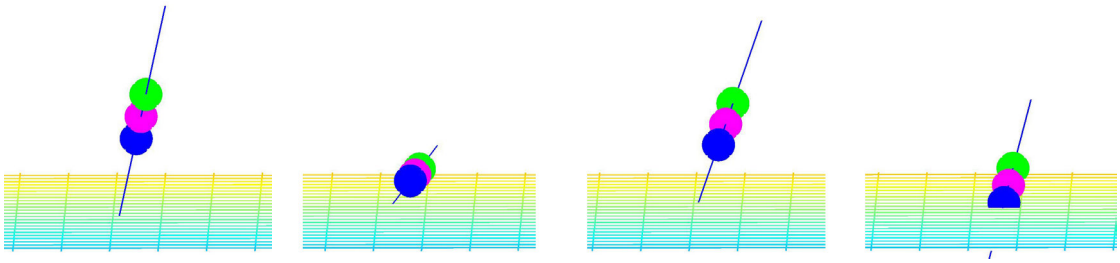


Figure 3.7: Major axis calculated using data obtained by the Asus Xtion pro



(a) Person in an upright position (b) Person lying on the floor (c) Person sitting on a chair (d) Person sitting on the floor

Figure 3.8: Abstract visualization of different poses with respect to the ground floor

position (shown in Figure 3.8a), a person lying on the floor is shown in Figure 3.8b. Therefore the similarity of the body orientation is used as a feature to distinguish between these poses. To be able to distinguish between similar activities, e.g. sitting down in the bed/chair (Figure 3.8c) and sitting on the ground (Figure 3.8d), the height of the spine is used as additional feature since in both scenarios the orientation of the body is the same. Therefore the following approaches are proposed: (1) mapping the 3D body joint coordinates to the 2D depth image and calculating the features using image coordinates; (2) analyzing features directly in the 3D space using world coordinates. For both approaches, two different thresholds are used: a similarity threshold as well as a threshold for the height of the spine in order to differentiate between a fall and a normal ADL.

### Body Orientation calculated in Image Coordinates

This approach uses the 3D skeleton information and maps the coordinates to a 2D image space. The orientation of the major axis is calculated using the coordinates of the head, shoulder, spine, hip and knee joints. Using the least squares algorithm to fit a straight line to the data points results in the orientation of the major axis. Afterwards, the angle between this line and the horizontal line is calculated. For calculating the height of the spine, an estimation of the ground plane is needed. The ground plane is estimated using the v-disparity map [Labayrade et al., 2002, Zhao et al., 2007]. The basic idea of this approach is that the depth linearly increases on the ground floor. Hence, the depth information of all pixels is analyzed and those pixels having a linear increase of depth are part of the ground plane. This approach assumes that the ground plane is visible in the depth map. After creating the ground plane estimation (which only needs to be done once per scene), the distance of the spine to the ground plane is calculated.

### Body Orientation calculated in World Coordinates

The proposed fall detection algorithm obtains the major orientation of the person's body in 3D space by using the skeleton information. For calculation of the orientation the

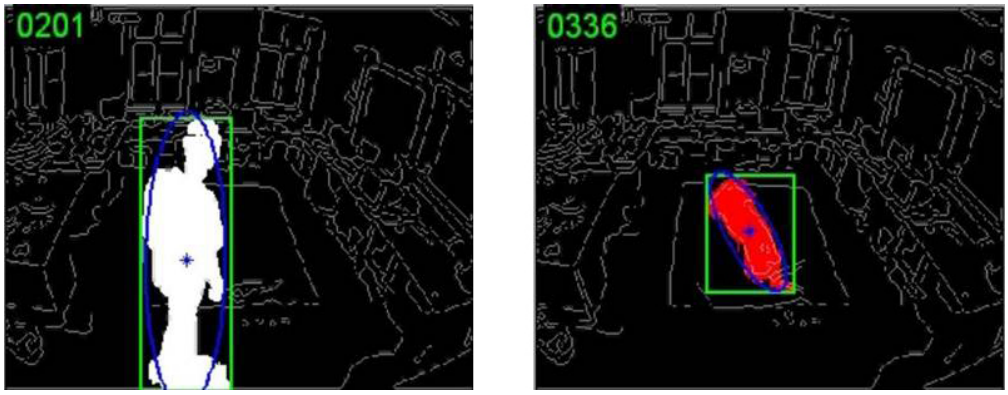



Figure 3.9: Fall in direction of the camera [Zambanini et al., 2010]

head, shoulder (center), spine, hip and the mean position of the knees are taken into consideration. Furthermore, the 3D ground floor is estimated and the spine distance to the ground floor is calculated. If the major orientation of the person is parallel to the ground floor and the height of the spine is near the ground floor, a fall occurred.

Using 3D depth data and world coordinates overcomes limitations of 2D camera approaches, e.g. the problem of falling in direction of the camera as the distance to the camera is analyzed. Figure 3.9 depicts the similarity of a person in an upright position and a fall in direction of the camera. Using a 2D single-camera approach and the orientation of the major axis as single feature, a fall in direction of the camera cannot be recognized, because the orientation of the major axis does not change.

### 3.2.2 Fuzzy Logic

Similar to Anderson et al. [Anderson et al., 2009] and Zweng et al. [Zweng et al., 2010], the proposed approach analyzing the body orientation is extended with a fuzzy framework, where pose estimation is based on confidence values for the poses “upright”, “in between” and “lying on the floor”. In contrast to Zweng et al. [Zweng et al., 2010], pose estimation and fall detection is only based on features previously introduced and motion speed is not taken into consideration. This is done in order to not constrain the fall event, but to be able to detect a variety of falls - even those, which occur slowly.

To classify different poses, trapezoidal functions [Zadeh, 1965] are constructed in order to map a feature (e.g. distance to the ground plane) to a probability value, according to the defined function. For the construction of the function, thresholds $t_{dist1}, t_{dist2}$ and $t_{dist3}$ (distances to the ground plane) and $t_{sim1}, t_{sim2}$ and $t_{sim3}$ (similarity of orientation) are found empirically. Fuzzy logic functions are calculated for each pose individually and are defined as follows:

- Orientation similarity:

$$P_{lying} = \begin{cases} 1, & \text{if } |\vec{b} \cdot \vec{n}| < t_{sim1} \\ \frac{t_{sim2} - |\vec{b} \cdot \vec{n}|}{t_{sim2} - t_{sim1}}, & \text{if } t_{sim1} < |\vec{b} \cdot \vec{n}| \leq t_{sim2} \\ 0, & \text{else} \end{cases} \tag{3.4}$$

$$P_{inbetween} = \begin{cases} 0, & \text{if } |\vec{b} \cdot \vec{n}| \leq t_{sim1} \\ \frac{|\vec{b} \cdot \vec{n}| - t_{sim1}}{t_{sim2} - t_{sim1}}, & \text{if } t_{sim1} < |\vec{b} \cdot \vec{n}| \leq t_{sim2} \\ \frac{t_{sim3} - |\vec{b} \cdot \vec{n}|}{t_{sim3} - t_{sim2}}, & \text{if } t_{sim2} < |\vec{b} \cdot \vec{n}| \leq t_{sim3} \\ 0, & \text{else} \end{cases} \tag{3.5}$$

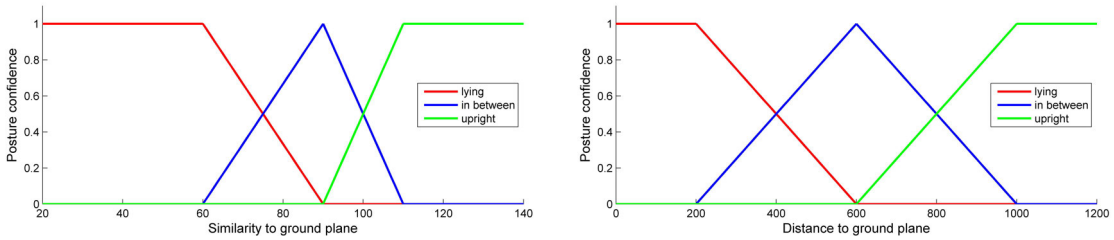
$$P_{upright} = \begin{cases} 1, & \text{if } |\vec{b} \cdot \vec{n}| > t_{sim3} \\ \frac{|\vec{b} \cdot \vec{n}| - t_{sim2}}{t_{sim3} - t_{sim2}}, & \text{if } t_{sim2} < |\vec{b} \cdot \vec{n}| \leq t_{sim3} \\ 0, & \text{else} \end{cases} \tag{3.6}$$

- Spine-ground distance:

$$P_{lying} = \begin{cases} 1, & \text{if } d(S, \epsilon) < t_{dist1} \\ \frac{t_{dist2} - d(S, \epsilon)}{t_{dist2} - t_{dist1}}, & \text{if } t_{dist1} < d(S, \epsilon) \leq t_{dist2} \\ 0, & \text{else} \end{cases} \tag{3.7}$$

$$P_{inbetween} = \begin{cases} 0, & \text{if } d(S, \epsilon) \leq t_{dist1} \\ \frac{d(S, \epsilon) - t_{dist1}}{t_{dist2} - t_{dist1}}, & \text{if } t_{dist1} < d(S, \epsilon) \leq t_{dist2} \\ \frac{t_{dist3} - d(S, \epsilon)}{t_{dist3} - t_{dist2}}, & \text{if } t_{dist2} < d(S, \epsilon) \leq t_{dist3} \\ 0, & \text{else} \end{cases} \tag{3.8}$$

$$P_{upright} = \begin{cases} 1, & \text{if } d(S, \epsilon) > t_{dist3} \\ \frac{d(S, \epsilon) - t_{dist2}}{t_{dist3} - t_{dist2}}, & \text{if } t_{dist2} < d(S, \epsilon) \leq t_{dist3} \\ 0, & \text{else} \end{cases} \tag{3.9}$$




(a) Orientation similarity (b) Spine-ground distance

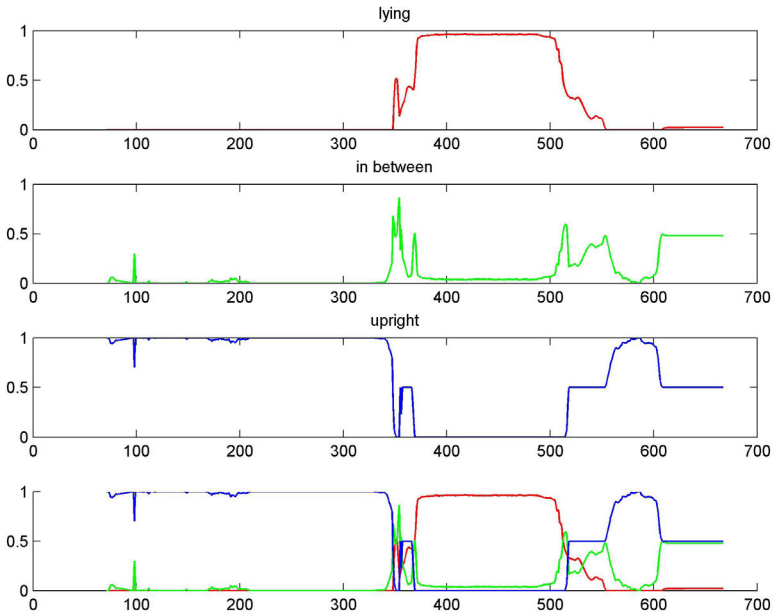
Figure 3.10: Definition of fuzzy boundaries

The trapezoidal functions for the posture confidence depending on the body orientation and the spine distance to the ground plane are depicted in Figure 3.10a and Figure 3.10b respectively. Posture confidences for the orientation and the height of the body are calculated independently in the first step. To get an overall decision whether a fall occurred, the confidence values are combined by calculating the arithmetic average. This combination results in three confidence values for the poses "upright", "in between" and "lying on the floor". The final decision whether a fall occurred is made by applying thresholds to the confidence values. Empirical results indicates that the "in between" pose does not result in additional information and thus, only the poses "upright" and "lying" are analyzed. A fall is detected if the upright probability is below a defined threshold and the lying probability is above a pre-defined threshold respectively.

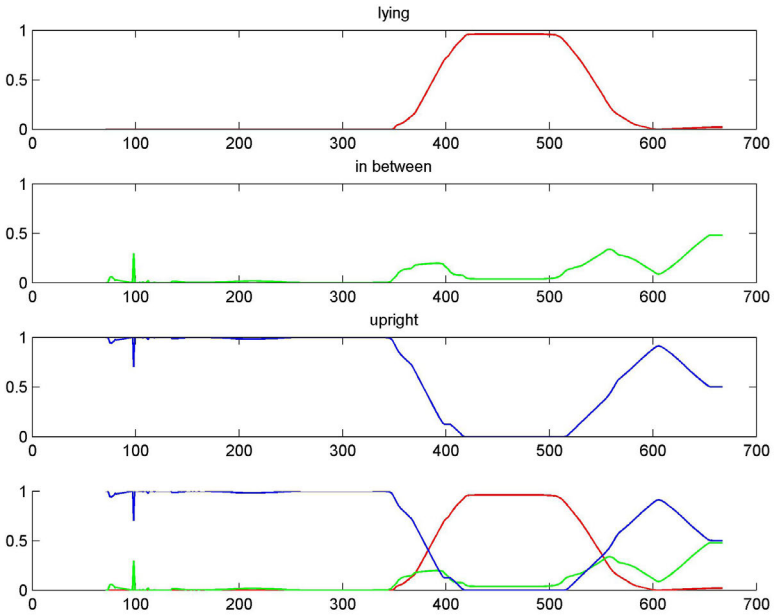
Eliminating outliers is achieved by analyzing the average of pose confidences over time (e.g. 50 frames). Figure 3.11a and Figure 3.11b illustrates and example of a simulated fall (starting at frame 350) and the getting up of the person (around frame 550). The probability for the lying pose is marked in red, the in-between pose is illustrated in green and the upright pose is depicted in blue. Exact and smoothed probabilities are visualized for each probability individually and a combined version illustrates the interdependency of these poses. Figure 3.11a depicts the exact probabilities for the three poses upright, lying and in between. As can be seen, glitches and outliers occur and thus a smoothed version, averaging the values over the duration of 50 frames, is used for analysis - depicted in Figure 3.11b.

## 3.3 Temporal Modeling

The focus of inactivity modeling within the context of AAL is the detection of critical events and changes within daily routines. However, long-term health changes resulting in mobility changes are not considered by state-of-the-art approaches. Hence, this work proposes an efficient approach in order to detect changes of mobility by introducing a new approach focusing on long-term analysis. However, not only reduced mobility can be detected but also increased mobility and thus the proposed approach cannot only be applied to detect deteriorations of the health conditions, but also to verify the success of countermeasures, trying to increase mobility (e.g. physiotherapy, social inclusion).




(a) Exact probabilities




(b) Smoothed probabilities (average of 50 frames)

Figure 3.11: Example of the fuzzy logic output during a fall event, depicting the probabilities of the poses lying, upright and in between over consecutive frames

However, increased mobility during the night is an indicator for the beginning of dementia and can be differentiated from increased mobility during the day. Since this work uses a 3D sensor to obtain data instead of ambient sensors, a novel approach to model activity is proposed. It is based on the temporal modeling using activity histograms instead of inactivity profiles in order to enhance the robustness when using vision-based data.

### 3.3.1 Long-Term Inactivity Analysis

This work proposes the analysis of changes in the long-term behavior, in order to detect a change of mobility. By performing long-term behavior analysis, long-term aspects are incorporated in the proposed behavior model. The algorithm of Cuddihy et al. [Cuddihy et al., 2007] detects an abnormal inactivity if the inactivity level is above the trained alert line. Hence, mid-term changes (i.e. in the range within one day) are detected (e.g. illness, fall). The alert line is calculated using a rolling window, i.e. the last 45 days are considered during the calculation of the alert line. This ensures that the algorithm is able to adapt to changes and thus reduces the number of false alarms. However, due to the adaption, a slow change of activity (e.g. over the course of a year) cannot be detected since the algorithm is adapted on the basis of the rolling window. In order to overcome this limitation, the comparison of alert lines (e.g. on a monthly basis) in order to detect general trends and significant changes of mobility is proposed.

The proposed method uses the calculation of an alert line described in [Cuddihy et al., 2007], resulting in an inactivity threshold $ALERT_i$ for every time interval $i$. The interval is set to one minute, resulting in 1440 intervals per day. Alert lines are stored at regular intervals (e.g. one alert line per month), resulting in $t$ different alert lines. The arithmetic average $\mu_i$ and standard deviation $\sigma_i$ of all alert lines $t$ are calculated for each interval $i$. During the training phase, all alert lines are incorporated to the calculation of the average alert line. After the training phase (e.g. three months), the deviation of the alert line to be added is calculated using the following rule: if the deviation of more than 25% (i.e. six hours) of the alert line intervals is within the range of $\mu \pm 2\sigma$, the average alert line is updated. If more than 25% of the alert line intervals are outside the range of $\mu \pm 2\sigma$, a significant change in long-term mobility is detected. Depending on the direction of the change (i.e. higher or lower inactivity compared to the reference alert line), a reduction respectively increase of mobility is reported. Moreover, since time intervals with increased or decreased mobility are specified, increased mobility during night (beginning of dementia) can be differentiated from increases during the day.

An example of a deviation is shown in Figure 3.12: the trained average alert line is shown as thick dashed line, whereas the other dashed lines indicate the range of $\mu \pm 2\sigma$. The alert line to be tested is visualized as cyan solid line and time intervals outside the range of $\mu \pm 2\sigma$ are visualized as magenta parts of the alert line. In this case it can be also visually verified that less than 25% of the alert line are outside the boundary and hence, no deviation in comparison to the normal mobility is detected.

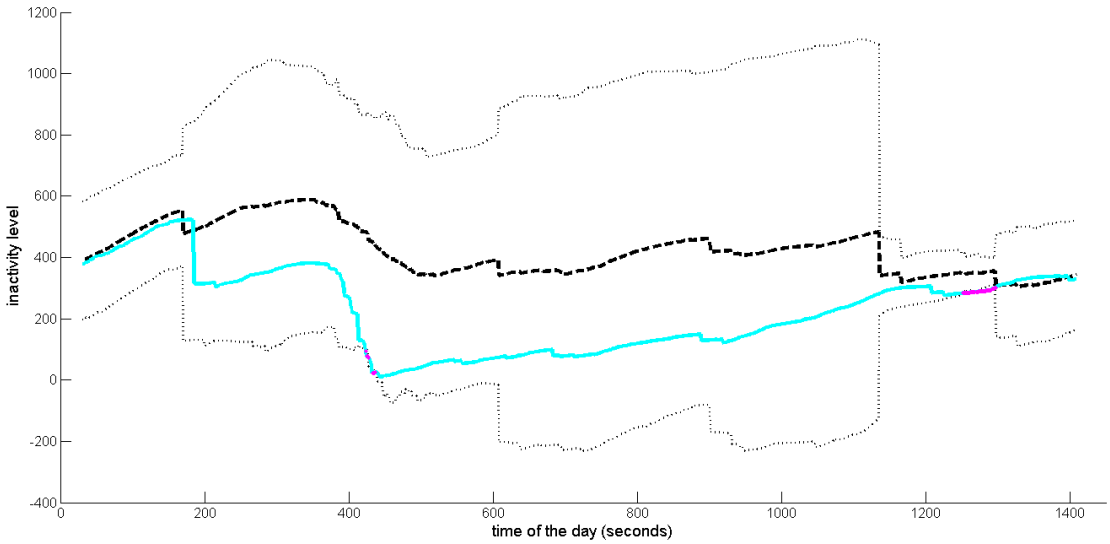



Figure 3.12: Example of an alert line with only minor deviations from the average alert line

### 3.3.2 Activity Histograms

Activity data is obtained by using tracking information from the OpenNI tracker NITE [OpenNI, 2011] and the Asus Xtion pro, but can be obtained by any arbitrarily tracking algorithm or sensor type (e.g. motion sensor). Tracking information consists of the center of mass of a person and the timestamp when motion is detected. Tracking is performed with approximately 30 fps (the frame rate of the Asus Xtion pro is not constant), resulting in a maximum of 30 events per second. In order to reduce the amount of data, activities are aggregated on a level of minutes. Hence, for each minute within an hour it is detected whether motion occurred within this minute or not. In the next step, the histogram is calculated based on the motion information per minute. Hence, the histogram depicts the amount of motion within a specific hour per minute (e.g. a histogram value of 30 is interpreted as motion has been detected for 30 minutes within this hour).

Floeck & Litz [Floeck and Litz, 2008] and Cuddihy et al. [Cuddihy et al., 2007] use inactivity profiles since they argue that it is difficult to combine different signals (start/stop signals and discrete events) to one common profile. This work focus on discrete events (i.e. motion detected / no motion detected), hence no inactivity profiles need to be calculated to fuse the sensor data. Instead, activity histograms are used to detect abnormal inactivity. Since histogram comparison is widely used in the area of image processing (e.g. image retrieval), activity histograms are introduced in this work in order to detect abnormal inactivity. Activity data is aggregated in histograms of 24 bins representing one day, resulting in one bin per hour. The number of bins was chosen to achieve a trade-off regarding the granularity of the approach, i.e. the activity is not analyzed in detail (e.g. per minute) and not per day, but on an hourly basis. Figure 3.13

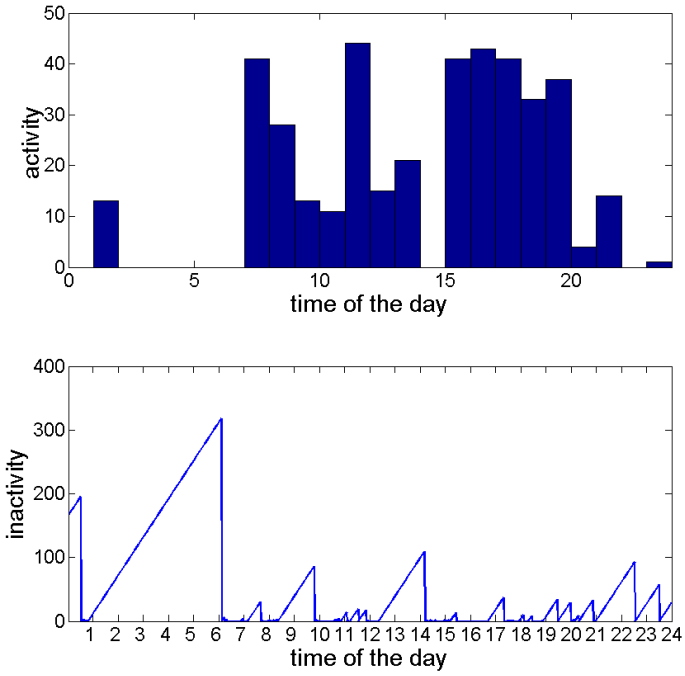



Figure 3.13: Example of an activity and corresponding inactivity profile

depicts an example of an activity histogram (top) and the corresponding inactivity profile (bottom). Since motion was detected during the night between one and two AM, the inactivity dropped to zero. The inactivity between 8:30 and 9:30 AM is better reflected in the inactivity profile, since only a smaller amount of activity is depicted in the activity histogram. But since a temporal buffer need to be added to detect abnormal inactivity, both representations are feasible.

During the training phase, the histograms $H_n$ for all $n$ training days are calculated. The average histogram $H_{ref}$ of all histograms $H_n$ is calculated and used as a reference for normal behavior. In order to model the variability of the training data, the distances $d_n$ between the $n$th histogram and the reference $H_{ref}$ are calculated.

The distance matrix $D_n$ represents the distances between all bins and the distance $d_n$ is the sum of all distances $D_{ij}$, shown in Equation 3.10.

$$d_n = \sum_{i,j=1}^{24} D_{ij} \tag{3.10}$$

The average distance $\overline{d}$ and standard deviation $\sigma$ are calculated from the training set and used as decision criteria during the test phase. A deviation from a normal daily routine is detected if

$$|d_t| \geq \overline{d} + \sigma \tag{3.11}$$

where $d_t$ denotes the histogram distance of the day to be analyzed to the reference histogram $H_{ref}$.

In order to provide a lateral buffer, the histograms are compared on a daily basis resulting in a delay of an alarm in comparison to the approach introduced by Floeck & Litz [Floeck and Litz, 2008], but reducing the number of false alarms. Since distance metrics are used in different applications to achieve different goals (e.g. Bhattacharyya distance is used to extract and select features [Choi and Lee, 2003], earth mover's distance is used for color histogram comparisons within content-based image retrieval [Rubner et al., 2000]), distance metrics are evaluated in order to analyze their performance and suitability for the proposed approach. For the calculation of the distances, the Euclidean, chi-square [Cha, 2008], earth mover's distance [Rubner et al., 2000], Bhattacharyya distance [Comaniciu et al., 2000, Bhattacharyya, 1943] as well as intersection [Swain and Ballard, 1991] and the Pearson Product-Moment Correlation Coefficient [Rodgers and Nicewander, 1988] are analyzed during the evaluation. They are defined as follows:

- Chi-square distance:

$$d(H_1, H_2) = \frac{1}{2} \sum_i \frac{(H_1(i) - H_2(i))^2}{H_1(i) + H_2(i)} \tag{3.12}$$

- Earth mover's distance:

$$d(H_1, H_2) = \frac{\sum_i \sum_j d_{ij} f_{ij}}{\sum_i \sum_j f_{ij}} \tag{3.13}$$

  where $f_{ij}$ denotes the optimal flow and $d_{ij}$ the ground distance

- Bhattacharyya distance:

$$d(H_1, H_2) = \sqrt{1 - \sum_i \frac{\sqrt{H_1(i) \cdot H_2(i)}}{\sqrt{\sum_j H_1(j) \cdot \sum_j H_2(j)}}} \tag{3.14}$$

- Intersection of histograms:

$$d(H_1, H_2) = 1 - \sum_i min(H_1(i), H_2(i)) \tag{3.15}$$

- Pearson Product-Moment Correlation Coefficient:

$$d(H_1, H_2) = \frac{\sum_i (H_1(i) - \overline{H}_1) \cdot (H_2(i) - \overline{H}_2)}{\sqrt{\sum_i (H_1(i) - \overline{H}_1)^2 \cdot \sum_i (H_2(i) - \overline{H}_2)^2}} \tag{3.16}$$

  where

$$\overline{H}_k = \sum_i \frac{H_k(i)}{n} \tag{3.17}$$

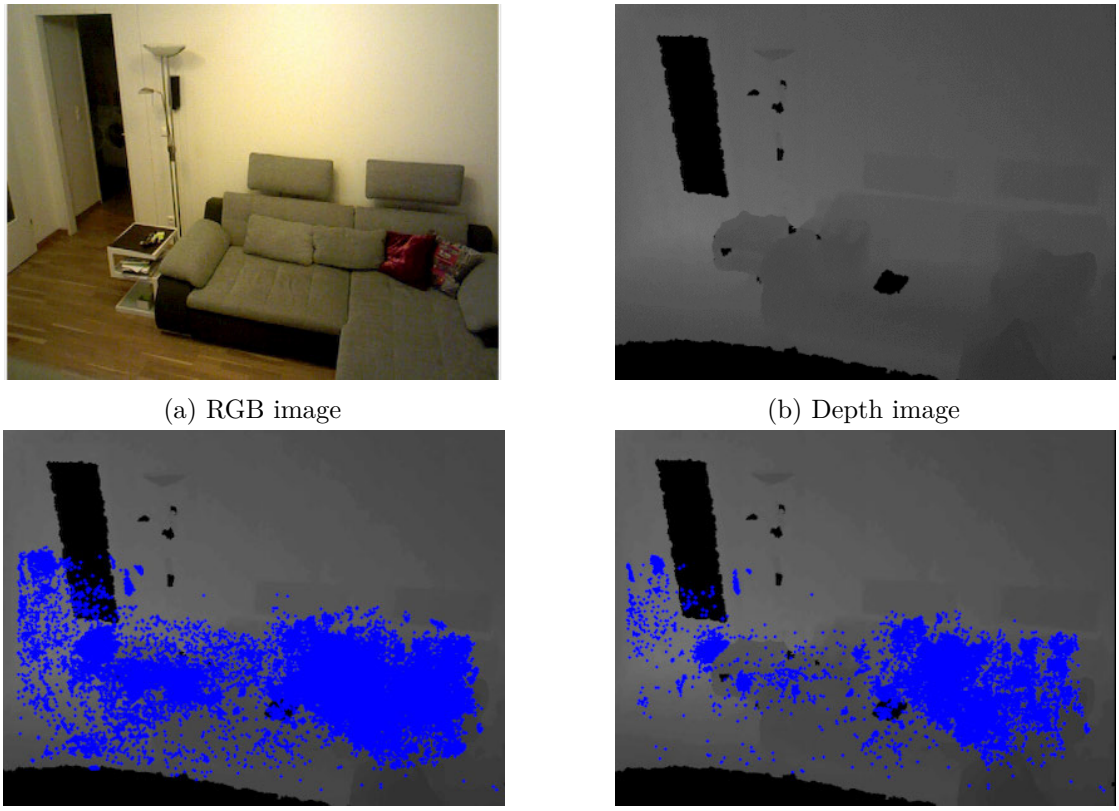

Figure 3.14: Workflow of the spatial modeling approach

## 3.4 Spatial Modeling

Spatial modeling in the initial behavior model is performed by the detection of ROI according to the amount of motion within different areas. Although this approach is feasible, only regions with a high amount of motion are considered, independently from their functionality. Hence, a more advanced approach incorporating knowledge about the scene itself is proposed. The proposed approach combines the advantages of 3D depth data together with long-term tracking and introduces a novel human-centered scene understanding, solely based on noisy tracking information. One of the aims of human-centered scene understanding is the detection of areas within the room, offering different functions for humans. The proposed approach focus on the functions "walking" and "sitting", since these are common functions within indoor scenes [Delaitre et al., 2012] and are used in the literature [Gupta et al., 2011, Fouhey et al., 2012, Jiang et al., 2013]: if an area is considered as walking area, no objects are present within this area (otherwise people could not walk there). If objects are present, people can either use them to relax by sitting or lying on them (e.g. bed, chair, sofa) or interact with them (e.g. table, wardrobe). Since the interaction with all possible types of objects is not in the scope of this thesis, this work focus on the detection of "walkable" and "sitable" areas within a scene, being supported by the room layout. In contrast to related work, the proposed approach is purely based on 3D long-term tracking information from a depth sensor, hence no geometric information is used and no manual annotations are needed in any step.

The workflow of the proposed approach is depicted in Figure 3.14: continuous depth data is obtained by the use of an Asus Xtion pro, the detection and skeleton tracking of the person is performed using the OpenNI SDK [OpenNI, 2011]. The 3D CoM position of a person within a frame is obtained from long-term tracking data, where filtering mechanisms to reject noisy and unreliable tracking data are applied. The filtered long-term tracking data is clustered according to the height (distance to the ground floor) into walking and sitting clusters. KDE together with non-maxima suppression and the calculation of a convex hull yields in hotspots of each activity region, representing areas being commonly used by humans. Please note that the proposed approach is only based on indoor tracking information, obtained by analyzing the skeleton of humans being tracked on the long-term range, i.e. over the duration of weeks and months.

Figure 3.15 and Figure 3.16 summarize the proposed workflow and illustrate a real-world example: the dataset is visualized as RGB (Figure 3.15a) and depth image



(a) RGB image

(b) Depth image

(c) Clustered CoM tracking data (unfiltered)

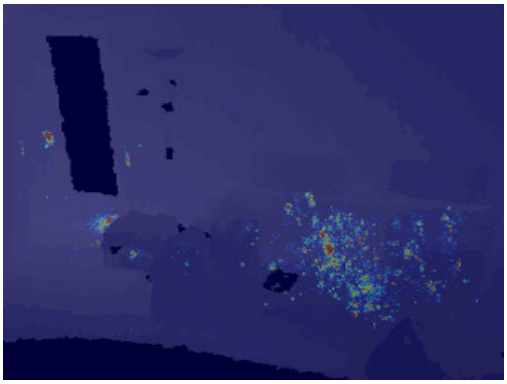
(d) Clustered CoM tracking data (filtered)

Figure 3.15: Illustration of the proposed workflow (dataset and filtering of tracking data)

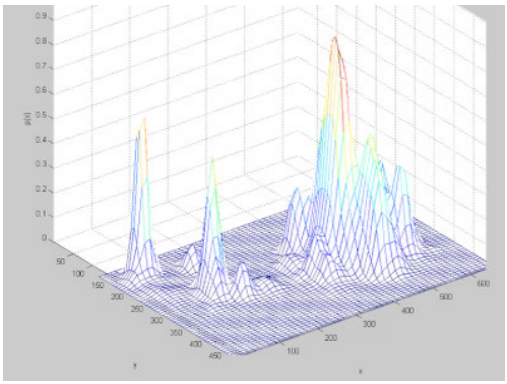
(Figure 3.15b) and is shown together with unfiltered and filtered tracking data of the CoM (Figure 3.15c respectively Figure 3.15d). Figure 3.16a visualizes hotspots within the tracking data. The use of a KDE together with non-maxima suppression are illustrated in Figure 3.18 and Figure 3.16c. The final result is depicted in Figure 3.16d. In the following sections, detailed information about the proposed filtering, clustering and modeling of the regions are provided.

### 3.4.1 Filtering

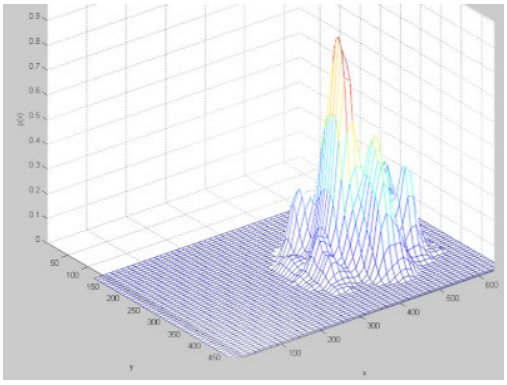
Long-term tracking data contains noise and tracking errors (e.g. furniture is incorrectly tracked as a person), influencing the results of the proposed approach. Hence, tracking data is filtered according to plausibility rules in order to ensure robust results. Although the filtering step removes a high amount of tracking information (i.e. all other activities and tracking information except walking and sitting), this does not influence the overall performance since the approach focus on long-term tracking, thus ensuring a sufficient number of reliable information is available. Filtering is based on the following three features: First, it can be assumed that a person being either walking or sitting is in an



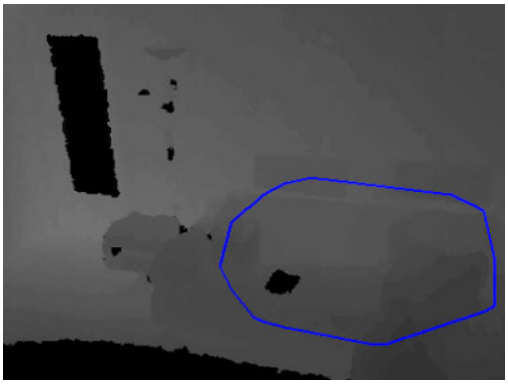
(a) Hotspots within filtered tracking data



(b) KDE



(c) KDE with non-maxima suppression



(d) Convex hull of KDE contour

Figure 3.16: Illustration of the proposed workflow (KDE and finding a ROI)

upright pose (body orientation). Second, the tracked CoM is within a plausible range of height and third, the confidence values of OpenNI are used to eliminate unreliable CoM values in order to ensure correct tracking data. During the initial filtering step the skeleton data is used in order to detect the orientation of the body, based on the features used to detect falls.

Since the OpenNI [OpenNI, 2011] tracker is optimized for the active interaction of the user with the sensor, as this is the case in commercial entertainment applications, tracking results contain jitter, are noisy or objects are tracked as humans by mistake (Figure 3.17). By the use of a filtering step as pre-processing and incorporating additional knowledge and constraints, the results of the obtained walking and sitting areas are significantly improved. The filtering process is based on the following three criteria:

- **Tracking confidence** (provided by OpenNI): a confidence value is provided by OpenNI and is used to eliminate tracking errors - if tracking data is not robust (e.g. person is occluded), it is discarded. The influence of this value and its efficiency in removing outliers is analyzed in the evaluation section.

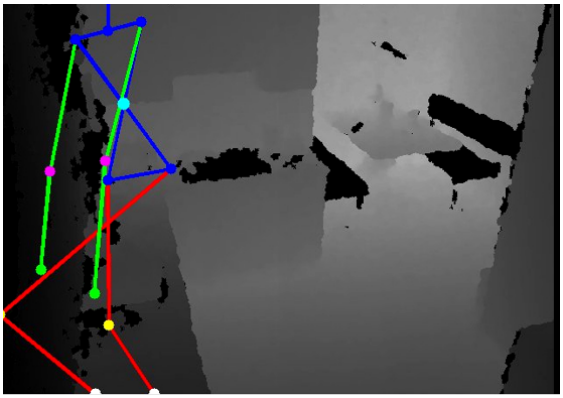


Figure 3.17: Tracking error: furniture is tracked as a person

- **Body orientation**: the orientation of the body indicates if the person is walking or sitting (based on the features proposed for the detection of critical events), since during walking and sitting an upright pose is assumed. The body orientation is calculated by the analysis of 3 body joints within the upper body with respect to the ground floor. A line is fitted to the center of the shoulder, the spine as well as the hip center and represents the orientation of the upper body. By using this feature, all tracking information where the person is not in an upright pose is filtered out (including tracking errors where an object is incorrectly tracked).

- **Person's height** (distance to the ground floor): when the CoM is tracked, the height of the CoM while a person is sitting or walking can be restricted, since it can be assumed that the distance is not higher than 2 meters or less than 20 cm from the ground floor. This broad range of thresholds (valid data from 20 cm to 2 m) can be narrowed down by automatically learning the optimal thresholds. However, since these thresholds depend on the scene and application, the use of a wide range is proposed in order to not being too restrictive and provide a generalizable approach.

Although these criteria may be straightforward, applying the proposed filtering allows to eliminate most tracking errors, especially where parts of the furniture are tracked as humans and thus improves the results of the consecutive processing steps. An example of this tracking error is shown in Figure 3.17, where a kitchen door is tracked resulting in a skeleton being tracked on the kitchen top (although the person is far away from the ground floor).

It should be noted that filtering outliers and focusing on only high qualitative data significantly reduces the amount of data since a high number of measurements is not considered. However, since the filtering is used for a long-term tracking approach, a high number of data points (over the duration of weeks or months) exist and thus, eliminating data points does not influence the result, since still enough data points are available. For example, in one dataset 18.000 data points were recorded over the duration of 18 days,

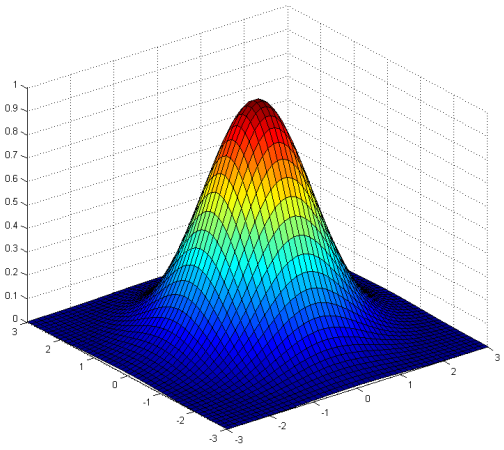
whereas in another dataset 120.000 data points were recorded within 34 days. After applying the proposed filtering mechanisms, dataset one was reduced to 7.000 data points and the second dataset was reduced to 30.000 data points. Although a high number of data points is discarded (11.000 and 90.000 respectively), the resulting sample is still big enough in order to obtain feasible results.

### 3.4.2 Clustering

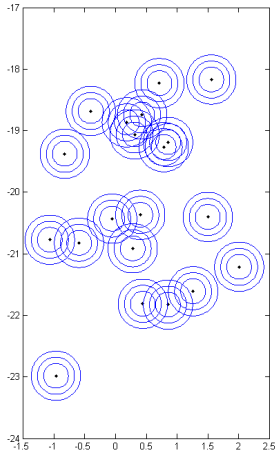
Interesting functional areas within indoor environments are sitting and walking areas, and thus are in the focus of research [Gupta et al., 2011,Delaitre et al., 2012,Fouhey et al., 2012]. Hence, tracking data is clustered on a per frame basis according to the height (i.e. distance to the ground floor) of the CoM (i.e. distance from the CoM to the ground floor) using the k-means algorithm to divide the tracking data into a sitting and a walking class ($k = 2$). Although tracking data does not only contain sitting or walking data, but also additional activities (e.g. picking something up from the floor, loading and unloading the dishwasher, etc.), it can be assumed that these activities are only minor activities (with respect to the duration) and that a walkable surface is mostly used for walking and that sofas and chairs are mainly used for sitting. When using long-term tracking information, these minor activities can be ignored and only major activities (walking and sitting) remain. Thus, the use of long-term tracking data is feasible to indicate walking and sitting areas within an indoor environment, without geometric information or training data. K-means is chosen since it is fast and the number of cluster is known in advance. However, this approach assumes that both, a sitting and walking area are within the field of view. If only one type of area is within the field of view (i.e. either sitting or walking), the obtained clusters need to be merged (e.g. by analyzing the distance of the cluster centers). Changing environments (e.g. moving chairs) are considered in two ways: first, temporarily changes (i.e. for a short period of time) does not influence the result, since long-term tracking data is obtained and thus tracking data represents the long-term behavior (i.e. is this area used as walkable area most of the time?). Second, permanent changes can be handled by a self-adapting approach and using a rolling window and thus only considering the last $n$ days/weeks, resulting in an adapted model.

### 3.4.3 Kernel Density Estimation (KDE)

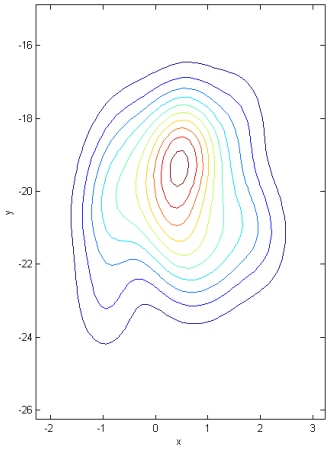
A KDE with a bivariate Gaussian kernel is performed in order to estimate the probability density function of the clustered data. Figure 3.18 illustrates the basic principle of the KDE with random data points: a Gaussian kernel, depicted in Figure 3.18a, is used. This kernel is applied to all data points, visualized in Figure 3.18b. Depending on the number of overlapping kernels, a probability density function is estimated - similar to histograms: the higher the number of overlapping kernels, the higher the result of the probability density function. The outcome can be visualized either with contour plots (Figure 3.18c where the height of the density function is represented by different colors) or as a 3-dimensional function, shown in Figure 3.18d.



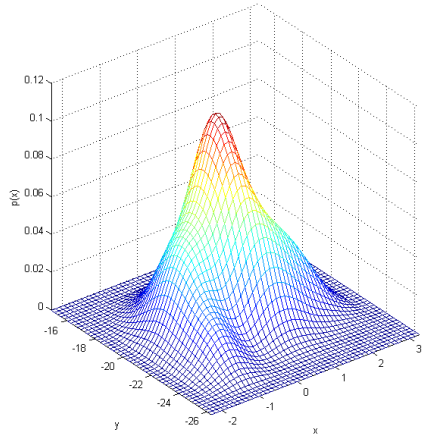
(a) Gaussian Kernel



(b) Random data points with their kernels



(c) KDE contour



(d) Probability density function

Figure 3.18: Illustration of KDE using random data points

The estimation is performed on both classes (sitting and walking) separately in order to model the probability density function of both classes. This step is performed to detect hotspots within the clustered data, i.e. areas being relevant for this class. Non-maxima suppression is applied to suppress irrelevant areas and allows to focus on the main areas, being representative for each class. The relevant area is obtained by applying a fixed threshold to the probability density function, where the resulting contour describes the functional areas within the scene. In order to aggregate smaller but similar areas, the convex hull of all contours is calculated to ensure a coherent area.

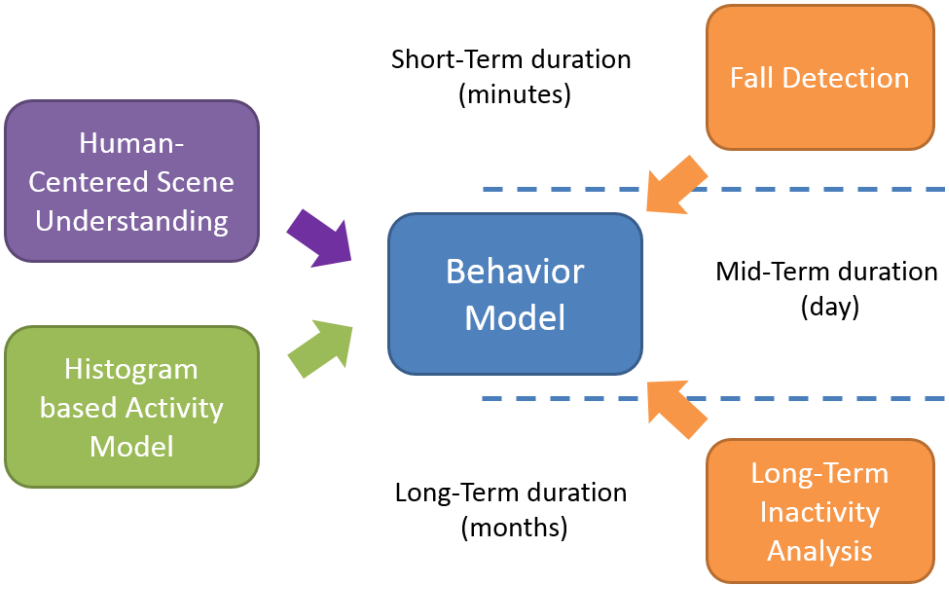



Figure 3.19: Combination of all modules to obtain the behavior model

## 3.5 Advanced Spatio-Temporal Behavior Modeling

Combining the proposed human-centered scene understanding approach together with the comparison of activity histograms results in a behavior model, focusing on mid-term duration (i.e. duration of days). Due to the integration of long-term analysis of inactivity, the proposed behavior model detects deviations of mobility on the long-term (i.e. over the duration of months). By complementing the behavior model with the proposed fall detection system, the model also considers short-term events (i.e. duration of minutes), interrupting or changing daily routines of elderly people.

Figure 3.19 summarizes the proposed behavior model: the novel histogram based activity modeling is combined together with the proposed human-centered scene understanding approach and form the foundation of the introduced behavior model. This model focus on the detection of abnormal events in mid-term duration, i.e. on a daily basis, allowing to detect deviations within different days. By incorporating the proposed fall detection approach, the model is extended to incorporate short-term events. Finally, by including the long-term detection of changes in mobility, the proposed behavior model is able to model spatio-temporal behavior on short-, mid- and long-term and thus provides a holistic approach to detect changes in the elderly's daily routines robustly.

The proposed approach combines the advantages of 3D depth data together with long-term tracking information and introduces a new local behavior model in order to detect health-related changes. Depth data is obtained by the use of an Asus Xtion pro, the detection and tracking of the person is performed using the OpenNI SDK [OpenNI, 2011]. The 3D position of a person within a frame is obtained from long-term tracking data,

where filtering mechanisms to reject unreliable tracking data are applied. The filtered long-term tracking data is clustered according to the height (distance to the ground floor) into walking and sitting clusters. KDE together with non-maxima suppression and the calculation of a convex hull yields in hotspots of each activity region. Finally, a behavioral model is learned locally within in each activity region.

### 3.5.1 Spatial Knowledge

Since tracking data is noisy, data need to be filtered in order to obtain reliable data. Filtering is based on the following three features:

1. it can be assumed that a person being either walking or sitting is in an upright pose (body orientation),
2. the tracked Center of Mass (CoM) is within a plausible range of height and
3. the confidence values of OpenNI are used to eliminate unreliable CoM values in order to ensure correct tracking data.

The filtering process is introduced in order to eliminate tracking errors, which are caused by objects being recognized as person and being incorrectly tracked. The height of the CoM is clustered for each frame using the k-means algorithm, resulting in two cluster: sitting and walking. A KDE with a bivariate Gaussian kernel is performed in order to estimate the probability density function of the clustered data. The estimation is performed on both classes (sitting and walking) separately to model the probability density function of the walking and the sitting class. This step is performed to detect hotspots within the clustered data, i.e. areas being relevant for this class. Non-maxima suppression is applied to suppress irrelevant areas and allows to focus on the main areas, being representative for each class. The relevant area is obtained by applying a fixed threshold to the probability density function, where the resulting contour describes the functional areas within the scene. In order to aggregate smaller but similar areas, the convex hull of all contours is calculated to ensure a coherent area.

### 3.5.2 Temporal Knowledge

Temporal Information is obtained from behavior monitoring based on histograms, allowing to model the behavior on a global level, considering only temporal but not spatial information. By extending this approach to model behavior on a local level, incorporating the knowledge about sitting and walking regions, a spatio-temporal approach is introduced. Hence, instead of calculating one global behavior histogram for the environment, a local behavior histogram is calculated for each detected region individually. At the beginning, a reference histogram $H^i_{ref}$ with 24 bins (one bin represents one hour) is trained for all regions $i$ separately. The reference histogram $H^i_{ref}$ is the average histogram of all $n$ training days within the respective region.

The distance between the histogram to be trained or tested $H_t^i$ and the reference histogram is calculated using the chi-square distance [Cha, 2008]. The average distance $\overline{d}_i$ represents the average distance between all $n$ training histograms and the reference histogram $H_{ref}^i$. In combination with the standard deviation $\sigma_i$, deviations from the reference histogram are detected according to the measurement previously introduced.

In other words, if the distance between the reference histogram $H_{ref}^i$ and the current histogram $H_t^i$ exceeds the threshold, a deviation from normal behavior is detected. This allows to model the behavior within sitting and walking areas separately and thus allows to get insights about the walking and sitting behavior individually. Moreover, a change of walking and sitting behavior can be detected (e.g. less walking and increased sitting behavior due to the reduction of mobility), which is not possible with the use of a global behavior model since actions and activities cannot be separated.

### 3.5.3 Privacy Aspects

When using autonomous systems within the context of AAL, privacy and the protection of data becomes an essential aspect to be considered. To ensure the dignity of the elderly, the anonymization of data is required. In order to consider privacy aspects, only depth data is processed. Hence, people and objects cannot be identified. Figure 3.20 depicts the depth images and thus automatically anonymized snapshots of the 3D Asus Xtion pro containing major body joints and the major orientation of the person. However, although privacy aspects need to be considered, a trade-off between anonymity and the possibility to react to emergency situations when being applied in practice need to be found. The more information is available for an emergency agent, the better and faster countermeasures can be taken. Figure 3.21 depicts different visualizations and discusses their relevance for usability in order to immediately react to alarms in practice. Although a video stream contains most information for humans, privacy is not protected at all and thus the use of video streams is not recommended. Depth images allows for a medium privacy protection while allowing to better understand the current situation than using a virtual top-down view or a more abstract visualization, where the person is represented by three dots, the major body orientation and the ground floor. From a computer vision perspective, depth images are a good trade-off ensuring privacy but allowing to interpret the situation. In practice, caretaker organizations offers a panic button in combination with a voice communication, allowing both, protecting the privacy and gathering information about the alarm.

## 3.6 Summary

**Fall Detection:** within this thesis, a novel fall detection algorithm based on high-level features in combination with a fuzzy framework is introduced. The proposed algorithm extracts features from depth information in order to calculate the major body orientation in 3D. In combination with the estimation of the ground floor, the distance and the orientation with respect to the floor is calculated. A fuzzy logic decision framework

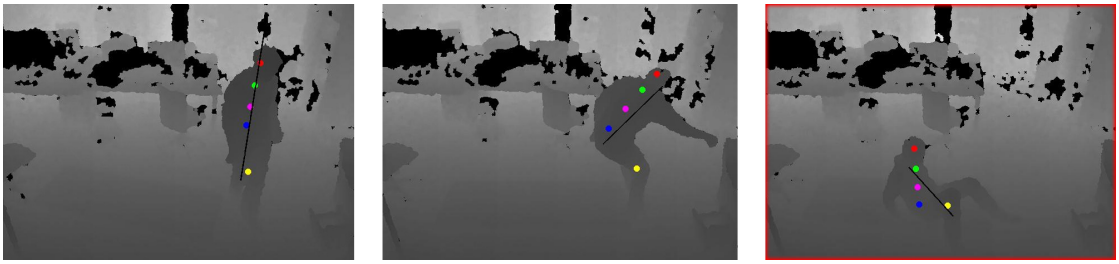


Figure 3.20: Anonymized snapshots using the 3D sensor (Asus Xtion pro)

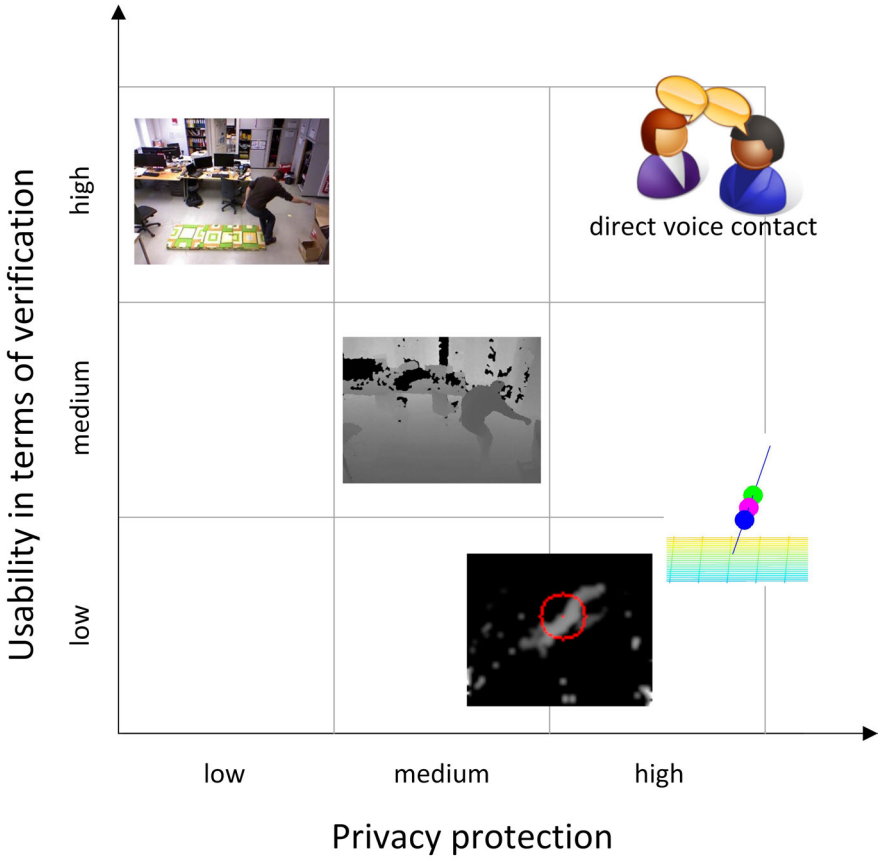



Figure 3.21: Privacy vs. usability aspects

is introduced in order to robustly differentiate the poses "upright" and "lying on the floor". The proposed fall detection approach based on depth information obtained by a 3D sensor was previously published in [Planinc and Kampel, 2012a, Planinc and Kampel, 2012b], is further evaluated in Stone and Skubic [Stone and Skubic, 2014] and integrated into the survey of Webster and Celik [Webster and Celik, 2014].

**Temporal Modeling:** state-of-the-art approaches focus on the detection of abnormal inactivity, by modeling the daily behavior and constructing an inactivity profile. With these approaches only critical events and rapid changes of the health status can be detected (e.g. falls, illness), but slow deteriorations cannot be detected. Hence within this work a long-term temporal model is introduced in order detect health changes, which occur slowly. Changes within the mobility of elderly people can be detected by monitoring their inactivity over a longer period of time (i.e. months) and comparison to previous models allows to detect decreased or increased mobility. A decrease of mobility is often health related (e.g. fear of falling after a fall), whereas an increase in mobility during the day indicates that the person is more active (e.g. due to successful rehabilitation). However, during night time, an increase in mobility indicates the beginning of dementia and thus is detected by the proposed approach, since the time of movement is considered in the model. Moreover, within this thesis it is shown that vision-based approaches perform better when using activity histograms rather than inactivity profiles. Hence, activity histograms are proposed and evaluated within this work, using different measurements for histogram comparisons. Long-term analysis of inactivity profiles is published in [Planinc and Kampel, 2013], the use of histogram comparisons is introduced in [Planinc and Kampel, 2014b].

**Spatial Modeling:** human-centered scene understanding allows to model the scene based on the functionalities, it is offering for humans. Within this thesis a novel approach using continuous depth data is introduced in order to detect "walkable" and "sitable" areas within a scene, since these are the most important areas when analyzing mobility. Moreover, filtering mechanisms in order to deal with noisy and unreliable tracking data from the Asus Xtion pro are presented, allowing to obtain a robust scene model based on long-term tracking data. The proposed approach does not consider short-term temporal changes within a scene, but is able to automatically adapt to long-term changes by using a rolling window approach. Moreover, no prior knowledge about the scene is needed, since the proposed approach automatically detects the most commonly used functional areas within a scene. The proposed filtering mechanisms are published in [Planinc and Kampel, 2014c], whereas the whole scene understanding approach is summarized in [Planinc and Kampel, 2015a].

**Spatio-Temporal Behavior Modeling:** the proposed spatio-temporal behavior model combines spatial and temporal approaches in order to obtain a generic and holistic model. Moreover, due to the integration of the detection of critical events, also short-term aspects are considered in order to provide immediate help during ADL. In contrast to state-of-the-art approaches, long-term analysis of mobility is incorporated into the proposed model, allowing to detect mobility changes over the duration of months, indicating e.g. reduced mobility due to health related issues or increased mobility due

to the beginning of dementia. However, no matter if focusing on the short-term, mid-term or long-term, the proposed behavior model is able to detect abnormal deviations from the behavior of the elderly, indicating changes within the health status of the person. When used in practice, thus allows to inform caretaker (formal or informal) and ensures, that appropriate measures are taken already at an early stage. The basic spatio-temporal behavior model is introduced in [Planinc and Kampel, 2014a], the advanced spatio-temporal behavior model is presented in [Planinc and Kampel, 2015b].

CHAPTER 4

# Results

Results are obtained from an evaluation of the proposed approaches on appropriate datasets and are compared to state-of-the-art methods. The comparison between different approaches is ideally based on the same dataset, in order to ensure accurate comparisons of different approaches. However, since the proposed approaches using the 3D sensor are novel, no publicly available datasets were available when the experiments were performed. Publicly available datasets containing depth information focus on the detection of human activities, e.g. the UTKinect-Action Dataset [Xia et al., 2012], the Cornell Activity Dataset [Sung et al., 2011] or the DailyActivity3D [Wang et al., 2012]. The aim of these datasets is to detect human actions and activities and hence, different activities and actions are recorded in short sequences, while the person is standing in front of the sensor. On the other hand, scene understanding datasets (e.g. NYU Depth Dataset v2 [Silberman et al., 2012], Berkeley 3-D Object Dataset [Janoch et al., 2013]) do not contain tracking information since traditional scene understanding approaches are based on geometric information. The proposed approaches neither focus on the detection of actions within short sequences, nor on the incorporation of geometric information. In contrast, the introduced approaches focus on the detection of critical events, human behavior on the long-term and thus, no datasets are available. In order to allow objective comparison in the future, the self-recorded datasets used for the detection of critical events and the evaluation of the spatial modeling are made publicly available[1] and are used by other authors [Pramerdorfer, 2013].

Different aspects of the scene need to be considered, e.g. the room layout, the placement of the sensor during the recording, as well as the duration in order to obtain sufficient data. Since the proposed approach to detect critical events is compared to a similar approach proposed by Zweng et al. [Zweng et al., 2010], a similar setup need to be chosen for the fall detection dataset. Moreover, an adapted version of fall scenarios is used

[1] http://fall-dataset.planinc.eu and http://tracking-dataset.planinc.eu, last accessed: March 31st, 2015

to obtain similar scenarios. It should be noted that the evaluation of proposed approaches on self-recorded datasets might lead to overfitting, i.e. obtaining high accuracy, because the approach perfectly fits the dataset - but obtaining low accuracy on other datasets. In order to minimize this effect, different strategies to evaluate the performance are chosen (e.g. cross validation, repeated sub-sampling validation, etc.), depending on the size of the dataset. The datasets as well as the methodology used for the evaluation are described in the respective sections.

The performance of the proposed approaches are evaluated using quantitative as well as qualitative methods, allowing to discuss the results in detail and to provide different views on the results. For quantitative results, either standard measures (e.g. f-score) or the number of false alarms are used to compare different approaches. For qualitative results, the visual outcome of the proposed approach is discussed, explaining and confirming quantitative results. Starting from the evaluation of a basic spatio-temporal behavior model, this model is extended, and the results of extensions are presented before finally the results of the advanced spatio-temporal behavior model are discussed.

## 4.1 Basic Spatio-Temporal Behavior Model

The evaluation of the basic spatio-temporal behavior model is performed on two different datasets obtained by an Asus Xtion pro in combination with the tracker provided by OpenNI. Depth data is available up to ten meters and thus the use of the Asus is feasible for most rooms in practice. For the analysis, an elderly couple was monitored over the duration of more than 100 days and tracking data was captured (dataset 1). For dataset 2, activity data of an elderly man was recorded over the duration of more than 50 days. Days without activity are removed in a first step in order to remove outliers, resulting in a dataset of 100 resp. 50 days. The depth image of the scene in dataset 1 is visualized in Figure 4.1, indicating two ROI. Figure 4.2 illustrates the depth image of the scene of dataset 2. The only ROI is detected at a table, indicate a high amount of motion close to the table (i.e. sitting down and getting up). Candidate ROI are illustrated in Figure 4.3, where only one ROI is detected to be significant.

Two regions of interest are detected in dataset 1 (Figure 4.1) and the alert lines obtained by training over a duration of 46 days are shown in Figure 4.4: Figure 4.4c shows the global alert line, whereas Figure 4.4a-4.4b show the alert lines to their corresponding regions 1 and 2. Since all alert lines use the same scale (but different offsets), the alert line in region 1 shown in Figure 4.4a is similar to the global alert line shown in Figure 4.4c, but provides a more distinctive change of inactivity, i.e. the difference from the maximal to the minimal value is larger. Furthermore, the alert line from region 2, depicted in Figure 4.4b, indicate activity between 9 and 11 AM. This information is completely lost in the global alert line and thus illustrates the advantage of the proposed region based approach, where each region is analyzed individually and thus spatial and more detailed temporal information is preserved.

Since the global alert line in dataset 1 is influenced by the high activity in region 1, further evaluation is performed to compare the performance of the global alert approach

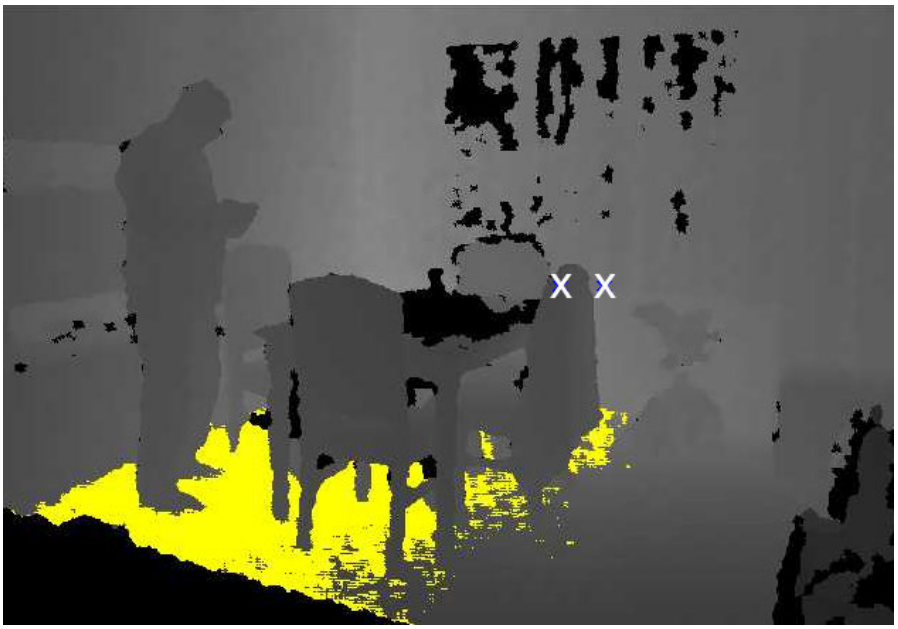


Figure 4.1: ROI (dataset 1): depth image (ground floor marked yellow)

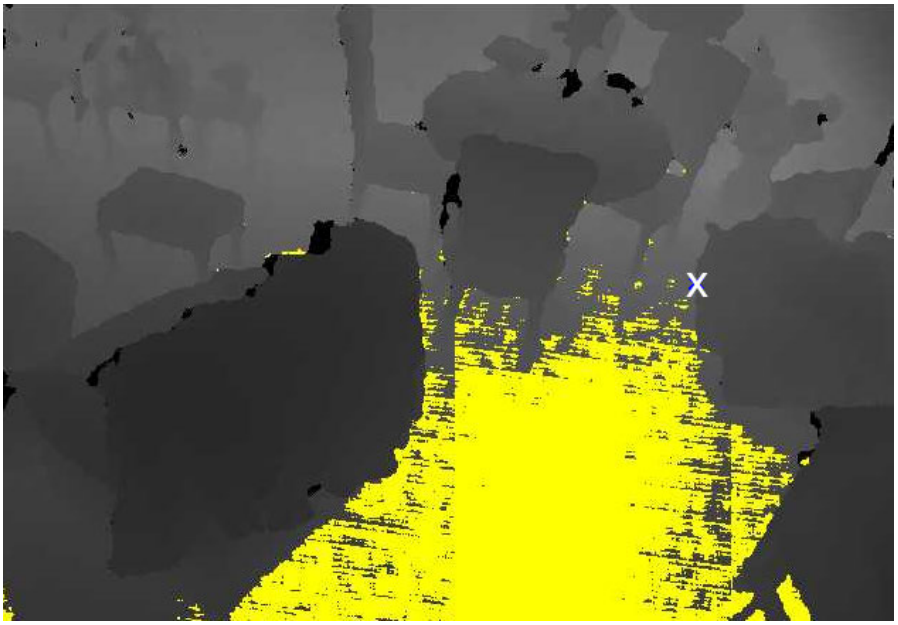


Figure 4.2: ROI (dataset 2): depth image (ground floor marked yellow)

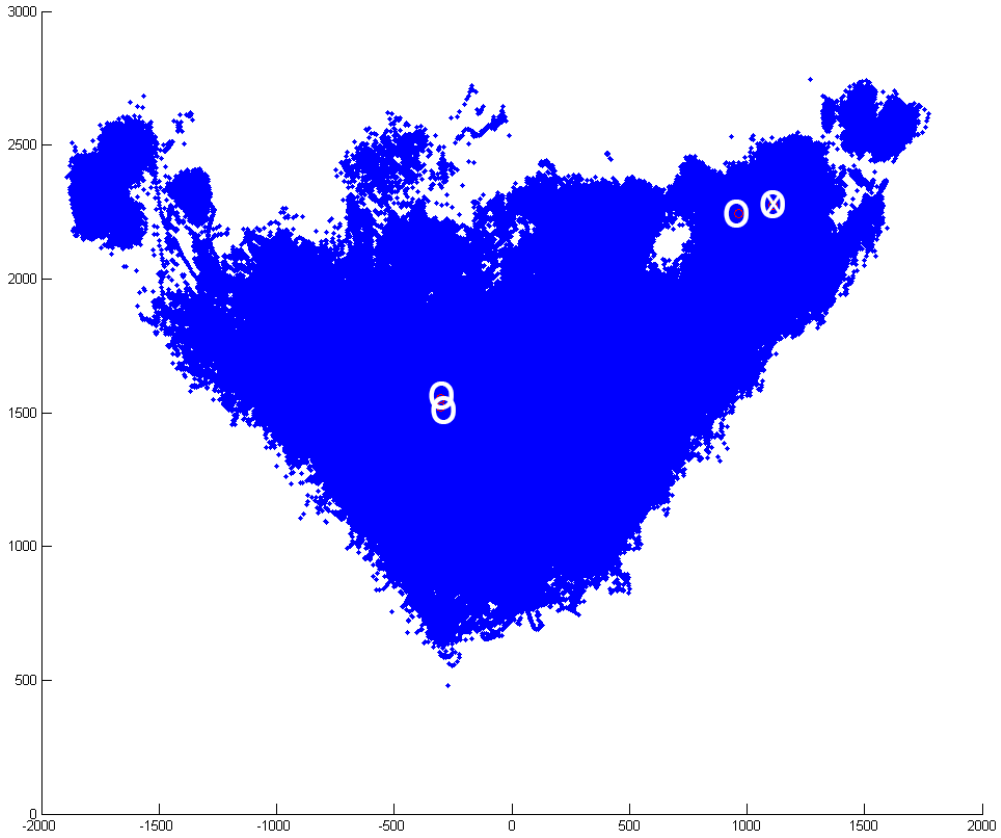
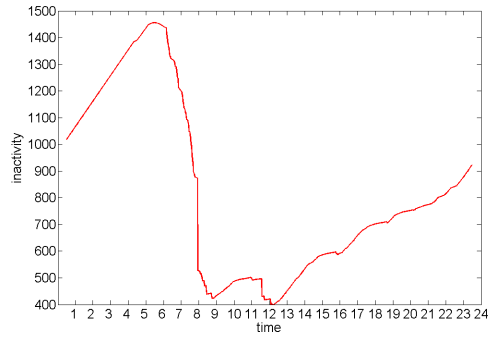


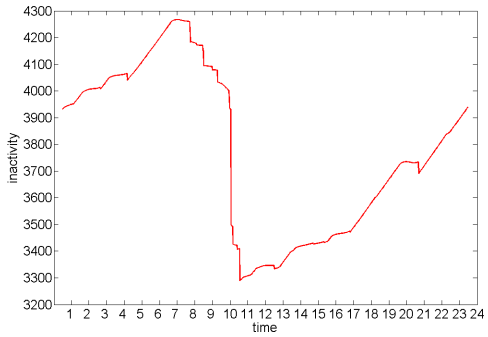

Figure 4.3: Top view of motion data and detected ROI (dataset 2): circles represent the initially calculated ROI by applying a threshold, crosses mark the final ROI centers after using non-maxima suppression

with the proposed region based approach in region 1. Evaluation is performed by calculating the number of alarms in region 1 and on a global level using cross validation and the number of false alarms are compared to the approach introduced by Cuddihy et al. [Cuddihy et al., 2007]. For dataset 1, nine rounds of cross validation were performed, for dataset 2 three rounds. For each round of cross-validation, the dataset is randomly split into a training and test set and the number of training data is varied from two days of training up to a training of 98 days. The rest of the data set is used as test set, hence resulting in a number of 98 test cases respectively a test set of two days. The results of all rounds are averaged and shown in Figure 4.5a (dataset 1) and Figure 4.5b (dataset 2). Multiple rounds of cross-validation are performed in order to avoid overfitting of data since the training data is randomly chosen multiple times.

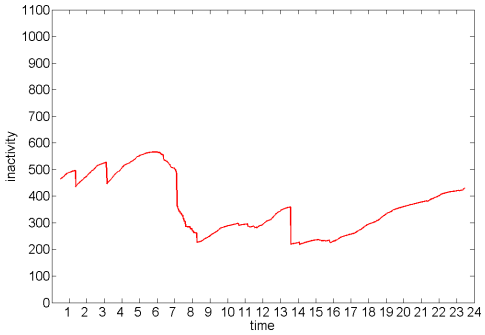
Figure 4.5a compares the proposed region based approach with the global approach introduced in [Cuddihy et al., 2007] based on dataset 1: the regional approach clearly outperforms the global approach since the number of false alarms is always lower. Please




(a) Alert line of region 1 (using the proposed approach)




(b) Alert line of region 2 (using the proposed approach)




(c) Global alert line without spatial information [Cuddihy et al., 2007]

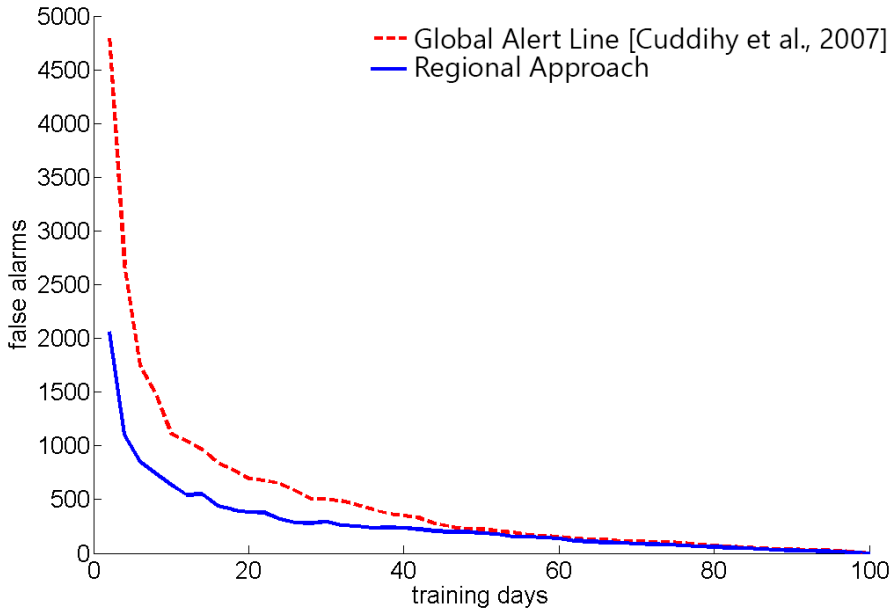
Figure 4.4: Comparison of global and regional alert lines

note that during the recording of the activity data no abnormal event was reported by the elderly couple, hence the number of false alarms should be zero. Thus, the number of false alarms indicates the performance of the algorithms.

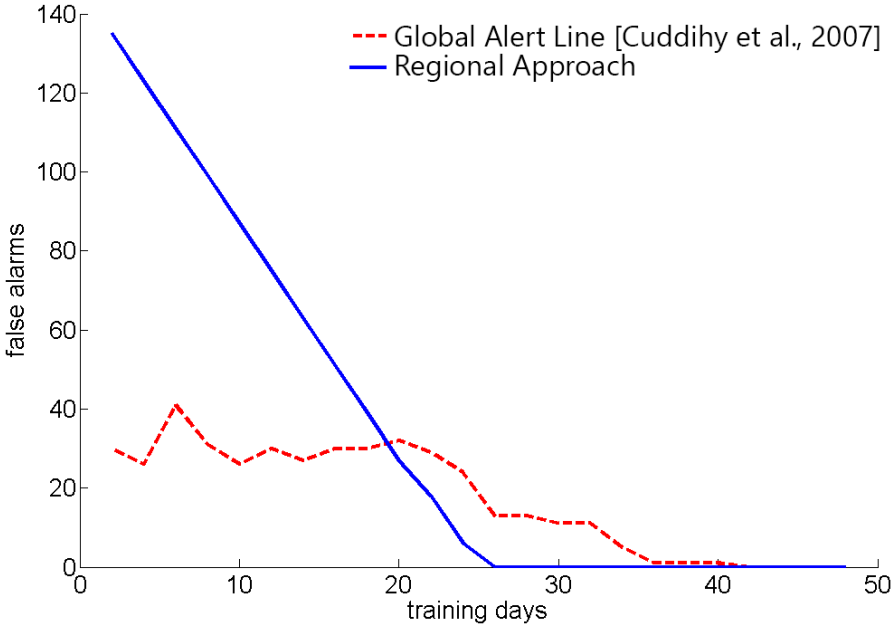
To verify the results, the region based and global algorithm are evaluated on a smaller dataset containing 50 days of activity data of an elderly man (dataset 2). The results, depicted in Figure 4.5b, again shows that the proposed algorithm outperforms the global approach introduced by Cuddihy et al. [Cuddihy et al., 2007], when using more than 20 training days. The false alarm rate drops to zero when using more than 25 days of training due to the small dataset available and possible overfitting (half of the data set is used for training).

## 4.2 Detection of Critical Events

Falls are simulated using an extended version of scenarios from Noury et al. [Noury et al., 2008], described in Table 4.1. The scenarios introduced by Noury et al. [Noury et al.,




(a) Dataset 1




(b) Dataset 2

Figure 4.5: Comparison of global and regional false alarm rate depending on the duration of the training

Table 4.1: Definition of scenarios similar to Noury et al. [Noury et al., 2008]

| Category | Description | Outcome |
|---|---|---|
| Backward fall | Ending sitting | **Positive** |
| | Ending lying | **Positive** |
| | Ending in lateral position | **Positive** |
| | With recovery | Negative |
| Forward fall | With forward arm protection | **Positive** |
| | Ending lying flat | **Positive** |
| | With rotation, ending in lateral position (left or right) | **Positive** |
| | With recovery | Negative |
| Lateral fall (to the left or right) | Ending lying | **Positive** |
| | With recovery | Negative |
| Neutral | To sit down on a chair, then to stand up | Negative |
| | To lie down on the bed, then to stand up | Negative |
| | Walking | Negative |
| | To bend down, pick something up, then to rise up | Negative |
| | To cough or sneeze | Negative |
| Additional sequences | To sit down on a chair, then fall while getting up | **Positive** |
| | To lie down on the bed, then to fall out of the bed | **Positive** |
| | Fall into camera direction | **Positive** |

2008] contain backward and forward falls, as well as lateral falls, ending in different positions. Moreover, ADL are described and classified as neutral scenarios, where no fall occurred. The additionally added scenarios are

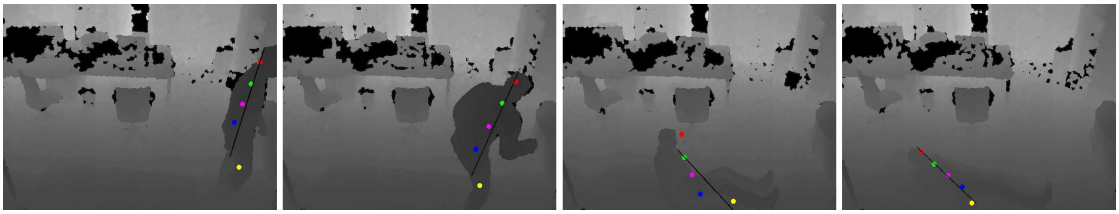
- "sitting down on a chair and fall while getting up",
- "to lie down to a bed and fall out of the bed" and
- "fall into camera direction".

These scenarios are added to enhance the quality of evaluation and are the outcome of discussions with end-user organizations. Furthermore, two scenarios were taken out from the original definition of Noury at al. [Noury et al., 2008], since the uniqueness of the outcome is not clear. The modification results in 18 different sequences, containing ten falls and eight no-falls. These scenarios are simulated by two subjects, simulating each scenario twice. This results in an overall set of 72 videos, containing 40 falls and 32 no-falls and are publicly available[2].

Figure 4.6a shows video frames taken out of a test sequence, showing a simulated fall according to the scenario "fall backward, ending lying on the ground". To prevent injuries, falls are simulated using a mat. The corresponding depth frames with skeleton points of the head, shoulder, spine, hip and the average of both knees are shown in

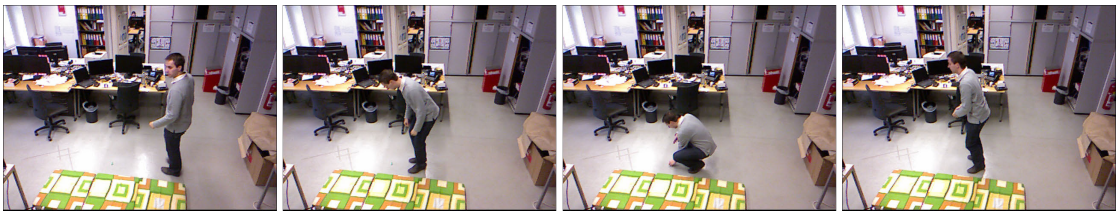
[2]http://fall-dataset.planinc.eu, last accessed: March 31st, 2015

(a) Video frames



(b) Depth and skeleton information

Figure 4.6: Frames of a scenario containing a simulated fall



(a) Video frames

(b) Depth and skeleton information

Figure 4.7: Frames of a scenario where the subject picks something up from the ground

Figure 4.6b. In contrast, video frames of an ADL are illustrated in Figure 4.7a, where a person picks something up from the floor. Due to the orientation of the body this scenario provoke false alarms, as the body may be close and parallel to the ground floor - the corresponding depth data of this sequence is depicted in Figure 4.7b.

To be able to evaluate the fall detection algorithm, tracking data is manually annotated to obtain ground truth information. Therefore the frame number where the fall begins and the frame number where the person is in a fully upright position are annotated. A True Positive (TP) is obtained if the algorithm detects the fall between the first frame where it begins and the last frame, where it ends. As the sequences do not only include

falls but also activities similar to falls, True Negative (TN) are marked (there is no fall and the fall detection algorithm does not detect a fall). Furthermore, False Positive (FP) and False Negative (FN) are analyzed by examining the results of the algorithm. FP occur if the person does not fall, but the algorithm detects a fall; a FN is obtained if the person falls, but the algorithm does not detect it. Each sequence contain one fall at most, but it is possible that the algorithm results in a TP (i.e. fall detected correctly) and a FP (i.e. fall detected without a person falling) within the same video sequence.

The quality of the algorithm is measured using the standard measurements recall, precision, F-score, true negative rate and accuracy. They are defined as follows [Van Rijsbergen, 1979]:

$$recall = \frac{TP}{TP + FN}, \tag{4.1}$$

$$precision = \frac{TP}{TP + FP}, \tag{4.2}$$

$$F - score = 2 \cdot \frac{recall \cdot precision}{recall + precision}, \tag{4.3}$$

$$truenegativerate = \frac{TN}{TN + FP}, \tag{4.4}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.5}$$

Within the context of AAL, TP are falls being detected by the proposed approach, whereas FP are ADL being incorrectly detected as a fall. ADL being not detected as fall results in TN and FN represent the number of falls, which were not detected by the system (i.e. missed). Recall describes the number of relevant falls which were detected, i.e. a recall of 1 indicates that all falls have been correctly detected (not stating how many ADL being detect as falls too). On the other hand, precision indicates the number of the detected falls to be relevant, i.e. a precision of 1 indicates that only real falls were detected (not stating how many falls were missed). The f-score is the harmonic mean between recall and precision and allows to provide a holistic view on the performance of the algorithm.

All tests are conducted under laboratory settings, the room setup is shown in Figure 4.8. The size of the laboratory is approximately 7×6 meters, whereas the camera field of view is set to an area of approximately 5.5×5.3 meters. The Asus Xtion pro sensor is placed in the middle of the wall at a height of 2.4 meters, which is a typical position for surveillance cameras. One frame of the room setup using the Asus Xtion pro to illustrate the camera field of view is shown in Figure 4.9.

### 4.2.1 Body Orientation

Results of the proposed approaches on 72 sequences are depicted in Table 4.2. The absolute values for TP, TN, FP and FN are shown and a comparison between image coordinates and world coordinates is presented. Using the previously introduced measures, a comparison of the results is shown in Table 4.3, indicating that the performance of both

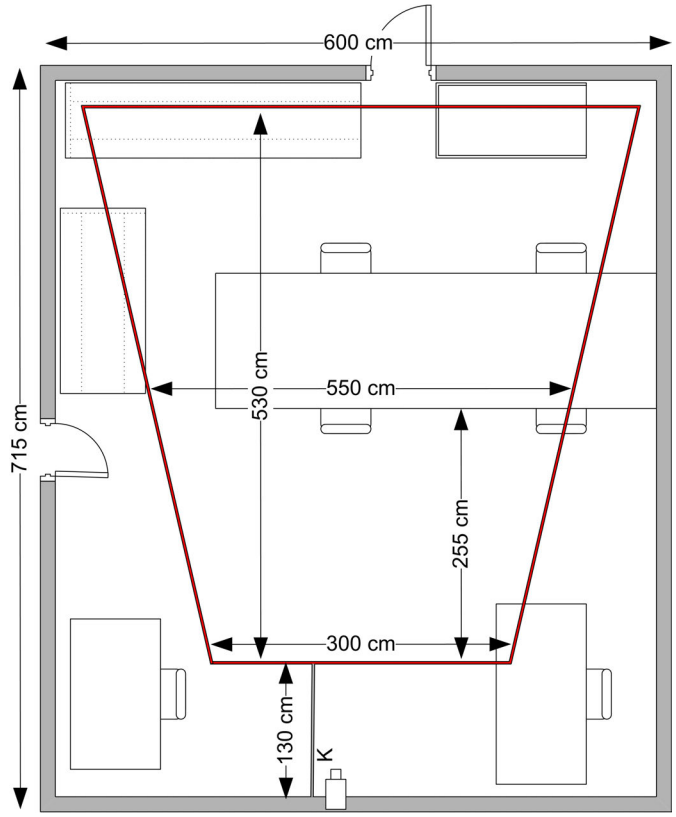



Figure 4.8: Room plan showing the room setup for the evaluation

Figure 4.9: One frame of the Asus Xtion pro Sensor, illustrating the camera field of view

Table 4.2: Results for evaluating our fall detection approaches

| | 3D sensor using image coordinates | | | | | 3D sensor using world coordinates | | | |
|---|---|---|---|---|---|---|---|---|---|
| | TP | TN | FP | FN | | TP | TN | FP | FN |
| Person 1 a | 4 | 8 | 0 | 6 | | 10 | 8 | 0 | 0 |
| Person 1 b | 10 | 7 | 1 | 0 | | 10 | 6 | 4 | 0 |
| Person 2 a | 9 | 8 | 0 | 1 | | 9 | 8 | 1 | 1 |
| Person 2 b | 8 | 8 | 0 | 2 | | 8 | 8 | 0 | 2 |
| | **31** | **31** | **1** | **9** | | **37** | **30** | **5** | **3** |

Table 4.3: Measures for evaluating our fall detection approaches

| | 3D sensor using image coordinates | 3D sensor using world coordinates |
|---|---|---|
| **recall** | 0.775 | 0.925 |
| **precision** | 0.969 | 0.881 |
| **f-score** | 0.861 | 0.902 |
| **true negative rate** | 0.969 | 0.857 |
| **accuracy** | 0.861 | 0.893 |

approaches is similar. However, analyzing the evaluation results in detail shows that at least in five sequences errors occur due to not correctly ending the tracking process when the person leaves the frame. Thus, improving the tracking of the skeleton will improve the obtained results. Assuming that tracking works correctly (i.e. filtering out the last frames of the videos where tracking problems occurred) lead to the results depicted in Table 4.4. Analog to Table 4.2, the absolute values for TP, TN, FP and FN are shown whereas the corresponding measures are shown in Table 4.5. After the elimination of the tracking errors, a comparison indicates that the use of the Asus Xtion pro as 3D sensor with world coordinates clearly outperforms the image coordinates based approach.

The evaluation of the introduced Asus Xtion pro based approach (together with image and world coordinates) is compared to results of the fall detection algorithm using a 2D sensor and a statistical model, introduced by Zweng et al. [Zweng et al., 2010]. The database used by Zweng et al. [Zweng et al., 2010] consists of 73 video sequences containing 49 falls and 24 video sequences with ADL (e.g. sneezing, picking something up, etc.). They tested a single as well as a multiple camera approach under laboratory conditions, resulting in the precision specified in Table 4.6.

The comparison of results is shown in Table 4.6. This comparison shows that the 2D sensor approach [Zweng et al., 2010] is outperformed by using the Asus Xtion pro

and exploiting depth information. The proposed algorithm is implemented in C++ and is able to detect falls in real-time, that is 30 fps on an Intel Core i7-2620M Quad Core CPU @2.7 GHz and 8 GB RAM.

### 4.2.2 Fuzzy Logic

The results of the extended approach combining the features body orientation and spine distance together with fuzzy logic is presented in this section. Experiments have shown that the pose "in between" is not needed for evaluation, as analyzing only two poses is sufficient. Results are presented using a Receiver Operating Characteristic (ROC) curve [Davis and Goadrich, 2006], depicted in Figure 4.10. Since the proposed algorithm is frame-based, "no fall" events occur in each sequence (even if it is a sequence containing a fall event), since most of the frames are "no falls" and only a few frames show the fall respectively a person lying on the floor. The ROC curve in Figure 4.10 is generated by varying the thresholds $t_{lying}$ and $t_{upright}$ and show, that the approach of Zweng et al. [Zweng et al., 2010] is outperformed. Although the evaluation is not based on exactly the same dataset (since the dataset of Zweng et al. [Zweng et al., 2010] is only camera based), the evaluation setting is very similar (laboratory setting is similar to Zweng et al. [Zweng et al., 2010]). Furthermore, similar to Zweng et al. [Zweng et al., 2010] the fall scenarios defined by Noury et al. [Noury et al., 2008] are used. The presented approach results in only one FP on the whole dataset whereas the ROC curve from Zweng et al. [Zweng et al., 2010] indicates a higher number of FP.

In summary, using fuzzy logic in combination with an Asus Xtion pro results in an accuracy of 98.6% on 72 videos, resulting in one FP in the whole dataset. This FP occurs due to a tracking error after a fall, since the person is not tracked correctly while getting up again. Hence, a second fall is detected within the same sequence but as this fall does not occur in the time interval specified in the ground truth annotation, it is marked as a FP.

## 4.3 Temporal Modeling

The proposed temporal approach of long-term mobility analysis is evaluated on the activity data of an elderly couple over the duration of 103 days for the activity histogram comparison and a subset of 100 days for the long-term inactivity analysis evaluation. The subset is chosen in order to split the training data into four equal parts, consisting 25 days of training data each, whereas the full dataset is used during the evaluation of the histogram comparison approach. The evaluation is based on tracking data obtained from the observation of the living room of an elderly couple over the duration of 103 days. The Asus Xtion pro was placed in the living room of the couple, monitoring the dining table and its surrounding area. This area was chosen since the couple performs regular food intake at the dining table and thus results in regular patterns. The monitored field of view is shown in Figure 4.11 and covers the area of the living room, where a table is used for food intake. Since only a small, but important and regularly visited area of the

Table 4.4: Results obtained after eliminating tracking errors

| | 3D sensor using image coordinates | | | | | 3D sensor using world coordinates | | | |
|---|---|---|---|---|---|---|---|---|---|
| | TP | TN | FP | FN | | TP | TN | FP | FN |
| Person 1 a | 4 | 8 | 0 | 6 | | 10 | 8 | 0 | 0 |
| Person 1 b | 10 | 8 | 0 | 0 | | 10 | 8 | 0 | 0 |
| Person 2 a | 9 | 8 | 0 | 1 | | 9 | 8 | 0 | 1 |
| Person 2 b | 8 | 8 | 0 | 2 | | 8 | 8 | 0 | 2 |
| | **31** | **32** | **0** | **9** | | **37** | **32** | **0** | **3** |

Table 4.5: Measures obtained after eliminating tracking errors

| | 3D sensor using image coordinates | 3D sensor using world coordinates |
|---|---|---|
| recall | 0.775 | 0.925 |
| precision | 1 | 1 |
| f-score | 0.873 | 0.961 |
| true negative rate | 1 | 1 |
| accuracy | 0.875 | 0.958 |

Table 4.6: Comparison of different technologies for fall detection

| | 2D sensor based approach [Zweng et al., 2010] | 3D sensor using image coordinates | 3D sensor using world coordinates |
|---|---|---|---|
| recall | | 0.775 | 0.925 |
| precision | 0.77 | 1 | 1 |
| f-score | | 0.873 | 0.961 |
| true negative rate | | 1 | 1 |
| accuracy | | 0.875 | 0.958 |

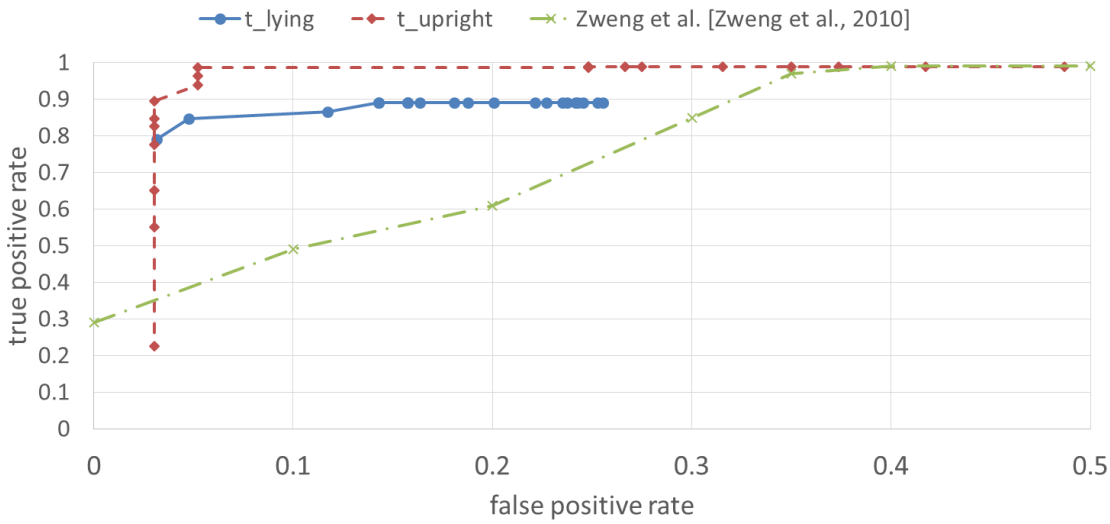



Figure 4.10: ROC curve of the proposed approach using fuzzy logic

flat was monitored, no additional devices (e.g. sensors, accelerometers) were used. There was no direct sunlight reported in this area, hence accurate depth data can be obtained. The couple is 72 resp. 66 years old and in a good health condition, i.e. no problems with mobility or balance were reported. Moreover, ADL are performed without additional help, hence both are able to live independently.

For the activity analysis using histograms, six of the monitored days were reported as "abnormal" by the couple, i.e. consist longer absence from home or dramatically changed daily routines. Hence, 97 days are considered as normal days where no alarm should be raised. Since this dataset is not artificially altered but acquired from a real scenario, it might be unbalanced with respect to the ratio of alarms and days without alarms.

### 4.3.1 Long-Term Inactivity Analysis

In order to evaluate results, alert lines based on 25 days intervals are generated, resulting in four long-term alert lines. For evaluation the leave-one-out cross-validation method is used, hence three alert lines are used for training whereas the fourth alert line is tested. The algorithm detected one significant decrease of mobility within the test period, depicted in Figure 4.12, where 64.7% of intervals are outside the specified range. During this interval, the couple was on a journey, thus resulting in increased inactivity. However, all three other alert lines are within the specified range, since only 21.1%, 3.7% and 1% of the time intervals are outside the range and thus considered to be outliers or only minor respectively temporary changes in mobility.

Figure 4.11: Part of the living room being monitored

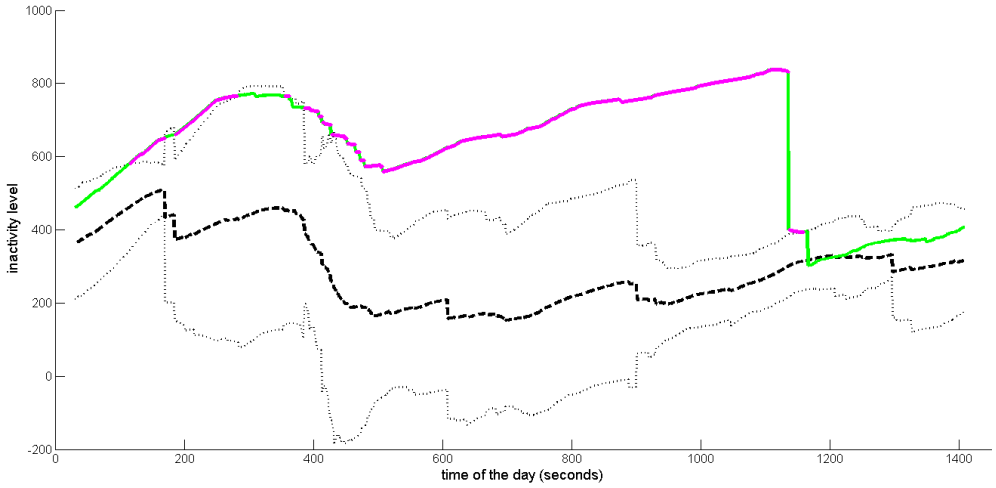



Figure 4.12: Deviation of alert line indicating higher inactivity and hence decreased mobility

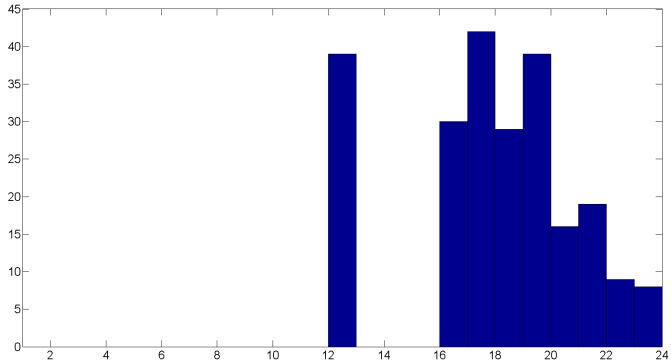



Figure 4.13: Example of a normal day 1 - activity in the morning is missing, but this day is considered as normal activity

### 4.3.2 Activity Histograms

The recorded dataset is challenging, since it represents the daily activities of real persons, not considering the change of daily activities during the week or on the weekend. Only the six days reported by the elderly where marked as alarms and thus being absent for half a day (Figure 4.13) is not reported as abnormal, since this is not abnormal for the couple. However, a typical histogram of activities is depicted in Figure 4.14: getting up in the morning between 6 and 7 AM followed by a peak of activities due to preparing and eating breakfast. Moreover, around noon, activity is increased due to typical activities performed during the morning and early afternoon (e.g. eating, playing cards, reading the newspaper). In the afternoon, no activity is detected due to watching TV in another part of the living room followed by activity due to preparing and eating dinner. Figure 4.15 depicts a similar histogram of activity, although the shape is different compared to Figure 4.14 due to a changed intensity of performing activities. Figure 4.16 illustrates abnormal behavior due to enhanced activity in the morning but decreased activity during the day (the amount of activity is significantly lower than normal). An abnormal shape of activity is depicted in Figure 4.17 and thus results in being categorized as abnormal activity. Absence for almost the whole day is also reported as abnormal since usually at least one person of the elderly couple is at home during the day (e.g. around noon), depicted in Figure 4.18.

Evaluation results are obtained by varying the number $n$ of randomly chosen training days from two up to 97 training days. The six days reported as abnormal behavior were not included in the training set, but in the test set. Thus, the size of the test set is 101 to six test samples, depending on the training set.

Since inactivity detection often results in false alarms, the goal is to reduce the number of FP while preserving TP. The proposed approach is evaluated using the toolbox provided by Piotr Dollár [Dollár, 2012] and compared to the approach using an alert line introduced by Cuddihy et al. [Cuddihy et al., 2007]. Evaluation results, depicted in Figure 4.19, visualize the number of alarms depending on the number of training

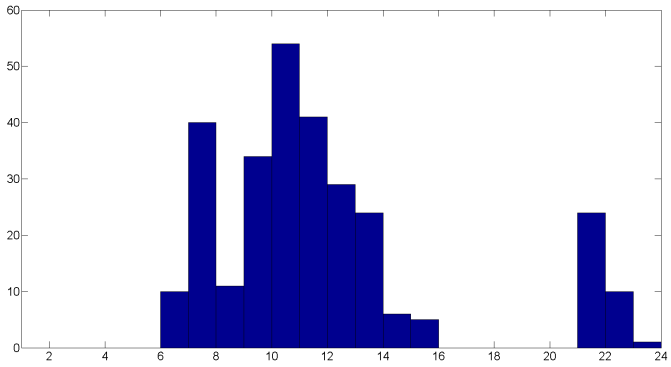



Figure 4.14: Example of a normal day 2 - activity is present throughout the day, except the afternoon

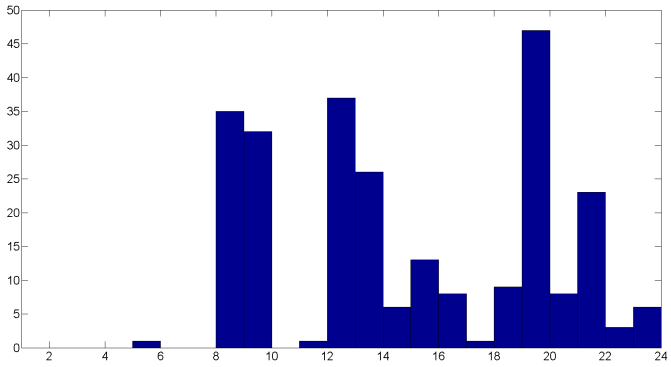



Figure 4.15: Example of a normal day 3 - activity and inactivity are present throughout the day

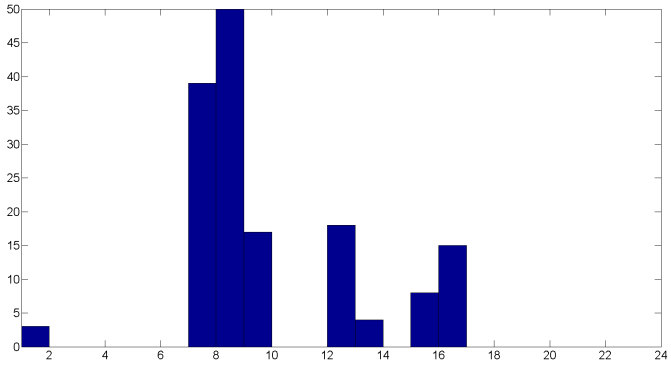



Figure 4.16: Example of abnormal activity 1 - activity is reduced in the afternoon/evening

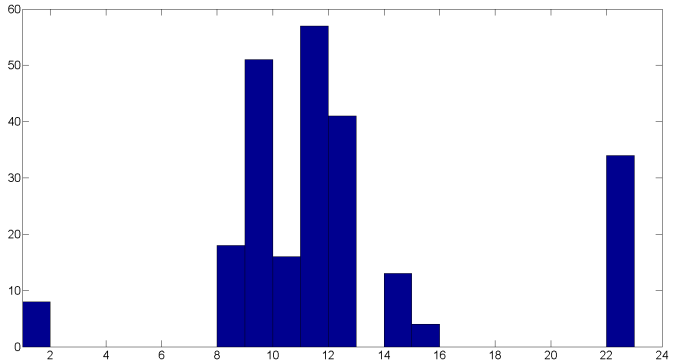



Figure 4.17: Example of abnormal activity 2 - no activity in the afternoon/evening

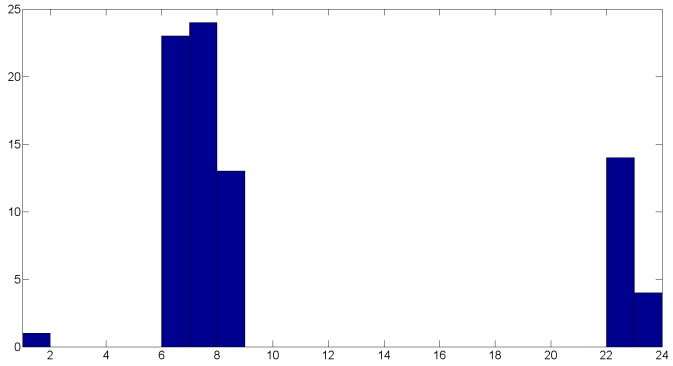



Figure 4.18: Example of abnormal activity 3 - absence during the day

days. As can be clearly seen, the number of false alarms using the alert line approach is high, especially with only few training data (over 500 alarms when using 2 days of training data). In comparison, using the proposed approach, the number of false alarms when using few training data is reduced to less than 100 false alarms. Please note that 100 resp. 500 false alarms on a test set including 101 test days results in one resp. five false alarms per day in average. Figure 4.20 shows a detailed view of Figure 4.19, where the maximum number of alarms is cut off at 100 in order to enhance the comparability between the approaches.

In order to improve the accuracy of the system, more training data is needed. However, even when increasing the training set to the size of 45 days, which is proposed by Cuddihy et al. [Cuddihy et al., 2007], the proposed approach still reduces the number of alarms from 35 when using the alert line approach to less than 16 alarms using the proposed approach. Since six alarms are included in the test set, the number of FP is even lower.

In order to evaluate the accuracy of the system, the f-score [Van Rijsbergen, 1979] is calculated and plotted depending on the size of the training set. Figure 4.21 depicts the f-scores of the introduced approach using different distance measures and the f-score

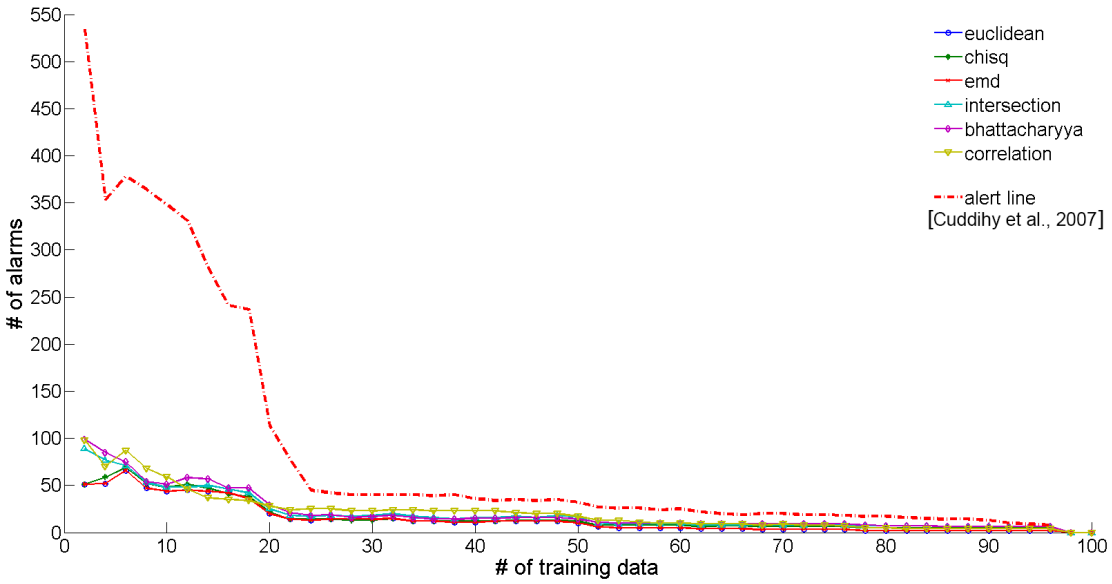



Figure 4.19: Alarm rate depending on the size of the training sample

of the alert line approach. All distances except the earth mover's distance and the correlation clearly outperform the alert line method [Cuddihy et al., 2007], not only in terms of less alarms but also in terms of better f-score values. The histogram correlation performs similar to the alert line approach, whereas the earth mover's distance results in a lower f-score than the alert line approach.

Table 4.7 depicts the number of alarms depending on the size of the training data. All histogram comparisons perform better in comparison to the alert line approach in terms of less FP. However, the number of alarms do not indicate whether the alarm is a TP or FP and thus the f-score is calculated and used for comparison of these approaches, e.g. the number of alarms using the Euclidean distance and the earth mover's distance result in the same number of alarms, but in different f-scores due to consideration of TP and FP when calculating the f-score.

Table 4.8 illustrates the accuracy of the proposed approach using different distance measures and compares the results to the alert line method introduced in [Cuddihy et al., 2007]. The highest f-score values are marked bold and thus can be seen that the chi-square distance performs best and increases the accuracy compared to the alert line approach.

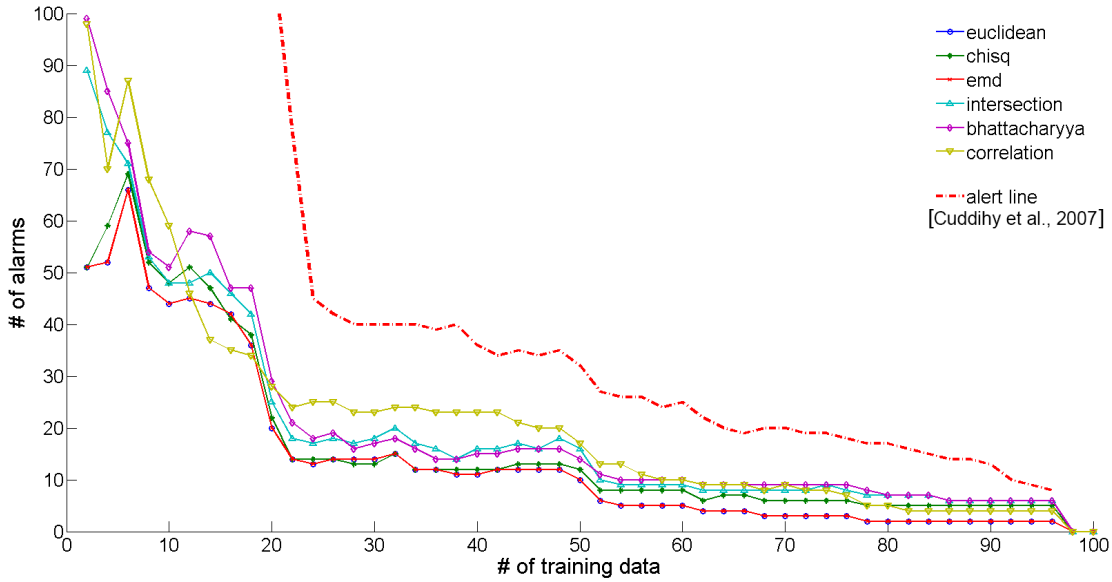



Figure 4.20: Detailed view of alarm rate (number of alarms $\leq 100$) depending on the size of the training sample

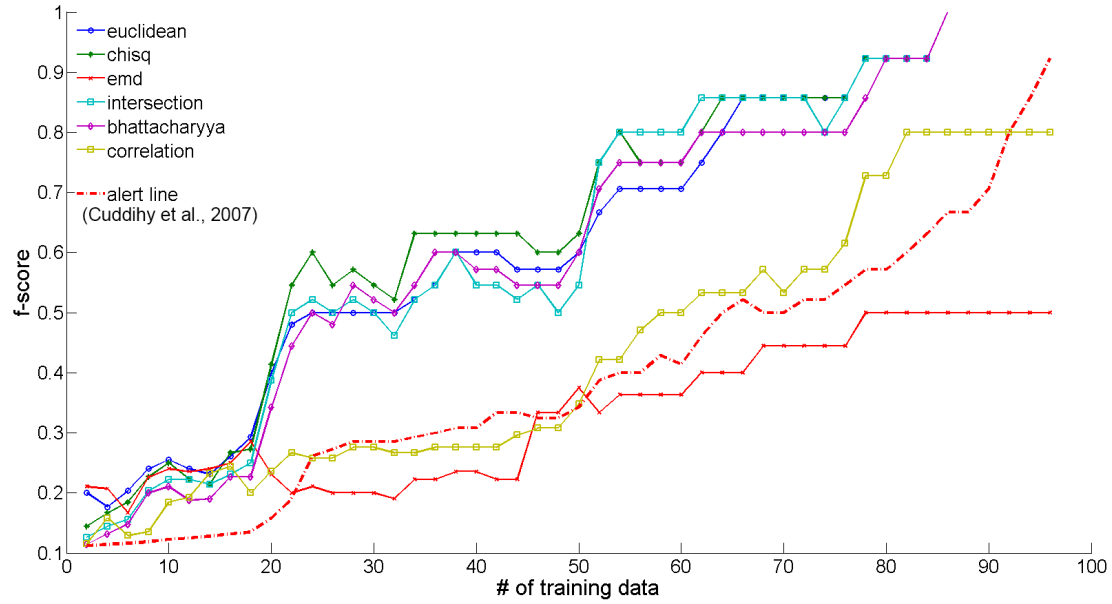



Figure 4.21: f-score depending on the size of the training sample

Table 4.7: Number of alarms

| training size | euclidean | chisq | emd | intersection | bhattacharyya | correlation | alert line [Cuddihy et al., 2007] |
|---|---|---|---|---|---|---|---|
| 10 | 44 | 48 | 44 | 48 | 51 | 59 | 348 |
| 20 | 20 | 22 | 20 | 25 | 29 | 28 | 113 |
| 30 | 14 | 13 | 14 | 18 | 17 | 23 | 40 |
| 40 | 11 | 12 | 11 | 16 | 15 | 23 | 36 |
| 50 | 10 | 12 | 10 | 16 | 14 | 17 | 32 |
| 60 | 5 | 8 | 5 | 9 | 10 | 10 | 25 |
| 70 | 3 | 6 | 3 | 8 | 9 | 9 | 20 |
| 80 | 2 | 5 | 2 | 7 | 7 | 5 | 17 |
| 90 | 2 | 5 | 2 | 6 | 6 | 4 | 13 |

Table 4.8: F-score

| training size | euclidean | chisq | emd | intersection | bhattacharyya | correlation | alert line [Cuddihy et al., 2007] |
|---|---|---|---|---|---|---|---|
| 10 | **0.255** | 0.250 | 0.240 | 0.222 | 0.211 | 0.185 | 0.122 |
| 20 | 0.400 | **0.414** | 0.231 | 0.387 | 0.343 | 0.235 | 0.158 |
| 30 | 0.500 | **0.545** | 0.200 | 0.500 | 0.522 | 0.276 | 0.286 |
| 40 | 0.600 | **0.632** | 0.235 | 0.545 | 0.571 | 0.276 | 0.308 |
| 50 | 0.600 | **0.632** | 0.375 | 0.545 | 0.600 | 0.348 | 0.343 |
| 60 | 0.706 | 0.750 | 0.364 | **0.800** | 0.750 | 0.500 | 0.414 |
| 70 | **0.857** | **0.857** | 0.444 | **0.857** | 0.800 | 0.533 | 0.500 |
| 80 | **0.923** | **0.923** | 0.500 | **0.923** | 0.923 | 0.727 | 0.571 |
| 90 | **1.000** | **1.000** | 0.500 | **1.000** | **1.000** | 0.800 | 0.706 |

## 4.4 Spatial Modeling

The proposed approach is evaluated on three different datasets (monitoring of a kitchen, living room and office environment), being recorded from a bird's eye view and are publicly available[3] in order to allow comparisons in the future. All scenes contain a sitting area as well as a walking area to be detected by the proposed approach. The evaluation of the proposed approach is performed on three different datasets[4]:

- **Living room dataset:** The dataset was recorded within a living room and consists of 90 days of tracking data. The scene is depicted in Figure 4.22 (left) and shows a living room with a sofa and free area, not only used to walk, but also during housecleaning.

- **Kitchen dataset:** The dataset was recorded within a kitchen and consists of 74 days of tracking data. The kitchen scene (Figure 4.22, middle) depicts a kitchen including cupboards, a kitchen top and a cook-top. Moreover, benches and a table for meal intake are within the scene and thus the benches are regularly used.

[3] http://tracking-dataset.planinc.eu, last accessed: March 31st, 2015

[4] Please note that filtering mechanisms are evaluated on a subset of this dataset, on 34 days of data within the living room and 18 days of tracking data within the kitchen. Humans were present for around 3-4 hours per day during the week and approximately 6 hours per day during the weekend.

- **Office dataset:** The dataset was recorded within an office and consists of 20 days of tracking data. The office scene is depicted in Figure 4.22 (right), showing that the proposed approach does not need to be applied within the context of AAL, but is a general approach to be used in different environments.

In all scenes, a sitting area as well as a walking area is present and to be detected by the proposed approach. Although the focus of this thesis is the monitoring of elderly people in home environments, the capability to extend the proposed approaches to new environments as well (e.g. office) is demonstrated. The advantages of monitoring the elderly at home are obvious (detect health deterioration, increased or decreased mobility), but are more subtle in the office environment. However, also in the office environment, the health status can be detected automatically: longer working hours than normal and less breaks are examples to indicate stress. Hence, the system can be adapted in order to detect stress at the workplace.

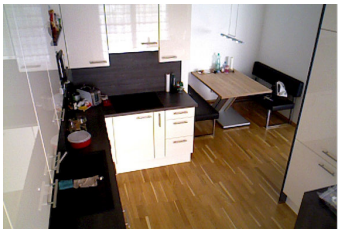
Figure 4.22 depicts RGB images and corresponding depth images of the scenes in combination with the annotated ground truth - walking areas are marked with a blue solid line, sitting areas are indicated by red dashed lines. Please note that within the office environment the definition of the sitting area is complex, since no fix constraints as in both other datasets are present, but a moveable chair is used in front of the desk. Hence, the sitting area is not exactly annotated since it is constantly changing. These specific field of views have been chosen in order to demonstrate the capability of the proposed approach - however, arbitrarily sensor positions are possible, as long as at least a free walking area is within the field of view.

In order to verify the accuracy, walking and sitting areas are annotated manually and the number of TP, FP, TN and FN are calculated to obtain the f-score. Figure 4.23 depicts the kitchen area as example for the evaluation process: the sitting area is annotated in red, the walking area in blue. Clustered tracking data is visualized with yellow circles representing the tracking information belonging to the sitting cluster, whereas cyan asterisks visualize tracking data of the walking cluster. Please note that the scenes are chosen since they are dynamic, i.e. objects are moved temporarily. Since the proposed approach is only human-centered, these movements does not influence the result of the scene analysis.

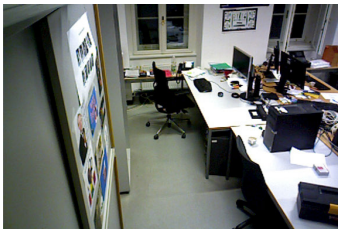
### 4.4.1 Skeleton Joint Analysis

In order to evaluate the influence of the proposed filtering mechanisms, the classification performance based on different skeleton joints is evaluated in a first step in order to choose the best skeleton joint for further processing. Figure 4.24 depicts the performance of different skeleton joints described by the f-score within sitting and walking areas of the kitchen and the living room dataset. The f-score is calculated based on the ground truth annotation and the height is clustered based on different skeleton joints (CoM, head, shoulder center, spine, hip left, hip right, knee left and knee right). For each skeleton joint, two clusters are obtained - the walking and the sitting cluster. However, since the cluster centers are detected automatically and skeleton tracking data is not

(a) Image of the living room

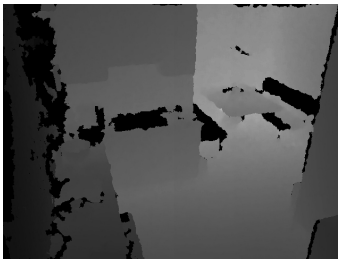


(b) Image of the kitchen



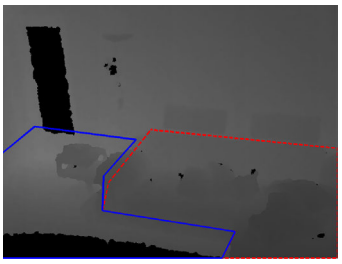
(c) Image of the office

(d) Depth image of the living room

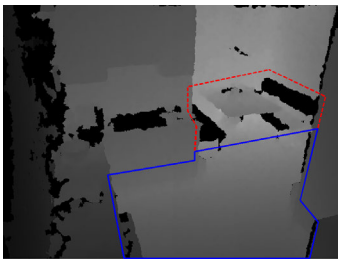


(e) Depth image of the kitchen

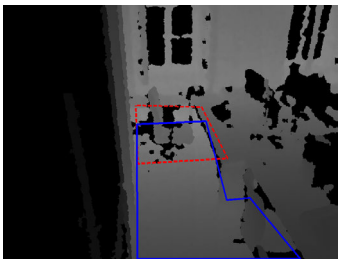
(f) Depth image of the office environment



(g) Ground truth annotation (living room)



(h) Ground truth annotation (kitchen)



(i) Ground truth annotation (office)

Figure 4.22: Evaluation dataset: RGB image, depth image and corresponding ground truth annotation (red dashed line indicates sitting area, walking area is marked with blue solid line)

robust, the same tracking information can result in different clusters, depending on the chosen skeleton joint. Most skeleton joints perform equally, except the knee joints where a significant reduction of the f-score is detected in the living room dataset. This is due to unstable tracking, since the upper body is tracked more robustly then body limbs. Moreover, the f-score within the kitchen area shows poor results due to massive tracking errors, where the skeleton is fit to furniture resulting in incorrect tracking information. This problem is discussed during the evaluation of the proposed approach in more detail.

### 4.4.2 Filtering

The evaluation of the proposed filtering mechanisms is performed on the kitchen and the living room dataset separately, since different parameter settings have different

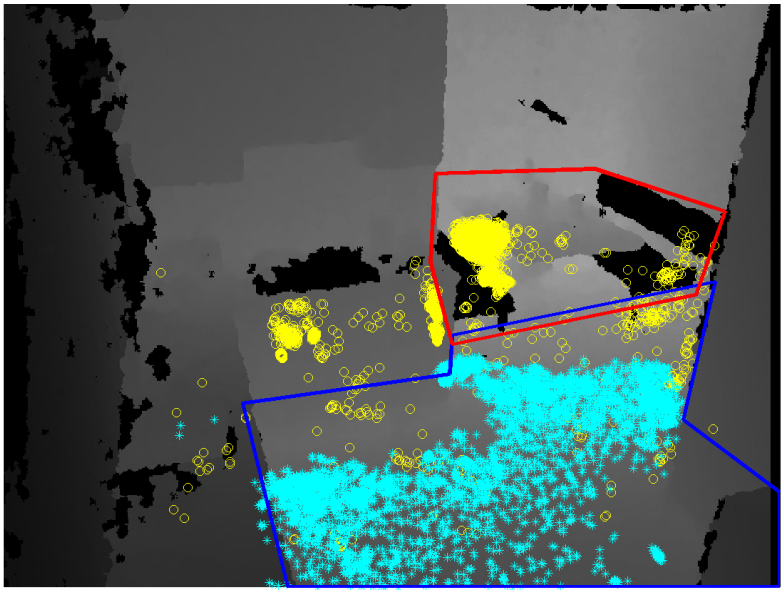


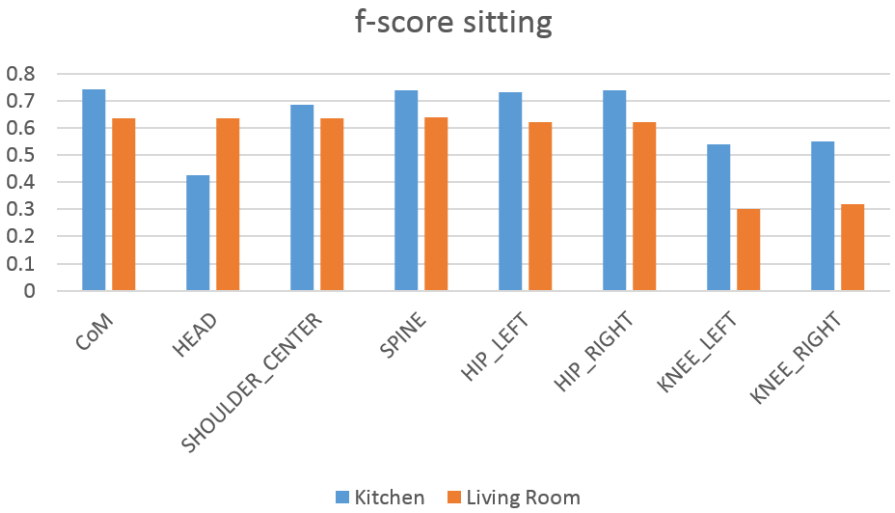
Figure 4.23: Kitchen scene: ground truth annotation of sitting (red) and walking area (blue) and center of mass tracking points clustered to walking (cyan asterisk) and sitting (yellow circle)

implications, depending on the scene. However, this does not mean that parameters are scene dependent, but a combination of parameters to fit general scenes is obtained. Evaluation if performed similar to the evaluation of the skeleton joints: according to the ground truth annotation and the clustering results with (filtered) data, the corresponding f-score is calculated for different thresholds applied during the filtering process.

Figure 4.25 illustrates the influence of the body orientation, the use of a height threshold as well as the use of confident data (marked by OpenNI). The value of body orientation is calculated according to the proposed fall detection approach, where the similarity of the orientation to the ground plane is calculated based on the angle between ground plane and upper body. The range of the body orientation index is from 0 (parallel to the ground floor) to 1 (orthogonal to the ground floor).

As can be seen clearly, only using upright positions (body orientation) does not influence the result of the kitchen scene, since the f-score is constant over time (Figure 4.25a and 4.25b). However, the use of a height threshold significantly increases the performance of the system from an f-score of 0.1 to an f-score of 0.98 within the walking area of the kitchen (Figure 4.25a). This increase of performance is gained due to eliminating incorrect tracks as depicted in Figure 3.17. These incorrect tracks occurred due to the movement of the cupboard doors, which are incorrectly tracked by OpenNI, resulting in tracks on the kitchen top with a high distance to the ground floor. Hence, these incorrect tracking results can be filtered out using the proposed approach, since the height of the CoM is much higher than 2 meters. Within the sitting area of the kitchen


(a) Comparison of the performance of different skeleton joints in the walking area




(b) Comparison of the performance of different skeleton joints in the sitting area

Figure 4.24: Comparison of f-scores for different skeleton joints

Table 4.9: F-score indicating the influence of different parameter combinations

| | | | kitchen | | living room | |
|---|---|---|---|---|---|---|
| body orientation | height threshold | confidence | **F1 walk** | **F1 sit** | **F1 walk** | **F1 sit** |
| no | no | no | 0.10 | 0.74 | 0.80 | 0.63 |
| yes | no | no | 0.10 | 0.75 | 0.83 - 0.99 | 0.63 - 0.75 |
| no | yes | no | 0.96 | 0.9 | 0.74 - 0.79 | 0.64 |
| yes | yes | no | 0.98 | 0.88 - 0.9 | 0.83 - 1 | 0.63 - 0.82 |
| yes | yes | yes | 0.98 - 0.99 | 0.92 | 0.9 - 1 | 0.77 - 0.95 |

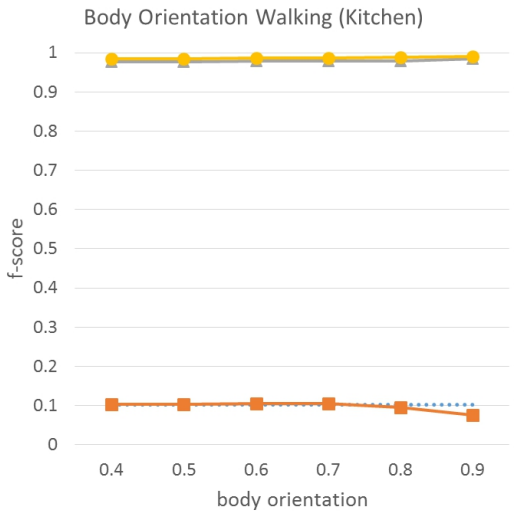
a performance improvement from an f-score of 0.75 to an f-score of around 0.9 can be gained by introducing a height threshold, illustrated in Figure 4.25b.

In the living room dataset, an improvement when applying the proposed body orientation threshold is showed in Figure 4.25c and 4.25d. In both, the walking and the sitting area, focusing on upright positions significantly improves the results. In combination with the proposed height thresholds, high f-scores are achieved.

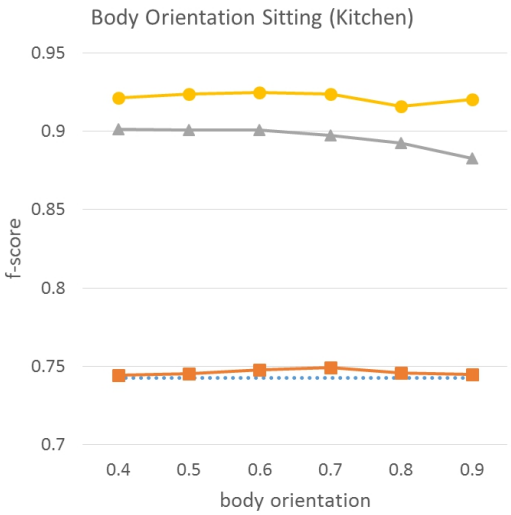
In summary, a body orientation of 0.8 should be chosen to obtain the best results. The use of a body orientation threshold does not change the results in the kitchen dataset, but improves the results of the living room dataset. However, restricting the body orientation too much results in a reduction of the f-score, depicted in Figure 4.25d.

The influence of different height thresholds is depicted in Figure 4.26: all datasets indicate that choosing the value is not critical since results are constant and thus not depending on the chosen height threshold. However, results decreases when the height threshold is too relaxed, i.e. higher than 1.9 meters. As already shown before, filtering the body orientation results in higher f-score, especially in the living room dataset - illustrated in Figure 4.26c and 4.26d.

Table 4.9 summarizes the results and the influence of the evaluated parameters. A range of f-score in this table indicates that the results are within the specified range and depend on the defined threshold. As can be seen, the use of thresholds for eliminating outliers significantly increases the accuracy of the detection of walking and sitting areas. Due to the use of fast filtering mechanisms, the performance of the classification is improved, without increasing the computational demands since only thresholds are applied. Results show that all three proposed filtering mechanisms are feasible to improve the results of the classification step. The choice of the height threshold is not critical, since results are mainly independent from the chosen height threshold. However, the choice of the body orientation threshold is more critical, but focusing only on upright positions increases the results. Moreover, the use of only confident joint values (marked by OpenNI) is also recommended, since this combination leads to the best results.

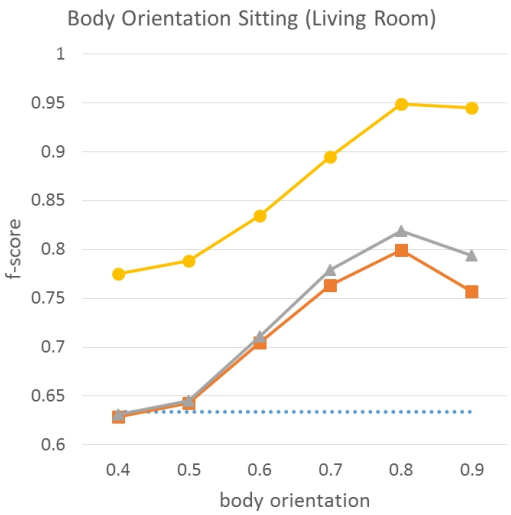



(a) Influence of the body orientation on the f-score in walking areas of the kitchen




(b) Influence of the body orientation on the f-score in sitting areas of the kitchen

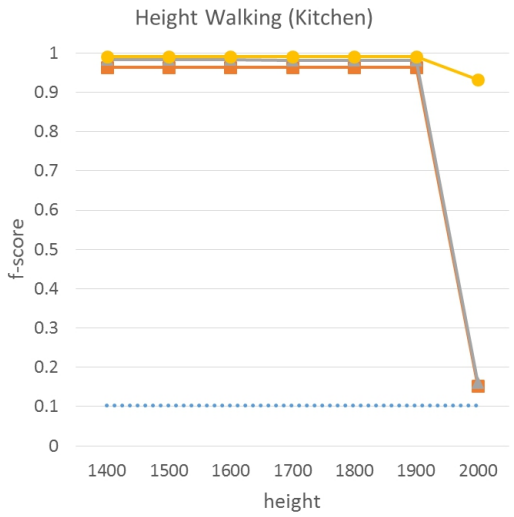

(c) Influence of the body orientation on the f-score in walking areas of the living room



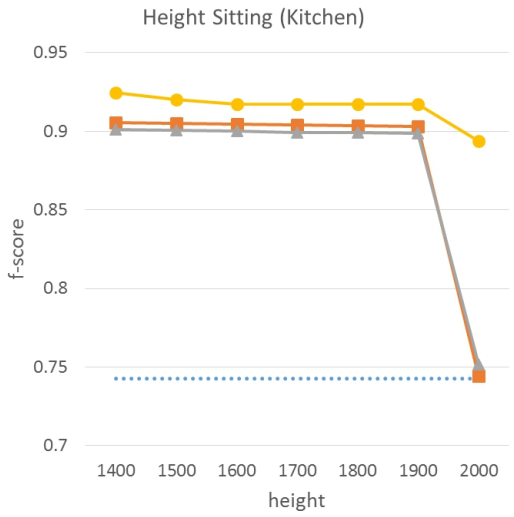

(d) Influence of the body orientation on the f-score in sitting areas of the living room


(e) Legend

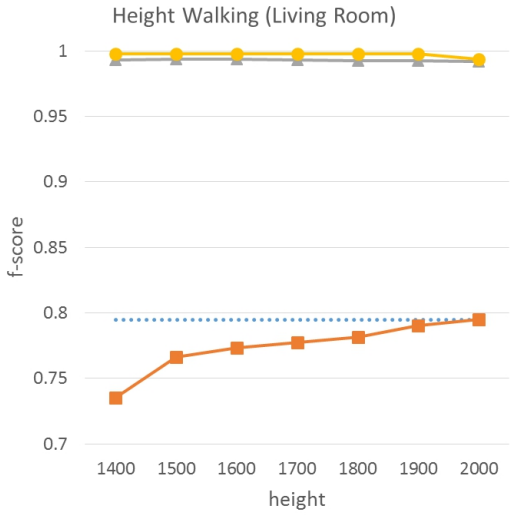
Figure 4.25: Influence of the body orientation threshold on the filtering results



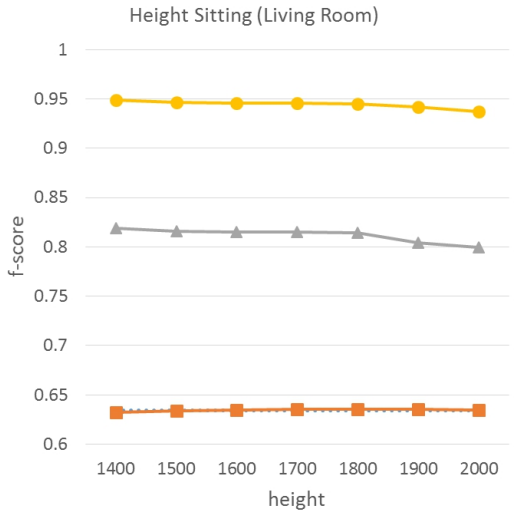

(a) Influence of the body height on the f-score in walking areas of the kitchen




(b) Influence of the body height on the f-score in sitting areas of the kitchen




(c) Influence of the body height on the f-score in walking areas of the living room




(d) Influence of the body height on the f-score in sitting areas of the living room


(e) Legend

Figure 4.26: Influence of the body height threshold on filtering results

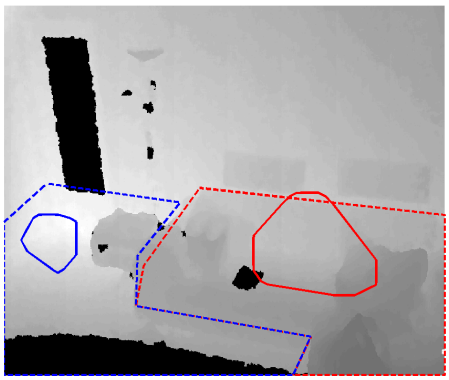
### 4.4.3 Modeling of Walking and Sitting Areas

Repeated sub-sampling validation is performed by randomly choosing the training set 6 times and averaging the results of all 6 runs. An example of the learning process of walking and sitting areas in the living room dataset is shown in Figure 4.27: starting with only 2 days of training data, the results improve until 89 days of training data is provided. This example shows that the proposed approach is highly depending on the quality of the training data, if only a small training set is available. However, since this analysis is proposed to be performed on the long-term, the approach eliminates outliers when being applied to large datasets and thus provides reasonable results.

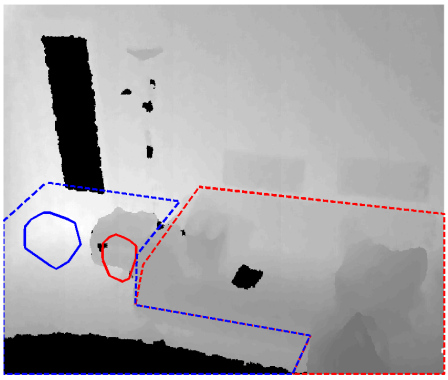
Figure 4.28 depicts the dependency of the f-score on the size of the training set (number of training days), where the sitting area of the kitchen is shown in the left picture and the walking area of the kitchen is shown in the right picture. The red dotted curve depicts the result of the unfiltered data, whereas the blue curve depicts the result if applying the filtering step. The walking area is modeled with a resulting f-score of 0.87 for the filtered data and 0.84 for unfiltered data. However, the f-score for the sitting area achieved only 0.44 (unfiltered) respectively 0.33 (filtered).

In order to perform further analysis on these quantitative results, qualitative evaluation of the results is shown in Figure 4.29b: during the 6 runs, two different models are obtained - the correct one is obtained four times (depicted on the left side) and an incorrect one is obtained two times (depicted on the right side). The incorrect model is generated due to incorrect tracking data and thus, the sitting area is modeled much bigger than it actually is. This strong influence of the incorrect tracking data is a result of too less training data, hence when using more training data, the influence of each training day is minimized. However, the low f-score is also a result of the ground truth labeling - as can be seen in the left image of Figure 4.29b, the detected sitting area considers only one bench while the other bench is ignored since the second bench is already outside the tracking range. Hence, a person sitting on the second bench is not tracked and thus, no tracking data within this area is available. Moreover, the table is considered as sitting area in the ground truth and thus an f-score of 1 cannot be achieved. Hence, ground truth annotation need to be adapted according to the range limits of the tracking algorithm and the table need to be excluded from the ground truth. However, qualitative analysis of the walking area in the kitchen dataset shows that the walking area is modeled very well in both cases and thus result in an f-score of greater than 0.8. Moreover it is shown, especially in the walking area of the kitchen, that the filtering step is able to enhance the robustness of the proposed approach, since a more stable and higher f-score is achieved.

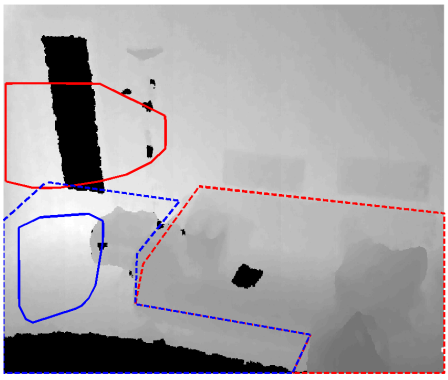
The quantitative results of the living room and office dataset are depicted in Figure 4.30 and Figure 4.31: in contrast to the kitchen dataset, higher f-scores are achieved. The f-score for the sitting and walking area in the living room is 0.75 respectively 0.73, both for the filtered dataset. On the office dataset, f-scores of 0.68 for the sitting area and 0.88 for the walking area are achieved by using the maximum number of training data. The office dataset only consists 20 days of monitoring, hence the influence of increasing the training set can be seen since the f-score raises with adding additional training data.



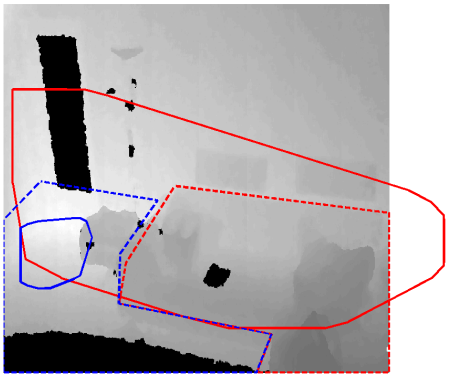
(a) Detected walking and sitting areas based on 2 training days



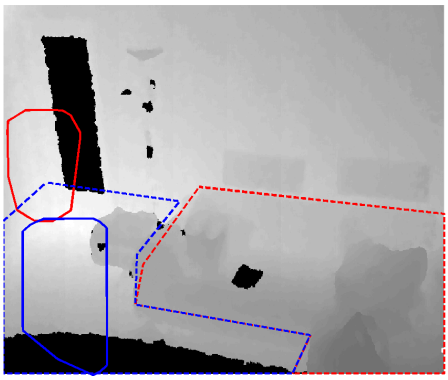
(b) Detected walking and sitting areas based on 4 training days



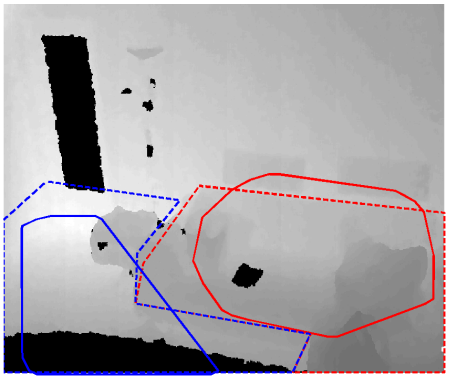
(c) Detected walking and sitting areas based on 9 training days



(d) Detected walking and sitting areas based on 18 training days



(e) Detected walking and sitting areas based on 34 training days



(f) Detected walking and sitting areas based on 89 training days

Figure 4.27: Example of the learning process, showing the results of detected walking (blue) and sitting areas (red) in the living room dataset with different amounts of training data where the ground truth is marked with blue respectively red dashed lines

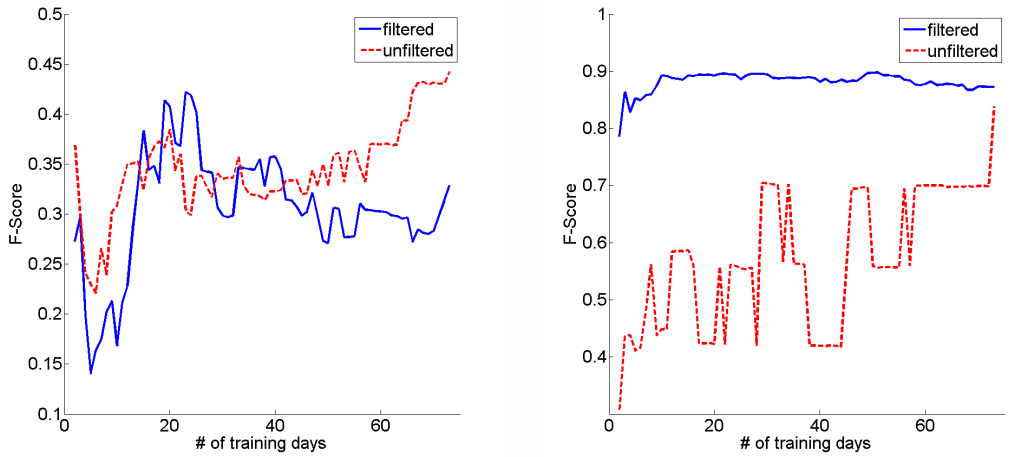
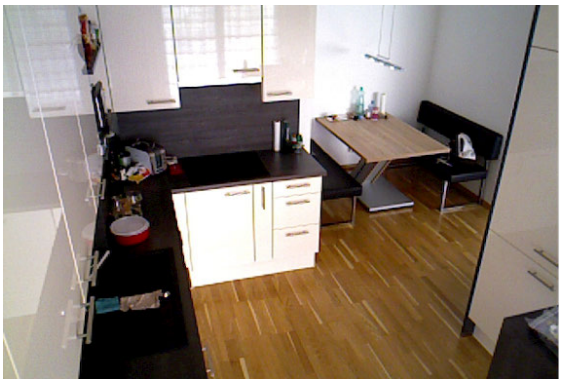
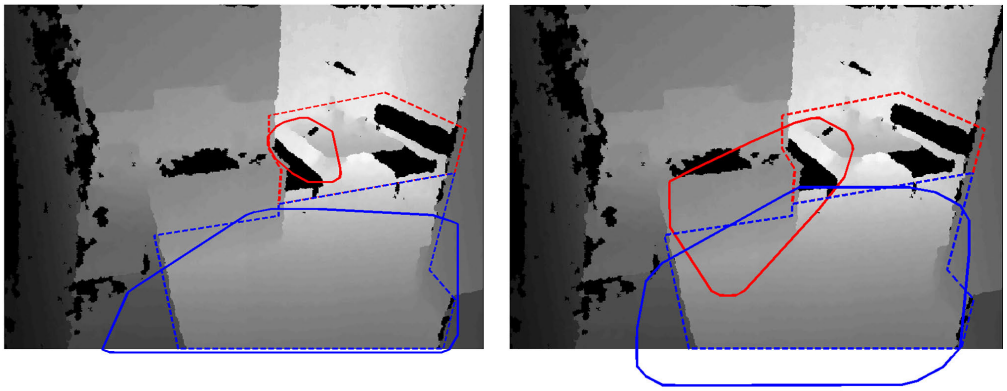



Figure 4.28: F-score of the sitting (left) and walking area (right) in the kitchen dataset depending on the number of training samples (average of 6 runs)



(a) RGB image of the kitchen



(b) Sitting (red) and walking areas (blue) in the kitchen dataset with respective ground truth marked with dotted lines

Figure 4.29: Results of the scene understanding approach on the kitchen dataset

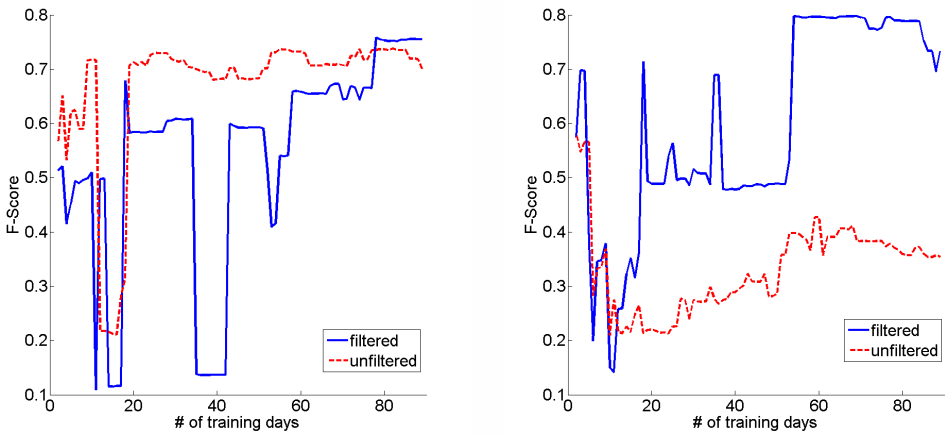



Figure 4.30: F-score of the sitting (left) and walking area (right) in the living room dataset depending on the number of training samples (average of 6 runs)

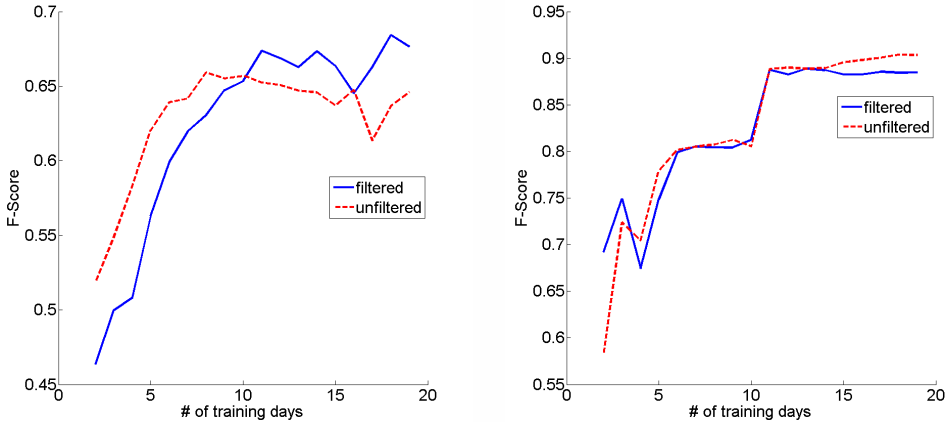
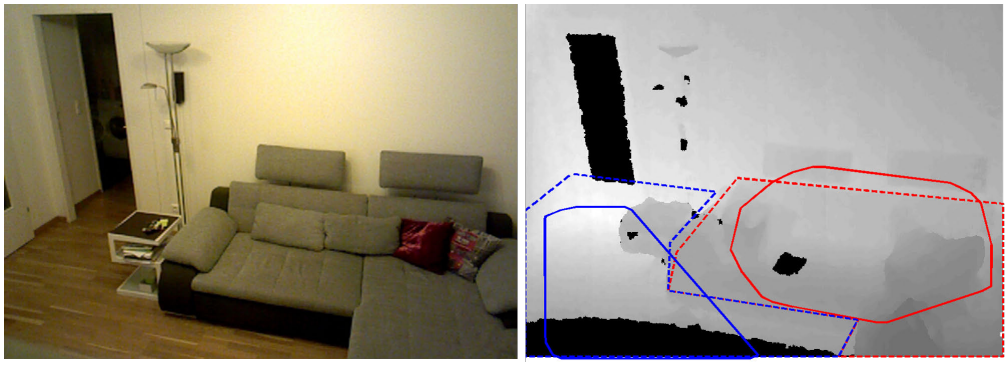



Figure 4.31: F-score of the sitting (left) and walking area (right) in the office dataset depending on the number of training samples (average of 6 runs)

Moreover, the office dataset is not as challenging as the kitchen and living room dataset since almost no tracking errors are present and thus, applying the filtering does not significantly change the results. This is due the fact that the most common tracking errors experienced with the living room and kitchen dataset are the fitting of the skeleton to doors and other objects. In the office scene, no doors are within the field of view and thus enhancing the robustness of the tracker, yielding in better results already with little training data.

Qualitative analysis of the results show that the regions are modeled accurately and confirm the quantitative results: Figure 4.32a and Figure 4.32b depict the scenes, detected sitting (red) and walking area (blue) of the living room (Figure 4.32a) and office dataset



(a) RGB and depth image of the living room data set, including sitting (red) and walking areas (blue) and respective ground truth marked with dotted lines

(b) RGB and depth image of the office data set, including sitting (red) and walking areas (blue) and respective ground truth marked with dotted lines

Figure 4.32: Results of the scene understanding approach on the living room and office dataset

(Figure 4.32b). The corresponding ground truth is shown with dotted lines. Similarly to the kitchen dataset, also in the office dataset the ground truth of the sitting area is larger since all possible positions to sit are considered within the ground truth annotation - however, this does not mean that all possible positions to sit are actually used by the person, since people tend to usually sit on the same spots and do not change these spots.

## 4.5 Advanced Spatio-Temporal Behavior Modeling

The evaluation is based on a subset of the dataset previously introduced: the monitoring of a kitchen (90 days) and a living room (74 days). Evaluation of the behavior modeling approach is not performed on the office data set since it only contains 20 days of monitoring and thus does not contain enough data for a long-term behavior analysis. Hence, it is only performed on the living room and kitchen dataset. Previous experiments

have shown that the system is sensitive and produces false alarms - thus the evaluation is based on the fact, that no changes of mobility are present and thus no alarms should be generated.

The aim of region based behavior modeling is to gain a more accurate local behavior model, representing different activities based on the actions sitting and walking. In order to evaluate the performance of the proposed approach, the number of false alarms is used as indicator since during the testing period no TP were obtained and thus a ROC curve cannot be constructed. Moreover, since behavior models are prone to false alarms, the performance of the system in order to reduce the false alarms is evaluated. For repeated sub-sampling, 20 different training samples are chosen randomly and the evaluation is performed 20 times, where results are averaged. The results of the region based behavior modeling are depicted in Figure 4.33 and Figure 4.34[5]: although the proposed approach is able to outperform the approach of Cuddihy et al. [Cuddihy et al., 2007] in terms of lower false alarms independently from the behavior model, the approach of using activity histograms is only outperformed when focusing on the sitting region. It can be seen that the walking area results in a higher number of false alarms than when using activity histograms on a global level, since the walking data is not as regularly organized as the sitting data is. This can be explained easily due to the fact, that people tend to sit at the same time every day, e.g. to eat lunch or dinner, or to watch TV in the evening. Hence, a more structured daily routine is found for the time spending sitting, whereas walking is more unstructured since people usually do not walk to e.g. clean the flat at the same time but in a more unstructured way. Hence the proposed approach is able to incorporate these aspects of the daily routines, whereas this information is lost when using a global approach.

## 4.6 Summary

**Fall Detection:** evaluation shows that the use of an Asus Xtion pro in order to obtain depth information is able to outperform 2D fall detection. Moreover, due to the combination of tracking information together with a fuzzy logic framework, a robust fall detection system is introduced. From a practical perspective, evaluation and discussion with end-user organizations has shown that it is not feasible to detect the fall event itself. Therefore situations where help is needed in general should be detected, rather than focusing on the fall event. These situations are detected by the information that a person is in an upright position, followed by lying on the floor and does not get up to an upright position within a specified amount of time. Especially for real-world applications it does not make any difference, why the person is not able to get up from the floor. Hence, it is not of interest whether a fall occurred or if the person intentionally lay down on the ground - if the person is not able to get up anymore, help is needed in any case. Furthermore, the use of the Asus Xtion pro offers practical advantages: it is robust to changing lighting conditions, also works also during the night and the installation in real homes is simplified by using only one sensor without the need for a complex calibration.

[5]Please note that the number of false alarms is cut off in order to focus on the details.

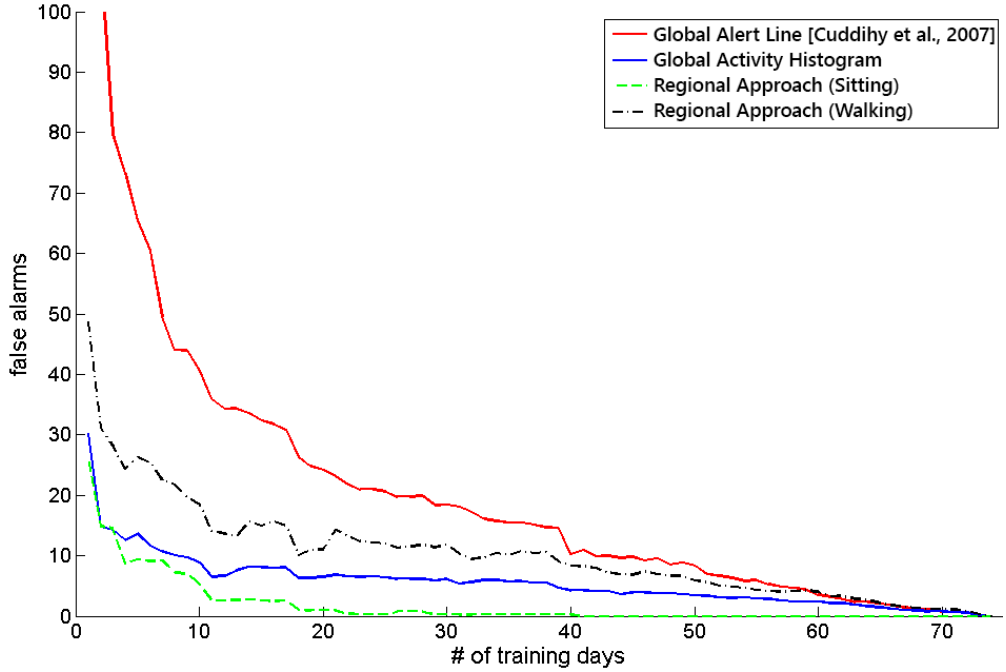



Figure 4.33: False alarm rate depending on the training (kitchen)

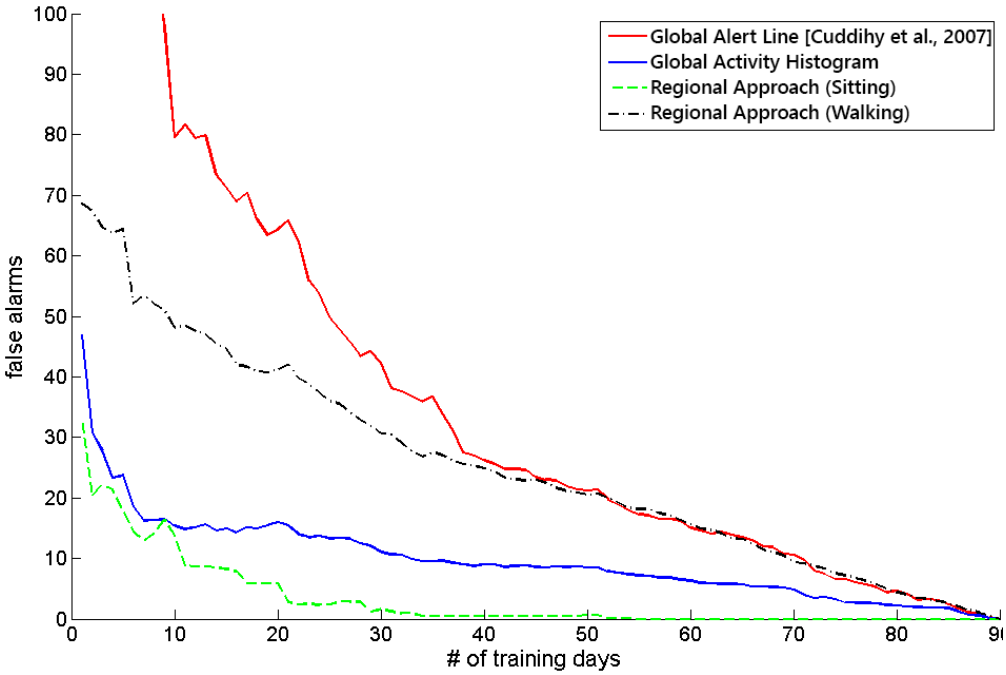



Figure 4.34: False alarm rate depending on the training (living room)

**Temporal Modeling:** the evaluation of the introduced long-term modeling in order to detect changes in behavior caused by mobility changes indicate that the use of this approach is feasible. Although only evaluated on a small dataset (4 long-term alert lines based on 100 days of data), one deviation was robustly detected and thus yields in feasible results. The introduction of activity histogram comparisons showed that the use of activity histograms allows to reduce the number of false alarms when using a vision-based approach, although the dataset is challenging. Different metrics for histogram comparisons were evaluated, results indicate that the chi-square metric is the most appropriate metric to be used within the proposed approach.

**Spatial Modeling:** filtering mechanisms in order to remove outlier from the dataset need to be applied as a pre-processing step, since otherwise no robust spatial modeling can be performed. The choice of the skeleton point to be analyzed was found to be not critical, as long as joints from the upper body and no limbs are chosen. The combination of body orientation, distance to the ground floor as well as the confidence value yields in promising results and indicates that this combination can be applied within other contexts as well. Quantitative results may suffer from the sub-optimal annotation of the ground truth, since sitting areas are annotated to broad, thus resulting in lower f-score values.

**Spatio-Temporal Behavior Modeling:** the combination of all proposed approaches showed that all approaches are able to reduce the number of false alarms in comparison to the state-of-the-art. However, region based modeling based on the obtained walking and sitting regions indicate that activity within walking areas is performed in a more unstructured way than in sitting areas. This results in a higher number of false alarms within walking areas, since temporal modeling within this area is more complex. Moreover, false alarms using global approaches are therefore mainly caused by walking areas and thus confirming the need for a spatio-sensitive behavior model.

**Sensors:** the use of a 3D sensor providing only depth information is able to respect the privacy of elderly people more appropriately than using cameras. Hence, the acceptance of 3D sensors is higher than cameras, since cameras are always connected with the attitude of surveillance. The development of 3D sensors like the Asus Xtion pro results in the low price of the sensor, since it is available on the mass market. Although the accuracy depends on the specific scenario, the use of 3D sensors within the context of AAL is feasible. However, drawbacks of current technologies are the limited field of view and limited range (10 m) of the sensors. The use of these sensors is also restricted to indoor environments, since interference of the infrared light with direct sunlight affects the accuracy.

**Computer Vision:** from a computer vision point of view, proposed algorithms outperformed state-of-the-art algorithms, only the advanced spatio-temporal behavior model indicates a better performance when using a different approach. However, since the proposed approach models sitting and walking areas separately, results show that the walking behavior of humans is less structured than sitting behavior and thus modeling is challenging. On the other hand, only considering results obtained within sitting areas clearly shows, that the proposed model is feasible to model behavior correctly. Not

only the overall behavior model, but also its modules provided interesting results to the community, especially the use a 3D sensor to detect falls was novel at this time. Moreover, advances within human-centered scene understanding were made in order to enhance the focus on humans and to propose alternatives to traditional scene understanding approaches.

**AAL:** results of the proposed approaches show, that the approaches are able to model behavior correctly, thus reducing the number of false alarms. Technology being applied in the context of AAL in practice need to be affordable, easy to install and reliable. Since off-the-shelf sensors are used, sensors can be afforded by elderly people. Moreover, since the introduced systems are self-calibrating, no external calibration is needed, ensuring the system being easy to install. Finally, the number of false alarms is reduced, allowing to provide help for the elderly without stressing the time of caretaker. However, modeling the behavior of humans is a complex task and thus it is impossible to model and predict the behavior of humans to 100%.

CHAPTER 5

# Conclusion and Future Work

Motivated by the demographic change, the emerging 3D sensor technologies and behavior monitoring approaches, a novel behavior model framework within the context of AAL was proposed. This thesis introduced a generic spatio-temporal behavior model, focusing on short-term (duration of minutes), mid-term (duration of days) and long-term (duration of months) aspects and thus allows to model the behavior in a holistic way. Moreover, privacy issues were discussed, since especially in the context of AAL, the dignity and privacy of elderly people need to be protected. Although the model was introduced within the context of AAL, it was shown that the proposed approaches can also be applied to other contexts as well, since they are based on long-term tracking data of humans.

Based on the state-of-the-art, a combination of temporal and spatial knowledge was introduced in order to enhance the information within the model. Hence, ROI based on tracking data obtained from the Asus Xtion pro were modeled and temporal modeling was performed using a regional alert line within each ROI individually. Results showed that this approach is able to model the behavior in more detail, since a global model aggregates spatial information. Hence, detailed temporal details are preserved when using this region based approach. Already the evaluation of the basic spatio-temporal behavior model indicates, that modeling ROI within a scene yields in feasible results describing the behavior of elderly people on a mid-term range, i.e. on a per days basis.

In order to incorporate short-term information into the spatio-temporal behavior model, a fall detection approach was introduced within this work. The introduced approach is based on the analysis of skeleton data obtained from depth images, calculating features based on the skeleton joints. The proposed features are the major body orientation as well as the spine distance to the ground floor. With these two features, the pose of the person as well as its distance to the ground floor is estimated. A fall is detected if the spine is close to the ground floor and the body orientation is parallel to the floor, indicating that a person is lying on the floor. However, due to the combination of these features, it can be differentiated whether the person is lying on the bed or on the floor. Moreover, sitting on the floor is also detected and thus does not result in false alarms.

The final decision whether a fall occurred or not is based on fuzzy rules, resulting in confidence values for each pose. Evaluation results showed that the approach is able to robustly detect falls. However, it was also shown that the results are highly depended on the quality of tracking data and thus false alarms are caused by incorrect tracking information. Hence, future work need to deal with the improvement of tracking data, in order to provide reliable and robust tracking, before fall detection is performed.

Incorporating long-term knowledge into the behavior model is achieved by the analysis of changes in mobility over the duration of months. The introduced approach compares the reference inactivity model with previous models in order to detect significant deviations. This is important, since slow changes in mobility cause most systems to adapt to new behavior, but not detecting these changes accordingly. From a practical point of view, detection of mobility changes is crucial, since enhanced mobility during the night is an indicator for dementia, whereas reduced mobility over time indicates a loss of strength. Evaluation on a small dataset yielded in promising results and showed that the proposed approach is able to detect trends in mobility over a long-term duration. Future work within this area need to perform a large-scale evaluation of this approach over the cause of several years, which was beyond the scope of this thesis.

To improve the robustness of temporal modeling based on computer vision, activity histograms were introduced and compared to inactivity profiles. Activity histograms model the activity within indoor scenes, instead of constructing an inactivity profile. Since histogram comparisons are widely used within the field of computer vision, different metrics for comparison are available. During the training phase, a reference histogram of 24 bins (one bin per hour) is learned and the standard deviation from the training data is modeled in order to allow flexibility depending on the context. During the test phase, the histogram modeling the behavior of the day is compared with the reference histogram and deviations of abnormal behavior are detected, if the histogram of the current day is significantly different from the pre-trained reference histogram. Results illustrated that the use of histogram comparisons is feasible, since the number of false alarms is reduced in comparison to the alert line approach using inactivity profiles. Hence, histograms are able to accurately model activities throughout the day.

Spatial modeling is improved by the introduction of a novel human-centered scene understanding approach based on continuous depth data, focusing on functionalities the scene is offering for humans. The proposed approach models sitting and walking areas within a scene, solely based on long-term tracking data of humans. Since noisy skeleton data obtained from OpenNI is used, filtering mechanisms were introduced in order to remove outlier. These filtering mechanisms are based on the pose and the distance to the ground floor, since during sitting and walking an upright pose is assumed. Thus allows to eliminate incorrect tracking data and enhances the quality of the tracking data significantly. The scene model is obtained by clustering the CoM of long-term tracking data, resulting in walking and sitting cluster. By performing a KDE, sitting and walking areas are accurately modeled. Quantitative results indicate that the ground truth annotation was sub-optimal, since sitting areas were annotated larger than they actually are in order to consider all sitting possibilities. However, qualitative analysis

showed that sitting and walking areas were modeled accurately and thus indicate areas, where people are actually sitting and walking and not areas, where people can walk or sit (as annotated in the ground truth data).

The combination of the improved temporal and spatial modeling together with the incorporation of short-term as well as long-term knowledge results in the proposed behavior model. This model obtains the information about spatial scenes by combining the proposed human-centered scene understanding approach with the activity histogram comparisons in order to provide a spatio-temporal behavior model. Evaluation showed astonishingly results, since the spatio-temporal behavior model is outperforming the inactivity based temporal model, but only the sitting area is outperforming the global activity based temporal model. The temporal model within walking areas perform worse than the global model, indicating that activities within sitting areas are more structured on a time base than within walking areas and thus, obtaining an accurate behavior model within walking areas is more challenging than within sitting areas.

Limitations of the sensor (e.g. Asus Xtion pro) only allow to model the behavior within indoor environments, since due to the functionality, depth data can only be obtained indoors. Moreover, the range of the sensor is limited to 10 meters, causing problems within larger rooms. However, due to a combination of multiple sensors, this challenge can be solved. Moreover, the proposed model is based on tracking data and thus highly relies on the quality of this data. If tracking data is inaccurate, also the behavior model is inaccurate. These limitations and challenges can be circumvented by the use of a more sophisticated sensor, allowing a higher range as well as a better performance in outdoor areas. However, privacy issues need to be considered since no RGB but only anonymized depth information should be used. In combination with improved tracking algorithms, the proposed spatio-temporal behavior model is not only able to model behavior within the context of ADL, but also can be used in different scenarios.

Within the context of AAL, the proposed approach allows elderly people to stay in their own homes longer, since it provides a holistic system, focusing on the needs of the elderly. Using the proposed behavior model allows to detect falls and provide help immediately, reducing the number of deaths and days needed for rehabilitation. Moreover, due to the detection of long-term mobility changes, early stages of dementia or a general reduction of mobility are detected and help can be provided. Overall, a holistic health status of the person can be obtained by analyzing the behavior of the elderly and thus, appropriate support can be provided.

# List of Figures

# List of Tables

# Acronyms

**AAL** Ambient Assisted Living. 1, 3, 6, 8, 9, 12–17, 19–21, 34, 40, 41, 48, 49, 51, 52, 63, 77, 89, 102, 116, 117, 119, 121, 123

**ADL** Activities of Daily Living. 2, 4, 12–15, 25, 27, 30, 41–44, 49, 57, 60, 79, 87–89, 91, 94, 121

**CAR** Circadian Activity Rythmus. 44

**CoM** Center of Mass. 7, 69–71, 73, 76, 102, 104, 120

**FN** False Negative. 89, 91, 102

**FP** False Positive. 89, 91, 92, 96, 98, 99, 102

**fps** frames per second. 19, 66

**GPS** Global Positioning System. 15

**HMM** Hidden Markov Model. 42, 43

**HOG** Histogram of oriented gradients. 32

**ICT** Information and Communication Technology. 12

**KDE** Kernel Density Estimation. 7, 69–71, 73, 74, 76, 120, 124

**KLT** Kanade-Lucas-Tomasi. 34

**MRF** Markov Random Field. 38

**POMDP** Partially Observable Markov Decision Process. 42

**RFID** Radio-Frequency Identification. 14, 15

**ROC** Receiver Operating Characteristic. 92, 94, 114, 125

**ROI** Region of Interest. 5–7, 49, 53–56, 69, 71, 82–84, 119, 124

**SDK** Software Development Kit. 20, 53, 57, 69, 75

**SIFT** Scale-invariant feature transform. 31, 35

**SURF** Speeded-up robust features. 31

**TN** True Negative. 89, 91, 102

**ToF** Time-of-Flight. 16, 17, 21, 22, 29, 30, 47

**TP** True Positive. 88, 89, 91, 96, 99, 102, 114

# Bibliography

[Aarts and Dijksterhuis, 2003] Aarts, H. and Dijksterhuis, A. (2003). The silence of the library: environment, situational norm, and social behavior. *Journal of Personality and Social Psychology*, 84(1):18–28.

[Abbate et al., 2012] Abbate, S., Avvenuti, M., Bonatesta, F., Cola, G., Corsini, P., and Vecchio, A. (2012). A smartphone-based fall detection system. *Pervasive and Mobile Computing*, 8(6):883–899.

[Aghajan et al., 2008] Aghajan, H., Wu, C., and Kleihorst, R. (2008). Distributed Vision Networks for Human Pose Analysis. In Mandic, D.; Golz, M.; Kuh, A.; Obradovic, D.; Tanaka, T., editor, *Signal Processing Techniques for Knowledge Extraction and Information Fusion*, pages 181–200. Springer US.

[Alwan et al., 2006] Alwan, M., Rajendran, P. J., Kell, S., Mack, D., Dalal, S., Wolfe, M., and Felder, R. (2006). A Smart and Passive Floor-Vibration Based Fall Detector for Elderly. In *Proceedings of the International Conference on Information & Communication Technologies: from Theory to Applications (ICTTA)*, volume 1, pages 1003–1007, Damascus, Syria. IEEE.

[Anderson et al., 2006] Anderson, D., Keller, J., Skubic, M., Chen, X., and He, Z. (2006). Recognizing falls from silhouettes. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, pages 6388–6391, New York, USA. IEEE.

[Anderson et al., 2009] Anderson, D., Luke, R. H., Keller, J. M., Skubic, M., Rantz, M., and Aud, M. (2009). Linguistic Summarization of Video for Fall Detection Using Voxel Person and Fuzzy Logic. *Computer Vision and Image Understanding*, 113(1):80–89.

[Bagalà et al., 2012] Bagalà, F., Becker, C., Cappello, A., Chiari, L., Aminian, K., Hausdorff, J. M., Zijlstra, W., and Klenk, J. (2012). Evaluation of accelerometer-based fall detection algorithms on real-world falls. *PLoS ONE*, 7(5):1–9.

[Bao et al., 2011] Bao, S. Y., Sun, M., and Savarese, S. (2011). Toward coherent object detection and scene layout understanding. *Image and Vision Computing*, 29(9):569–579.

[Belbachir et al., 2010] Belbachir, A. N., Lunden, T., Hanák, P., Markus, F., Böttcher, M., and Mannersola, T. (2010). Biologically-inspired stereo vision for elderly safety at home. *e & i Elektrotechnik und Informationstechnik*, 127(7):216–222.

[Bhattacharyya, 1943] Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society*, 35(99–109):4.

[Candamo et al., 2010] Candamo, J., Shreve, M., Goldgof, D., Sapper, D., and Kasturi, R. (2010). Understanding Transit Scenes: A Survey on Human Behavior-Recognition Algorithms. *IEEE Transactions on Intelligent Transportation Systems*, 11(1):206–224.

[Cardinaux et al., 2011] Cardinaux, F., Bhowmik, D., Abhayaratne, C., and Hawley, M. S. (2011). Video based technology for ambient assisted living: A review of the literature. *Journal of Ambient Intelligence and Smart Environments*, 3(3):253–269.

[Cha, 2008] Cha, S.-H. (2008). Taxonomy of nominal type histogram distance measures. In *Proceedings of the American Conference on Applied Mathematics (MATH)*, pages 325–330, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

[Chaaraoui et al., 2012] Chaaraoui, A. A., Climent-Pérez, P., and Flórez-Revuelta, F. (2012). A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living. *Expert Systems with Applications*, 39(12):10873–10888.

[Chan et al., 2009] Chan, M., Campo, E., Estève, D., and Fourniols, J.-Y. (2009). Smart homes - current features and future perspectives. *Maturitas*, 64(2):90–97.

[Chao et al., 2013] Chao, Y.-W., Choi, W., Pantofaru, C., and Savarese, S. (2013). Layout Estimation of Highly Cluttered Indoor Scenes Using Geometric and Semantic Cues. In *Proceedings of the Conference on Image Analysis and Processing (ICIAP)*, pages 489–499, Naples, Italy. Springer Berlin Heidelberg.

[Choi and Lee, 2003] Choi, E. and Lee, C. (2003). Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36:1703–1709.

[Chung and Liu, 2008] Chung, P.-C. and Liu, C.-D. (2008). A daily behavior enabled hidden Markov model for human behavior understanding. *Pattern Recognition*, 41(5):1572–1580.

[Cialdini and Trost, 1998] Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In *The Handbook of Social Psychology*, pages 151–192. McGraw-Hill, New York, USA.

[Comaniciu et al., 2000] Comaniciu, D., Ramesh, V., and Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 142–149, Hilton Head Island, USA. IEEE.

[Cuddihy et al., 2007] Cuddihy, P., Weisenberg, J., Graichen, C., and Ganesh, M. (2007). Algorithm to automatically detect abnormally long periods of inactivity in a home. In

*Proceedings of the International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments (HealthNet)*, pages 89–94, New York, USA. ACM.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, San Diego, USA. IEEE.

[Davis and Goadrich, 2006] Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 233–240, New York, USA. ACM Press.

[Delage et al., 2006] Delage, E., Lee, H., and Ng, A. Y. (2006). A Dynamic Bayesian Network Model for Autonomous 3D Reconstruction from a Single Indoor Image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2418–2428, New York, USA. IEEE.

[Delaitre et al., 2012] Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Gupta, A., and Efros, A. (2012). Scene semantics from long-term observation of people. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–298, Florence, Italy. Springer Berlin Heidelberg.

[Dellaert et al., 2000] Dellaert, F., Seitz, S., Thorpe, C., and Thrun, S. (2000). Structure from motion without correspondence. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 557–564, Pittsburgh, USA. IEEE.

[Deng and Yu, 2014] Deng, L. and Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*, 7:197–387.

[Deshpande et al., 2008] Deshpande, N., Metter, E. J., Bandinelli, S., Lauretani, F., Windham, B. G., and Ferrucci, L. (2008). Psychological, Physical, and Sensory Correlates of Fear of Falling and Consequent Activity Restriction in the Elderly : The InCHIANTI Study. *American Journal of Physical Medicine & Rehabilitation*, 87(5):354–362.

[Deshpande et al., 2009] Deshpande, N., Metter, J. E., Lauretani, F., Bandinelli, S., and Ferrucci, L. (2009). Interpreting Fear of Falling in the Elderly: What Do We Need to Consider? *Journal of Geriatric Physical Therapy*, 32(3):91–96.

[Dice, 1945] Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

[Diraco et al., 2010] Diraco, G., Leone, A., and Siciliano, P. (2010). An active vision system for fall detection and posture recognition in elderly healthcare. In *Proceedings of the Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1536–1541, Dresden, Germany. IEEE.

[Doukas et al., 2007] Doukas, C., Maglogiannis, I., Tragas, P., Liapis, D., and Yovanof, G. (2007). Patient Fall Detection using Support Vector Machines. In Boukis, C., Pnevmatikakis, A., and Polymenakos, L., editors, *Artificial Intelligence and Innovations 2007: from Theory to Applications*, volume 247 of *IFIP International Federation for Information Processing*, pages 147–156. Springer US.

[Duong et al., 2005] Duong, T., Bui, H., Phung, D., and Venkatesh, S. (2005). Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 838–845, San Diego, USA. IEEE.

[Dutta, 2012] Dutta, T. (2012). Evaluation of the Kinect sensor for 3-D kinematic measurement in the workplace. *Applied Ergonomics*, 43(4):645–649.

[Felzenszwalb et al., 2010] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object Detection with Discriminatively Trained Part-Based Models. *Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1627–1645.

[Floeck and Litz, 2008] Floeck, M. and Litz, L. (2008). Activity- and Inactivity-Based Approaches to Analyze an Assisted Living Environment. In *Proceedings of the International Conference on Emerging Security Information, Systems and Technologies (SECURWARE)*, pages 311–316, Cap Esterel, France. IEEE.

[Fofi et al., 2004] Fofi, D., Sliwa, T., and Voisin, Y. (2004). A comparative survey on invisible structured light. In Price, J. R. and Meriaudeau, F., editors, *Electronic Imaging 2004*, pages 90–98. SPIE.

[Fontecha et al., 2013] Fontecha, J., Navarro, F. J., Hervás, R., and Bravo, J. (2013). Elderly frailty detection by using accelerometer-enabled smartphones and clinical information records. *Personal and Ubiquitous Computing*, 17(6):1073–1083.

[Fouhey et al., 2012] Fouhey, D. F., Delaitre, V., Gupta, A., Efros, A. A., Laptev, I., and Sivic, J. (2012). People Watching: Human Actions as a Cue for Single-View Geometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 732–745, Florence, Italy. Springer Berlin Heidelberg.

[Gallimore and Lopez, 2002] Gallimore, R. and Lopez, E. M. (2002). Everyday Routines, Human Agency, and Ecocultural Context: Construction and Maintenance of Individual Habits. *OTJR: Occupation, Participation and Health*, 22(1 suppl):70–77.

[Galna et al., 2014] Galna, B., Barry, G., Jackson, D., Mhiripiri, D., Olivier, P., and Rochester, L. (2014). Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease. *Gait & Posture*, 39(4):1062 – 1068.

[Gibson, 1979] Gibson, J. J. (1979). *The ecological approach to visual perception.* Lawrence Erlbaum Associates, Hillsdale, USA.

[Gupta et al., 2011] Gupta, A., Satkin, S., Efros, A. A., and Hebert, M. (2011). From 3D scene geometry to human workspace. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1961–1968, Providence, USA. IEEE.

[Gupta et al., 2013] Gupta, S., Arbelaez, P., and Malik, J. (2013). Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 564–571, Portland, USA. IEEE.

[Han et al., 2013] Han, J., Shao, L., Xu, D., and Shotton, J. (2013). Enhanced computer vision with Microsoft Kinect sensor: a review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334.

[Hans et al., 2002] Hans, M., Graf, B., and Schraft, R. (2002). Robotic home assistant care-o-bot: past-present-future. In *Proceedings of the International Workshop on Robot and Human Interactive Communication (RO-MAN)*, pages 380–385, Berlin, Germany. IEEE.

[Hartley and Zisserman, 2004] Hartley, R. I. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition.

[He et al., 2012] He, Y., Li, Y., and Bao, S.-D. (2012). Fall detection by built-in tri-accelerometer of smartphone. In *Proceedings of the International Conference on Biomedical and Health Informatics (BHI)*, pages 184–187, Hong Kong, China. IEEE.

[Hedau et al., 2009] Hedau, V., Hoiem, D., and Forsyth, D. (2009). Recovering the spatial layout of cluttered rooms. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1849–1856, Kyoto, Japan. IEEE.

[Hedau et al., 2010] Hedau, V., Hoiem, D., and Forsyth, D. (2010). Thinking inside the box: Using appearance models and context based on room geometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 224–237, Heraklion, Greece. Springer Berlin Heidelberg.

[Hedau et al., 2012] Hedau, V., Hoiem, D., and Forsyth, D. (2012). Recovering free space of indoor scenes from a single image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2807–2814, Providence, USA. IEEE.

[Herbert Bay et al., 2006] Herbert Bay, Tuytelaars, T., and Gool, L. V. (2006). SURF: Speeded Up Robust Features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 404–417, Graz, Austria. Springer Berlin Heidelberg.

[Heywood, 2011] Heywood, J. L. (2011). Institutional Norms and Evaluative Standards for Parks and Recreation Resources Research, Planning, and Management. *Leisure Sciences*, 33(5):441–449.

[Hoey et al., 2010] Hoey, J., Poupart, P., Bertoldi, A. V., Craig, T., Boutilier, C., and Mihailidis, A. (2010). Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process. *Computer Vision and Image Understanding*, 114(5):503–519.

[Hoiem et al., 2008] Hoiem, D., Efros, A. A., and Hebert, M. (2008). Putting Objects in Perspective. *International Journal of Computer Vision*, 80(1):3–15.

[Howland et al., 1998] Howland, J., Lachman, M. E., Peterson, E. W., Cote, J., Kasten, L., and Jette, A. (1998). Covariates of Fear of Falling and Associated Activity Curtailment. *The Gerontologist*, 38(5):549–555.

[Jain et al., 2006] Jain, G., Cook, D. J., and Jakkula, V. (2006). Monitoring Health by Detecting Drifts and Outliers for a Smart Environment. In *Proceedings of the International Conference On Smart Homes and Health Telematics (ICOST)*, pages 114–121, Ulster, UK. IOS Press.

[Janoch et al., 2013] Janoch, A., Karayev, S., Jia, Y., Barron, J., Fritz, M., Saenko, K., and Darrell, T. (2013). A category-level 3-d object dataset: Putting the kinect to work. In *Proceedings of the ICCV Workshop on Consumer Depth Cameras for Computer Vision*, pages 141–165, Barcelona, Spain. IEEE.

[Jansen et al., 2007] Jansen, B., Temmermans, F., and Deklerck, R. (2007). 3D human pose recognition for home monitoring of elderly. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, pages 4049–4051, Lyon. IEEE.

[Jiang et al., 2013] Jiang, Y., Koppula, H., and Saxena, A. (2013). Hallucinated Humans as the Hidden Context for Labeling 3D Scenes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2993–3000, Portland, USA. IEEE.

[Kleinberger et al., 2007] 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 *Proceedings of the International Conference on Universal Access in Human-Computer Interaction (HCI International)*, pages 103–112, Beijing, China. Springer Berlin Heidelberg.

[Koppula et al., 2013] Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*, 32(8):951–970.

[Labayrade et al., 2002] Labayrade, R., Aubert, D., and Tarel, J.-P. (2002). Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In *Proceedings of the Intelligent Vehicle Symposium (IVS)*, volume 2, pages 646–651, Versailles, France. IEEE.

[Legters, 2002] Legters, K. (2002). Fear of Falling. *Physical Therapy*, 82(3):264–272.

[Leikas et al., 1998] Leikas, J., Salo, J., and Poramo, R. (1998). Security Alarm System Supports Independent Living of Demented Persons. *Gerontechnology: A Sustainable Investment in the Future*, 48:402–405.

[Litvak et al., 2008] Litvak, D., Zigel, Y., and Gannot, I. (2008). Fall detection of elderly through floor vibrations and sound. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, pages 4632–4635, Vancouver, Canada. IEEE.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110.

[Lu and Wang, 2012] Lu, J. and Wang, G. (2012). Human-centric indoor environment modeling from depth videos. In *Proceedings of the ECCV Workshops and Demonstrations*, pages 42–51, Florence, Italy. Springer Berlin Heidelberg.

[Lubinski, 1991] Lubinski, R. (1991). *Dementia and Communication*. B.C. Decker, Inc.

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, Vancouver, Canada. Morgan Kaufmann Publishers Inc.

[Mastorakis and Makris, 2012] Mastorakis, G. and Makris, D. (2012). Fall detection system using Kinect's infrared sensor. *Journal of Real-Time Image Processing*, 9:635–646.

[McKenna and Charif, 2005] McKenna, S. J. and Charif, H. N. (2005). Summarising contextual activity and detecting unusual inactivity in a supportive home environment. *Pattern Analysis and Applications*, 7(4):386–401.

[McKenna and Nait-Charif, 2004] McKenna, S. J. and Nait-Charif, H. (2004). Learning spatial context from tracking using penalised likelihoods. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 4, pages 138–141, Cambridge, UK. IEEE.

[McKhann et al., 1984] McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–944.

[Mihailidis et al., 2002] Mihailidis, A., Carmichael, B., and Boger, J. (2002). The Use of Computer Vision in an Intelligent Environment to Support Aging-in-Place, Safety, and Independence in the Home. *Gerontechnology*, 2(2):173–189.

[Miskelly, 2001] Miskelly, F. G. (2001). Assistive technology in elderly care. *Age and Ageing*, 30(6):455–458.

[Monekosso and Remagnino, 2010] Monekosso, D. N. and Remagnino, P. (2010). Behavior Analysis for Assisted Living. *IEEE Transactions on Automation Science and Engineering*, 7(4):879–886.

[Moshtaghi and Zukerman, 2014] Moshtaghi, M. and Zukerman, I. (2014). Modeling the Tail of a Hyperexponential Distribution to Detect Abnormal Periods of Inactivity in Older Adults. In *Proceedings of the Conference on Trends in Artificial Intelligence (PRICAI)*, pages 985–997, Gold Coast, Australia. Springer International Publishing.

[Moshtaghi et al., 2013] Moshtaghi, M., Zukerman, I., Albrecht, D., and Russell, R. A. (2013). Monitoring personal safety by unobtrusively detecting unusual periods of inactivity. In *Proceedings of the Conference on User Modeling, Adaptation and Personalization (UMAP)*, volume 7899, pages 139–151, Rome, Italy. Springer Berlin Heidelberg.

[Mutch and Lowe, 2006] Mutch, J. and Lowe, D. G. (2006). Multiclass Object Recognition with Sparse, Localized Features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 11–18, New York, USA. IEEE.

[Nait-Charif and McKenna, 2004] Nait-Charif, H. and McKenna, S. (2004). Activity summarisation and fall detection in a supportive home environment. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 4, pages 323–326, Cambridge, UK. IEEE.

[Nguyen et al., 2005] Nguyen, N., Phung, D., Venkatesh, S., and Bui, H. (2005). Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 955–960, San Diego, USA. IEEE.

[Noury et al., 2003] Noury, N., Barralon, P., Virone, G., Boissy, P., Hamel, M., and Rumeau, P. (2003). A smart sensor based on rules and its evaluation in daily routines. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, volume 4, pages 3286–3289, Cancun, Mexico. IEEE.

[Noury et al., 2008] Noury, N., Rumeau, P., Bourke, A. K., OLaighin, G., and Lundy, J. E. (2008). A proposal for the classification and evaluation of fall detectors. *Biomedical Engineering and Research IRBM*, 29(6):340–349.

[Obdrzalek et al., 2012] Obdrzalek, S., Kurillo, G., Ofli, F., Bajcsy, R., Seto, E., Jimison, H., and Pavel, M. (2012). Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, pages 1188–1193. IEEE.

[Oggier et al., 2004] Oggier, T., Lehmann, M., Kaufmann, R., Schweizer, M., Richter, M., Metzler, P., Lang, G., Lustenberger, F., and Blanc, N. (2004). An all-solid-state

optical range camera for 3D real-time imaging with sub-centimeter depth resolution (SwissRanger). In *Optical Design and Engineering*, volume 5249, pages 534–545. SPIE.

[Planinc and Kampel, 2012a] Planinc, R. and Kampel, M. (2012a). Introducing the Use of Depth Data for Fall Detection. *Personal and Ubiquitous Computing*, 17(6):1063–1072.

[Planinc and Kampel, 2012b] Planinc, R. and Kampel, M. (2012b). Robust Fall Detection by Combining 3D Data and Fuzzy Logic. In *Proceedings of the ACCV Workshop on Color Depth Fusion in Computer Vision*, pages 121–132, Daejeon, Korea. Springer Berlin Heidelberg.

[Planinc and Kampel, 2013] Planinc, R. and Kampel, M. (2013). Detecting Changes in Elderly's Mobility Using Inactivity Profiles. In *Proceedings of the International Work-conference on Ambient Assisted Living (IWAAL)*, pages 100–103, Carrillo, Costa Rica. Springer Switzerland.

[Planinc and Kampel, 2014a] Planinc, R. and Kampel, M. (2014a). Combining Spatial and Temporal Information for Inactivity Modeling. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 4234–4239, Stockholm, Sweden. IEEE.

[Planinc and Kampel, 2014b] Planinc, R. and Kampel, M. (2014b). Detecting Unusual Inactivity by Introducing Activity Histogram Comparisons. In *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 313–320, Lisbon, Portugal. SCITEPRESS.

[Planinc and Kampel, 2014c] Planinc, R. and Kampel, M. (2014c). Human Centered Scene Understanding based on Depth Information - How to Deal with Noisy Skeleton Data? In *Proceedings of the International Symposium on Visual Computing (ISVC)*, pages 609–618, Las Vegas, USA. Springer International Publishing.

[Planinc and Kampel, 2015a] Planinc, R. and Kampel, M. (2015a). Human Centered Scene Understanding based on 3D Long-Term Tracking Data. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis (IbPria)*, page to appear, Santiago de Compostela, Spain. Springer-Verlag Berlin Heidelberg.

[Planinc and Kampel, 2015b] Planinc, R. and Kampel, M. (2015b). Local Behavior Modeling based on Long-Term Tracking Data. In *Proceedings of the International Conference on Machine Vision Applications (MVA)*, page to appear, Tokyo, Japan. MVA Organization.

[Plantard et al., 2015] Plantard, P., Auvinet, E., Pierres, A.-S., and Multon, F. (2015). Pose Estimation with a Kinect for Ergonomic Studies: Evaluation of the Accuracy Using a Virtual Mannequin. *Sensors*, 15:1785–1803.

[Pollefeys et al., 2007] Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., Salmi, C., Sinha, S.,

Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G., and Towles, H. (2007). Detailed Real-Time Urban 3D Reconstruction from Video. *International Journal of Computer Vision*, 78(2-3):143–167.

[Popoola, 2012] Popoola, O. P. (2012). Video-Based Abnormal Human Behavior Recognition - A Review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):865–878.

[Pramerdorfer, 2013] Pramerdorfer, C. (2013). Depth Data Analysis for Fall Detection. Master's thesis, Vienna University of Technology, Austria.

[Rangarajan et al., 2007] Rangarajan, S., Kidane, A., Qian, G., Rajko, S., and Birchfield, D. (2007). The design of a pressure sensing floor for movement-based human computer interaction. In *Proceedings of the Conference on Smart Sensing and Context (EuroSSC)*, pages 46–61, Kendal, England. Springer Berlin Heidelberg.

[Rashidi and Mihailidis, 2013] Rashidi, P. and Mihailidis, A. (2013). A survey on ambient-assisted living tools for older adults. *IEEE Journal of Biomedical and Health Informatics*, 17(3):579–590.

[Rodgers and Nicewander, 1988] Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66.

[Rougier et al., 2011] Rougier, C., Auvinet, E., Rousseau, J., Mignotte, M., and Meunier, J. (2011). Fall Detection from Depth Map Video Sequences. In *Proceedings of the Conference on Smart Homes and Health Telematics (ICOST)*, pages 121–128. Springer Berlin Heidelberg, Montreal, Canada.

[Rougier et al., 2006] Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2006). Monocular 3d head tracking to detect falls of elderly people. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, pages 6384 –6387, New York, USA. IEEE.

[Rougier et al., 2007] Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2007). Fall detection from human shape and motion history using video surveillance. In *Proceedings of the International Conference on Advanced Information Networking and Applications Workshops (AINAW)*, volume 2, pages 875–880, Niagara Falls, Canada. IEEE.

[Rubner et al., 2000] Rubner, Y., Tomasi, C., and Guibas, L. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121.

[Sansoni et al., 2009] Sansoni, G., Trebeschi, M., and Docchio, F. (2009). State-of-The-Art and Applications of 3D Imaging Sensors in Industry, Cultural Heritage, Medicine, and Criminal Investigation. *Sensors*, 9(1):568–601.

[Särelä et al., 2003] Särelä, A., Korhonen, I., Lotjonen, J., Sola, M., and Myllymaki, M. (2003). Ist vivago - an intelligent social and remote wellness monitoring system for the elderly. In *Proceedings of the EMBS Special Topic Conference on Information Technology Applications in Biomedicine (ITAB)*, pages 362 – 365, Birmingham, England. IEEE.

[Scanaill et al., 2006] Scanaill, C., Carew, S., Barralon, P., Noury, N., Lyons, D., and Lyons, G. (2006). A Review of Approaches to Mobility Telemonitoring of the Elderly in Their Living Environment. *Annals of Biomedical Engineering*, 34(4):547–563.

[Schodl and Essa, 2001] Schodl, A. and Essa, I. (2001). Depth layers from occlusions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 639–644, Kauai, USA. IEEE.

[Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, Colorado Springs, USA. IEEE.

[Silberman et al., 2012] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 746–760, Florence, Italy. Springer Berlin Heidelberg.

[Smisek et al., 2011] Smisek, J., Jancosek, M., and Pajdla, T. (2011). 3D with Kinect. In *Proceedings of the International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1154–1160, Barcelona, Spain. IEEE.

[Stone and Skubic, 2014] Stone, E. and Skubic, M. (2014). Fall Detection in Homes of Older Adults Using the Microsoft Kinect. *Journal of Biomedical and Health Informatics*, 19(99):290–301.

[Stoyanov et al., 2011] Stoyanov, T., Louloudi, A., Andreasson, H., and Lilienthal, A. J. (2011). Comparative evaluation of range sensor accuracy in indoor environments. In *Proceedings of the European Conference on Mobile Robots (ECMR)*, pages 19–24, Örebro, Sweden.

[Sun et al., 2009] Sun, H., Florio, V. D., Gui, N., and Blondia, C. (2009). Promises and Challenges of Ambient Assisted Living Systems. In *Proceedings of the International Conference on Information Technology: New Generations (ITNG)*, pages 1201–1207, Las Vegas, USA. IEEE.

[Sung et al., 2011] Sung, J., Ponce, C., Selman, B., and Saxena, A. (2011). Human activity detection from rgbd images. In *Proceedings of the AAAI workshop on Plan, Activity and Intent Recognition (PAIR)*, pages 842–849, San Francisco, USA. AAAI.

[Swain and Ballard, 1991] Swain, M. and Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

[Tacconi et al., 2011] Tacconi, C., Mellone, S., and Chiari, L. (2011). Smartphone-based applications for investigating falls and mobility. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 258–261, Dublin, Ireland. IEEE.

[Taylor and Mai, 2013] Taylor, G. and Mai, F. (2013). Behind the Scenes: What Moving Targets Reveal about Static Scene Geometry. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, pages 546–553, Sydney, Australia. IEEE.

[Tsai and Kuipers, 2011] Tsai, G. and Kuipers, B. (2011). Real-time indoor scene understanding using Bayesian filtering with motion cues. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 121–128, Barcelona, Spain. IEEE.

[van Diest et al., 2014] van Diest, M., Stegenga, J., Wörtche, H. J., Postema, K., Verkerke, G. J., and Lamoth, C. J. (2014). Suitability of kinect for measuring whole body movement patterns during exergaming. *Journal of Biomechanics*, 47(12):2925 – 2932.

[Van Rijsbergen, 1979] Van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworths, London, UK.

[Virone et al., 2008] Virone, G., Alwan, M., Dalal, S., Kell, S. W., Turner, B., Stankovic, J., and Felder, R. (2008). Behavioral patterns of older-adults in assisted living. *IEEE transactions on Information Technology in Biomedicine*, 12(3):387–398.

[Virone and Sixsmith, 2008] Virone, G. and Sixsmith, A. (2008). Monitoring activity patterns and trends of older adults. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, pages 2071–2074, Vancouver, Canada. IEEE.

[Wang et al., 2012] Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, Providence, USA. IEEE.

[Webster and Celik, 2014] Webster, D. and Celik, O. (2014). Systematic review of Kinect applications in elderly care and stroke rehabilitation. *Journal of Neuroengineering and Rehabilitation*, 11(108):1–24.

[Weisenberg et al., 2008] Weisenberg, J., Cuddihy, P., and Rajiv, V. (2008). Augmenting motion sensing to improve detection of periods of unusual inactivity. In *Proceedings of the International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments (HealthNet)*, pages 2:1–2:6, Breckenridge, USA. ACM Press.

[Wild et al., 1981] Wild, D., Nayak, U. S., and Isaacs, B. (1981). How dangerous are falls in old people at home? *British medical journal (Clinical research ed.)*, 282(6260):266–268.

[Xia et al., 2012] Xia, L., Chen, C., and Aggarwal, J. (2012). View invariant human action recognition using histograms of 3d joints. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 20–27, Providence, USA. IEEE.

[Xinguo, 2008] Xinguo, Y. (2008). Approaches and principles of fall detection for elderly and patient. In *Proceedings of the International Conference on e-health Networking, Applications and Services (HealthCom)*, pages 42–47, Singapore, Singapore. IEEE.

[Yang and Ramanan, 2011] Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, Colorado Springs, USA. IEEE.

[Zadeh, 1965] Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

[Zambanini et al., 2010] Zambanini, S., Machajdik, J., and Kampel, M. (2010). Early versus Late Fusion in a Multiple Camera Network for Fall Detection. In *Proceedings of the Annual Workshop of the Austrian Association for Pattern Recognition (ÖAGM)*, volume 819862, pages 15–22, Zwettl, Austria.

[Zhang et al., 2011] Zhang, Z., Kapoor, U., Narayanan, M., Lovell, N. H., and Redmond, S. J. (2011). Design of an Unobtrusive Wireless Sensor Network for Nighttime Falls Detection. In *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*, pages 5275–5278, Boston, USA. IEEE.

[Zhao et al., 2007] Zhao, J., Katupitiya, J., and Ward, J. (2007). Global Correlation Based Ground Plane Estimation Using V-Disparity Image. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 529–534, Rome, Italy. IEEE.

[Zweng et al., 2010] Zweng, A., Zambanini, S., and Kampel, M. (2010). Introducing a Statistical Behavior Model into Camera-Based Fall Detection. In *Proceedings of the International Symposium on Visual Computing (ISVC)*, volume 6453, pages 163–172, Las Vegas, USA. Springer Berlin Heidelberg.

# Online References

[AAL Joint Programme, 2012] AAL Joint Programme (2012). Objectives. http://www.aal-europe.eu/about/objectives/. [Online; accessed 31-March-2015].

[Asus, 2011] Asus (2011). Asus Xtion pro. https://www.asus.com/Multimedia/Xtion_PRO/. [Online; accessed 31-March-2015].

[Azimi, 2012] Azimi, M. (2012). Skeletal Joint Smoothing. http://msdn.microsoft.com/en-us/library/jj131429.aspx. [Online; accessed 31-March-2015].

[Dollár, 2012] Dollár, P. (2012). Piotr's Image and Video Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html. [Online; accessed 31-March-2015].

[McDonald and Levin, 2011] McDonald, K. and Levin, G. (2011). Kinect Workshop. http://futuretheater.net/wiki/Kinect_Workshop. [Online; accessed 31-March-2015].

[Microsoft, 2010] Microsoft (2010). Kinect for Xbox360. http://news.microsoft.com/2010/06/13/. [Online; accessed 31-March-2015].

[Microsoft, 2013] Microsoft (2013). Enhanced sensing in Xbox One. http://blogs.microsoft.com/blog/2013/10/02/collaboration-expertise-produce-enhanced-sensing-in-xbox-one/. [Online; accessed 31-March-2015].

[Microsoft, 2015] Microsoft (2015). Kinect for Windows. http://www.microsoft.com/en-us/kinectforwindows/. [Online; accessed 31-March-2015].

[Microsoft Developer Network, 2015] Microsoft Developer Network (2015). Skeletal Tracking. https://msdn.microsoft.com/en-us/library/hh973074.aspx. [Online; accessed 31-March-2015].

[OpenNI, 2011] OpenNI (2011). http://www.openni.org. [Online; accessed 10-April-2014].

[United Nations, Department of Economic and Social Affairs, 2001] United Nations, Department of Economic and Social Affairs (2001). World population ageing: 1950-2050. http://www.un.org/esa/population/publications/worldageing19502050/. [Online; accessed 31-March-2015].

[United Nations, Department of Economic and Social Affairs, 2013] United Nations, Department of Economic and Social Affairs (2013). World population ageing 2013. http://www.un.org/en/development/desa/population/publications/ageing/WorldPopulationAgeing2013.shtml. [Online; accessed 31-March-2015].

# Curriculum Vitae

**Rainer Planinc**
**Mühlhäufelgasse 27/5/4**
**A-1220 Vienna, Austria**

**Phone: +43 1 58801 18389**
**Email: rainer.planinc@tuwien.ac.at**

**Personal Information:**

| | |
|---|---|
| Nationality: | Austria |
| Date of Birth: | 15.8.1984 |
| Place of Birth: | Vienna, Austria |

**Education**:

| | |
|---|---|
| 10/2010 – 06/2015 | **Doctoral Programme in Technical Sciences**<br>Vienna University of Technology, Computer Vision Lab |
| 10/2007 – 06/2010 | **Vienna University of Technology**<br>*Field of study*: Media Informatics (Master), Computer Science Management (Master)<br>*Thesis*: Modeling Sources and Sinks in Crowded Scenes by Clustering Trajectory Points Obtained by Video-based Particle Advection,<br>Didaktische Aufbereitung bekannter Video-Kompressionsverfahren („Didactical treatment of video compression methods“) |
| 10/2004 – 06/2007 | **Vienna University of Technology**<br>*Field of study*: Media Informatics (Bachelor) |

**Work Experience:**

| | |
|---|---|
| since 06/2010 | **Vienna University of Technology, Computer Vision Lab**<br>Research Assistant in national and international projects |
| 02/2015 | **Castilla-La Mancha University, MAmI Research Lab, Spain**<br>Visiting PhD student, Advisor: Prof. Dr. José Bravo Rodriguez |
| 09/2008 – 06/2010 | **Austrian Institute of Technology, Mobility Department**<br>Work experience and diploma thesis in the area of automatic video analysis of dense pedestrian motion flows |