Video-Based Fall Detection by Multiview 3D Spatial Features

August 2014

Dao Huu Hung

A Thesis for the Degree of Ph.D. in Engineering

Video-Based Fall Detection by Multiview 3D Spatial Features

August 2014

Graduate School of Science and Technology
Keio University

Dao Huu Hung

Dedicated to my Parents
Dao Huu Hoc and Hoang Thi Thom
My Wife and Son
Nguyen Thi Hang and Dao Huu Nhat Hung
My Brother
Dao Huu Hiep

Abstract

In this dissertation, we address the problem of detecting fall incidents by using vision technology. This problem is critical to ensure the safety of the elderly who increasingly prefer to live alone at home but are prone to suffer from accidental falls. The aim of detecting falls instantly is to offer immediate help to fallen elderly, in turn, not worsening their injuries or even saving their lives. Even though a large body of literature have been dedicating to fall detection, many challenges still remain for further investigation. One of the major challenges is to discriminate carefully falls from various activities of daily living (ADL), especially like-fall ones. e.g., crouch on the ground and sit down brutally, etc. Secondly, fall detection seems to be meaningless without real-time performance. Other challenges include low image quality, cluttered background, illumination variations, appearance variations, camera viewpoints, and occlusion by furniture, etc.

We realize that falls are associated with fast body movements to change postures from upright to almost lengthened, followed by a sufficient duration of staying almost motionlessly on the ground. This is contrary to slow manners of doing ADL of the elderly. Hence, we propose using 3D spatial features which are efficiently estimated from multiple views and are highly discriminative to classify human states into standing, sitting, and lying. Once a sequence of human states is given, fall events can be reliably inferred by analyzing human state transition.

Firstly, we describe in this dissertation a combination of heights and occupied areas, extracted from 3D cuboids of the person of interest for human state classification. Lying people take larger areas than sitting and standing people. Standing people are higher than sitting and lying people. These three states intuitively lie in three separable region of the feature space which can be classified by SVM. Falls are inferred by time-series analysis of human state transition. For efficient feature estimation, we configure two cameras whose fields of view are relatively orthogonal. Thus, 2D bounding boxes of the person, extracted from two views, serve as two orthographic projections of the 3D cuboids. The features are normalized by using Local Empirical

Templates which are defined as foregrounds of standing people in local image patches and can be obtained automatically in unknown scenes. The normalization cancels the viewing perspective and makes the features invariant across viewing window. Our experiments on multiple camera fall dataset produce comparable performance with state-of-the-art methods, tested on the same dataset but demonstrate lower computational cost.

By using height and occupied area, we can distinguish lying from standing and sitting states. But the information of where the person lies either on a sofa, for example, in normal situations or on the ground after falls is unknown. Consequently, this method is able to detect a state change from standing to lying as a fall. Sit-to-stand-transfer falling type in which people change from sitting to lying states is not considered. Therefore secondly, we present in this dissertation a low-cost scheme of estimating Human-Ground Contact Areas (HGCA) for fall detection. Standing and sitting people make a little contact with the ground, mainly by feet. But lying people lie almost completely on the ground after falls. Hence, HGCA is a good feature for not only classifying human states but also indicating where the person lies either on the ground or on a sofa. To measure HGCA, we project foregrounds of the person from one view to another by using the homography of the ground between views. Overlap regions between the foreground in the latter view and the projected foreground that only exist where people contact with the ground, i.e., feet location, due to the plane parallax, are measured as HGCA. We also propose a human state simulation in which a virtual camera captures various 3D human models in different states from a variety of angles to generate training samples. View-invariant distributions of HGCA with respect to human states are built from the training samples to generalize a threshold to separate lying from standing and sitting states. Temporal analysis of human state transition is used to infer falls. We also test this method on multiple camera fall dataset, leading to competitive performance and lower computational cost than state-of-the-art methods, performed on the same dataset.

Recently, Bag-of-Video-Word (BoVW) approaches have been showing good performance on a wide range of human action recognition datasets. However, to the

best of our knowledge, there is no work evaluating BoVW approaches on a dataset, exclusively dedicated to fall detection. Hence, we carry out an empirical study to access the effectiveness of BoVW approaches to fall detection. The standard BoVW approach with Chi-square kernel SVM classifier are tested against multiple camera fall dataset in Leave-one-scene-out cross validation setup, resulting in favorably comparable performance with our proposed methods, except its heavy computational cost.

We do hope that our research outcomes will contribute significantly a step toward the commercialization of vision-based fall detection technology which not only enhances the quality of life, quality of care, safety of the elderly but also fosters their autonomy and freedom.

Acknowledgment

My three-year journey of completing the PhD degree is now going to come to an fruitful end. This journey is probably the most challenging that I have encountered in the first 30 years of my life with full of up and down moments and emotions. It is a great privilege to be with Department of Information and Computer Science at Keio University, Yokohama, Japan, particularly with the Hyper Vision Research Laboratory (Hideo Saito Lab) whose members will always engrave in my mind. This endeavor could not have been successfully finished without the supports of many people.

First and foremost, I would like to express my deep sense of gratitude to my academic advisor, Hideo Saito sensei (Professor in Japanese). He always patiently guides me through many challenges I have encountered in the journey by providing great vision, encouragement, hard discussions, and necessary advices to nurture inchoate ideas. Without his commitment, expertise and leadership, my endeavor in the doctoral program could not have been finished. I always feel to be in his debt.

Secondly, special thanks must be sent to committee professors of my dissertation defense, Yoshimitsu Aoki sensei, Yasue Mitsukura sensei, and Maki Sugimoto sensei. I deeply appreciate their time, support, guidance and commitment in reviewing this dissertation with precious comments. Without such helpful comments, this dissertation could not have reached to this high quality.

Thirdly, I am grateful to all of my lab members for their friendship and constant support. We have shared a lot of exciting moments and emotions not only in study but also in daily lives. Particularly, my thank should goes to Hirose-san and Takumi-san (Mr. in Japanese) for their welcome and hospitality to escort me at Musashi-Kosugi station in the first day I came to Japan. Without their helps, many legal documents and procedures in Japanese could not be finished quickly that made my new life in japan settle at ease. I should also mention Big-san, a nice guy and one of my best friends, who was escorted along with me by these two guys. I also need to thank Sandy, my senior who helped me a lot in this journey. I had a nice trip to conference in

Okinawa with Ikeda san and thank for his support. Lastly, special thanks to Tamaki-san, Honda-san, Shinozuka-san, Nakagawa-san, Nakayama-san, Kawasaki-san, and other lab members.

I also need to include staffs in YISH (Yokohama International Student House) in which I spent 2 years on living. They gave me a good opportunity to live in an international environment with many social activities with local people so that I not only discover Japanese culture and traditions but also experience those of other countries. YISH is also a special place for my family in which we welcomed our first son born in Japan.

Most importantly, I wish to thank my parents, Dao Huu Hoc and Hoang Thi Thom, my brother Dao Huu Hiep, my wife, Nguyen Thi Hang, and my first son, Dao Huu Nhat Hung. Their endless love, patience and sacrifice are the great inspiration and motivation for me to complete the doctoral program. This dissertation and the PhD degree are dedicated to all my dear family members. I hope that this work makes you proud.

Finally, I would like to acknowledge the financial support from MEXT scholarship for my doctoral program at Keio University.

Yokohama, July 12, 2014

Dao Huu Hung

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

1.1.1 Population aging

Population aging is the process by which older people[1] account for an increasingly high percentage of total population. According to a report of United Nation [United Nation, 2002], population aging is "unprecedented in human history", "enduring", and "pervasive", spreading nearly all over the world by different paces. Developed countries have been suffering from population aging for over a decade by a fast growing manner. The number of the elderly exceeding that of children[2] happened for the first time in these nations in 1998 [United Nation, 2007]. Nowadays, the ratios of the elderly to children in Japan, Germany, and Italy, to name a few as examples, are around 1.85 [The-World-Factbook, 2013c], 1.6 [The-World-Factbook, 2013a], and 1.56 [The-World-Factbook, 2013b], respectively. Although the percentage of older people in developing nations today is just around 8% of the population, it is predicted that by mid-century, population aging will progress quickly to the same current level of developed nations [United Nation, 2007].

Population aging leads to tremendous political, economical and social problems

[1] their age is over 65
[2] their age is under 15

and consequences among which the burden on the healthcare system, especially the elderly care, is in our central consideration. Since the elderly are prone to health deterioration, malnutrition, senility, depression and isolation, etc., the elderly care are therefore in high demand. However, it consumes much government fundings and family budgets, as well as requires a lot of human resources. Seeking solutions to reduce the cost and the dependence on the foreseen shortage of nurses attracts attention of the whole society, ranging from decision makers to researchers. It is urgent for us to have actions quickly because of rapid permeation of population aging across the world.

1.1.2 Home telecare and assistive technology

To cut down the cost and improve the quality of the elderly care, healthcare centers should be more specialized and centralized into a few places. Institution care should be shifted to more advanced home healthcare, thanks to advances in information technology and assistive technology facilitating this trend [Koch, 2006]. These centralized and specialized healthcare centers keep a weather eye on health conditions of the elderly from a distance and provide instant assistance upon detected abnormality and/or emergency. It creates an alternative and promising model of the elderly care, so-called *home telecare*. Home telecare promotes greatly independence of the elderly who prefer to live alone, separately from their relatives. These advantages of home telecare make it very prevalent among healthcare provision areas [Ruggiero et al., 1999].

Assistive technology (AT) "is an umbrella term for any device or system that allows an individual to perform a task they would otherwise be unable to do or increases the ease and safety with which the task can be performed" [Cowan and Turner-Smith, 1999]. Evidently, AT plays an important role in home telecare. It keeps a weather eye on everyday conditions of the elderly by plenty of smart devices and sensors, implanted into their houses [Chan et al., 2008], automatically detects accidents or emergency cases [Lee and Mihailidis, 2005], supports people with dementia [Bharucha et al., 2009], replaces human caregivers by virtual caregivers [Hossain and Ahmed, 2012],

and provides social interactions by assistive social robots [Broekens et al., 2009], etc. Advances in information technology and assistive technology ease the burden on human caregivers, guarantee security and safety, promote independence and autonomy, and lessen depression and isolation of the elderly, etc., in turn, improving significantly quality of life.

1.1.3 Importance of fall detection in home telecare

Among these crucial applications of AT to home telecare, security and safety assurance is at the heart of our interests since the elderly living alone are considered as an "at-risk" group [Kharicha et al., 2007]. They appear to be associated with higher risks of accidental falls that happen frequently and have profound implications [MacCulloch et al., 2007]. Falls are considered as the most common cause of injuries [Yu, 2008] and the sixth leading cause of death [MacCulloch et al., 2007] of the elderly. The severity of injuries is proportional to the delayed time in receiving medical treatments. Timely responses help fallen people not worsen the injuries or even save their lives. Hence, we should detect falls as soon as possible to offer immediate treatments, preventing injuries from further severe.

The study of accidental falls of the elderly can be broken into two categories, fall prevention and fall detection. There are many practical programs and research works in the literature, dedicating to the former. Exercise interventions to strengthen muscle and balance, in turn, reducing fall risks, a number of medications associated with fall risk reduction, and injury protection such as using hip fracture protector are investigated [MacCulloch et al., 2007]. However, the purposes of fall prevention researches merely seem to discover how to lessen fall risks but not completely. By its nature, it seems to be impossible to predict whether a fall happens for prevention.

In contrast, the latter detecting a fall right after it happened is more straightforward than the former. Moreover, fall detection methods are very important to ensure the safety of the elderly living alone in the consideration that accidental falls seem to be unavoidable or unpreventable. Healthcare industry has been realizing a very useful application of fall detection methods at the heart of Intelligent Personal Emergency

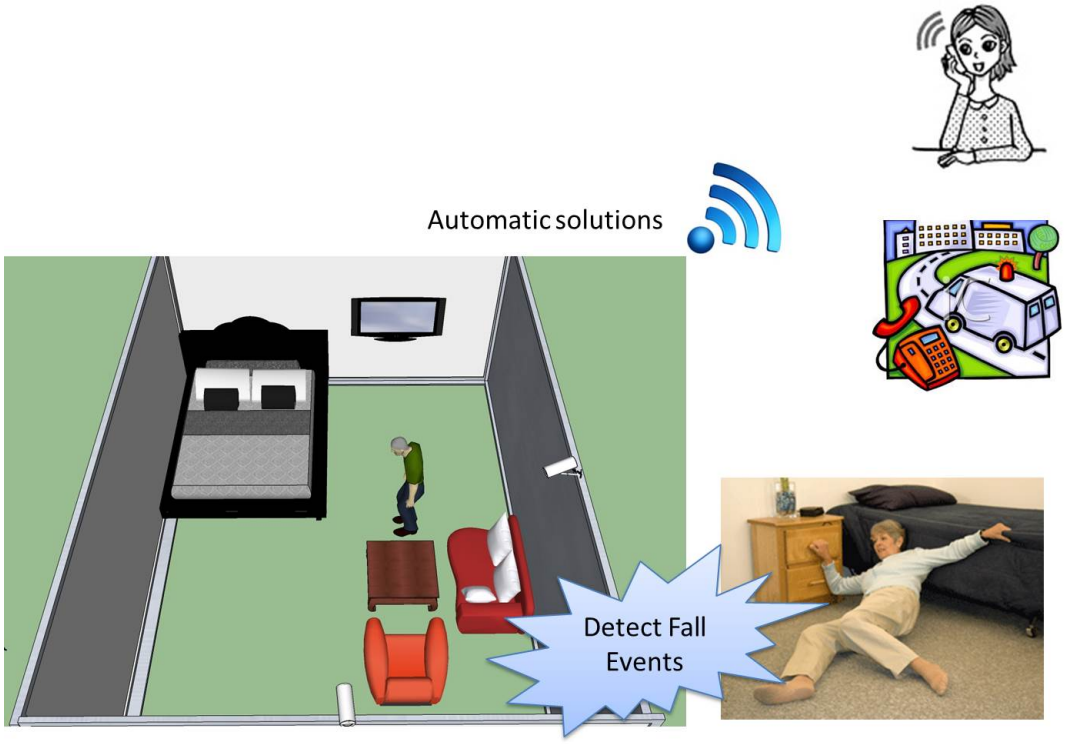


Figure 1-1: The application of iPERS to ensure the safety of the elderly living alone at home. Emergency situations like an accidental fall, are automatically sensed to trigger a console to make an instant notification to family members and/or to connect with a designated emergency response center for immediate help.

Response System (iPERS) to ensure safety of the elderly living alone at home [Lee and Mihailidis, 2005]. General speaking, iPERS is conventional PERS with an added capability of automatic sensing of emergencies. Conventional PERS composes of a small radio transmitter, a console connecting to users' telephone, and an emergency response center that handles this type of calls [Doughty et al., 1996]. In emergency situations, the users press HELP button, usually attached in an easily accessed place on users' body, to contact with a designated emergency response center to receive instant necessary assistance. However, conventional PERS exposes a major weakness that prevents it from practical usage of the elderly. Users must carry the HELP button 24 hours a day. It is not an easy task for the elderly since most of them suffer from dementia or deterioration of cognitive ability [Bharucha et al., 2009] and feel uncomfortable [Yu, 2008]. Moreover, the impact of shock after falls may force the elderly to experience unconscious states of mind as well as physical pain. Pressing

HELP button to call for emergency assistance seems to be inappropriate in practice. Therefore, iPERS [Lee and Mihailidis, 2005] that is capable of providing automatic sensing of emergencies is favorable in the elderly care. Falls are automatically detected to trigger the console to make an instant notification to family members and/or to connect with the designated emergency response center for immediate help, as illustrated in Fig. 1-1. That is why fall detection has been being an active research for recent years by a large body of literature [Yu, 2008; Noury et al., 2007; Ward et al., 2012; Mubashir et al., 2013; Spasova and Iliev, 2014]. *Our dissertation is also devoted to the problem of fall detection of the elderly.*

1.1.4 Technologies for fall detection

The common pipeline of fall detection methods, shown in Fig. 1-2, includes three main parts, sensors, feature extraction and classification, and fall inference [Yu, 2008]. This section summarizes a variety of sensors that can be used in fall detection. Depending on where to place sensors, we can break fall detection methods into three categories, corresponding to wearable-device, ambient-device and vision technologies. We only highlight strengths and weaknesses of each technology and also explain why vision technology should be treated separately from ambient-device technology. Please refer to comprehensive reviews [Yu, 2008; Ward et al., 2012; Mubashir et al., 2013; Spasova and Iliev, 2014] to have insights into methods.

Wearable-device technology

Wearable-device technology for fall detection includes motion and posture sensors, attached on the human body [Mathie et al., 2004; Wang et al., 2008] or on the garment [Lin et al., 2007; Nyan et al., 2006]. These sensors, i.e. accelerometers [Mathie et al., 2004; Wang et al., 2008; Nyan et al., 2006], and gyroscope [Nyan et al., 2006; Tamura et al., 2009] etc., measure motion, location, posture and Electromyography (EMG) signals [Ghasemzadeh et al., 2009] of human body for fall inference by thresholding [Tamura et al., 2009] or machine learning techniques [Ghasemzadeh et al., 2009;

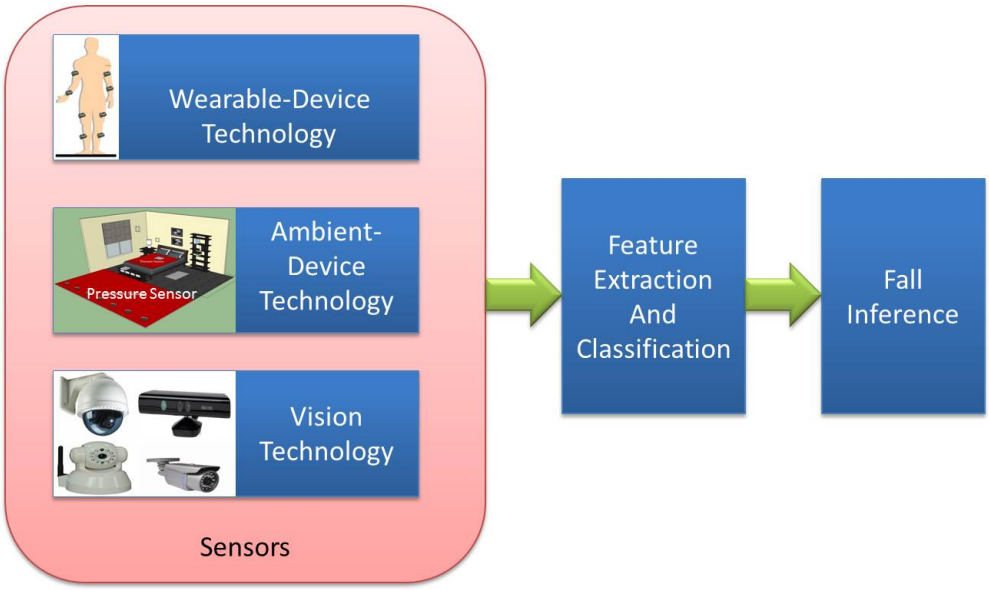


Figure 1-2: The common pipeline of fall detection methods.

Doukas et al., 2007]. The wearable-device technology offers cheap[3] solutions with simple initialization, setup, and operation. But wearing and attaching something on the body are intrusive and uncomfortable. As a result, many people are unwilling to accept this technology [Yu, 2008]. Moreover, sensors are designed to work effectively at some particular positions on the body that may be changed unexpectedly during operation, easily leading to low detection rates and high false alarms [Yu, 2008; Mubashir et al., 2013].

Ambient-device technology

Ambient-device technology refers to sensor-embedded environments that are sensitive and responsive to the presence of human [Yu, 2008]. Vibration sensors [Alwan et al., 2006], pressure sensors [Scott, 2000] either on the floor or under the bed mattress, thermal imaging sensors [Sixsmith and Johnson, 2004] and audio processing [Zhuang et al., 2009], etc. are utilized in fall detectors. The merits of ambient-device technology are cost-effective and non-intrusive. However, pressure sensors-based fall detectors, relying on pressure measured from environments, are very sensitive to environmental changes and people carrying objects, etc. leading to poor performance

[3]except wearable garment

[Mubashir et al., 2013]. Moreover, both ambient-device and wearable-device technologies do not provide post visual inspection and verification for caregivers in cases of alarms [Yu, 2008].

Vision technology

Vision technology has been flourishing for over decades. Cameras with increasingly high quality are cheaper and cheaper, along with a mature of vision algorithms. Revolution in computing devices makes vision technology feasible in real-time applications. As a consequence, cameras are nowadays very prevalent on our doorstep. Vision technology are progressively permeated into every corner of life, ranging from video surveillance, biometrics, virtual reality to medicine, etc. In particular, it has been revolutionizing today's telecare services and e-healthcare systems [Hossain et al., 2012].

In this section, we treat vision technology separately although vision sensors can be sorted into ambient-device technology. It is because vision sensors have numerous advantages over above sensors. Apart from its low-cost and non-intrusive properties, most importantly, information extracted from vision sensors is richer than that of other sensors. It allows us to perform not only fall detection but also other human action/activity recognition [Poppe, 2010], security surveillance [Hu et al., 2004], health diagnosis via gait recognition [Pogorelc et al., 2012], human emotion recognition [Fasel and Luettin, 2003], and virtual social interactions [Broekens et al., 2009], etc. Vision technology opens an interesting perspective of replacing human caregivers by intelligent virtual caregivers [Hossain and Ahmed, 2012] that non-intrusively monitor health conditions, safety and security of the elderly, understand their feelings, try to offer their needs, remind them how to perform a daily task, and call for human caregivers when necessary, etc. *Therefore in this dissertation, we only consider vision-based methods of fall detection.*

1.2 Research Objectives

Our research objectives are summarized as the following.

1. The most important objective is to find a solution that is capable of reaching high detection rates since the application of fall detection relates to human safety. False alarms also need to be kept as low as possible. Otherwise, emergency response centers will be frequently disturbed. In daily life, people exhibit a variety of actions/activities,[4] many of which and falls have some characteristics in common, for example, fast motion followed by a relatively motionless duration. Crouching, lying either on a bed, a sofa, or on the ground and sitting down brutally are among common confounding or like-fall actions. Desired solutions must distinguish precisely fall events from like-fall ones.

2. A major issue in video surveillance, particularly in multiview approaches, is about using complicated initialization, registration, calibration and site models, making the proposed solution unfeasible [Javed and Shah, 2008]. During operation, camera viewpoints may be changed unexpectedly probably due to natural and human factors, etc. Consequently, we need to recalibrate cameras, re-attain site models, re-run the initialization, or make the registration again. Otherwise, the performance will be deteriorated, leading to detrimental effects on human safety. Maintaining stable performance of such camera network during operation is a daunting task [Javed and Shah, 2008]. Thus, our objective is to ideally find an automatic solution to tackle with this issue that requires automatically-obtained initialization, registration and site models. With least human intervention, the desired solutions can adapt to unexpected changes in camera viewpoints during operation.

3. Our third objective is to deal with other common issues in video surveillance. Low image quality, cluttered background, view invariant, illumination variations, and occlusion by furniture, need to be taken into consideration. Since we

[4]Since there is no clear difference in definition between actions and activities, hereinafter we use action and activity interchangeably.

aim at designing features based on foreground images, the above factors have great influences on the final detection results. The desired solutions can work well in arbitrary oblique viewpoints,[5] under typical indoor lighting conditions, and with partial and even severe occlusion caused by furniture. For instance, the elderly may fall to a sofa that makes people occluded either partially or severely from a certain viewpoint.

4. Fourthly, our proposed solutions target to single-user applications, for example, to support the elderly living alone at home. We argue that the application of fall detection methods seem to be useless in case of having more than one person in the monitored spaces. The falling-down action of one person should be known or detected by the other ones.

5. Our final objectives include designing low-cost solutions and protecting users' privacy. The desired solution must be low computational cost so that it can be run in real-time on a common PC. We aim at declining the cost of home telecare services so that most older people can approach it. Since cameras are placed at fixed positions in video surveillance, our proposed solutions are able to make use of foreground images to protect users' privacy.

1.3 Proposed approaches and contributions

In this dissertation, we introduce multiview 3D spatial features-based approaches to fall detection. We realize that falls are associated with fast body movements to change postures from upright to almost lengthened, followed by a sufficient duration of staying almost motionlessly on the ground or on some objects. This is contrary to slow manners of doing usual activities of the elderly. Therefore, we propose using 3D spatial features which are highly discriminative in distinguishing human states or postures. Given a sequence of human states, falls can be reliably inferred by analyzing human state transition. Figure 1-3 illustrates the insights of our fall detection pipeline,

[5] Overhead viewpoint is not considered because oblique viewpoints not only provide richer information of detecting falls but also monitor wider areas than overhead one

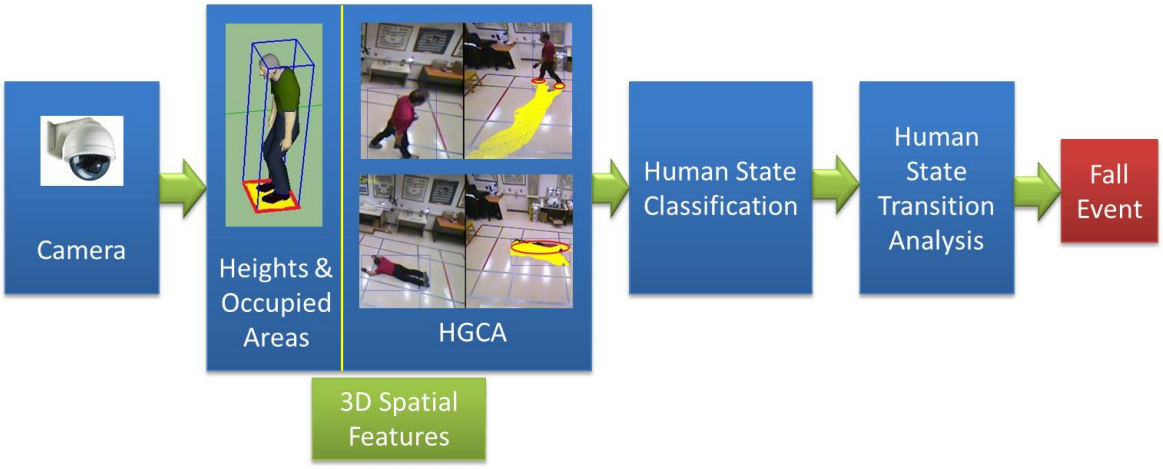


Figure 1-3: The insights of our fall detection pipeline. We propose using 3D spatial features, i.e., the combination of heights and occupied areas, and Human-Ground Contact Areas (HGCA). Heights and occupied areas are extracted from 3D cuboids of the person of interest. The bottom area of the 3D cuboid (the yellow area enclosed by a red rectangle) is defined as occupied area as shown in the left of the module *3D Spatial Features*. Meanwhile, HGCA is defined as the contact area between the person and the ground, a part of red ellipses, as depicted in the right of the module *3D Spatial Features*.

composing of three elements: 3D spatial feature extraction, human state classification, and human state transition analysis for fall inference.

The first two major contributions of this dissertation include our introduction of (1) the combination of heights and occupied areas, and (2) Human-Ground Contact Areas (HGCA) for fall detection. They are 3D spatial features which are highly discriminative in classifying human states and are efficiently estimated from multiple views. We present low-cost multiview schemes of estimating these 3D spatial features and also demonstrate empirically their good performance in fall detection.

Recently, Bag-of-Video-Word (BoVW) approaches to human action recognition have demonstrated good performance on a wide range of datasets. However, its performance in discriminating fall actions from other actions of daily living is unknown. To the best of our knowledge, there is no research work evaluating BoVW approaches on a dataset, exclusively dedicated to fall detection. Hence, the last contribution of this dissertation comes from our empirical study of accessing the effectiveness of BoVW approaches to fall detection, in comparison with our proposed as well as state-

of-the-art approaches.

1.3.1 Fall detection based on heights and occupied areas

Firstly, we describe in this dissertation the combination of heights and occupied areas, extracted from 3D cuboids of the person of interest as good 3D spatial features for human state classification. We realize that people in lying states occupy larger areas than people in standing and sitting states. Heights of people in standing states are greater than that of people in sitting and lying states. Intuitively, three typical human states, i.e., standing, sitting, and lying, fall into three separable regions of the proposed feature space, composing of heights and occupied areas. In order to reduce the computational complexity of approximating the person of interest by 3D cuboids, we configure two cameras whose fields of view are relatively orthogonal. Thus, 2D bounding boxes of the person extracted from two views serve as two orthographic projections of the 3D cuboids, making the 3D cuboid reconstruction straightforward. However, reconstructed 3D cuboids are not view-invariant across viewing windows due to the camera perspective. We propose using Local Empirical Templates (LET) that are originally proposed for counting people [Hung et al., 2010, 2012], to normalize the 3D cuboids. LET are defined as sizes of a standing person in local image patches. Two important characteristics of LET include (1) LET in unknown scenes can be easily extracted by an automatic manner, and (2) by its nature, LET hold the perspective information. Therefore, the reconstructed 3D cuboids become view-invariant across the viewing window after normalization by appropriate LET. Support vector machines are adopted to classify human states before inferring fall events by time-series analysis of human state transition.

1.3.2 Fall detection based on human-ground contact areas

By using height and occupied area in the first solution, we can distinguish lying from standing and sitting states. But the information of the lying positions such as on a sofa in resting states or on the ground after falls is unknown. Therefore, the above

solution can only detect a state transition from standing to lying as a fall. Sit-to-stand-transfer falling type in which the person changes from sitting to lying states is not considered. This kind of falling happens quite often when the person comes out of the resting states. Therefore, we propose another 3D spatial feature to overcome this limitation.

We argue that people always make a little contact with the ground during usual activities mainly by the feet but often lie completely on the ground after suffering from accidental falls. We come up with another good 3D spatial feature, so-called Human-Ground Contact Areas (HGCA). To measure HGCA, we project foregrounds of the person of interest from one view to another by using homography of the ground between two views. Overlap regions between projected foreground and the foreground in the latter view that only exist where people are in contact with the ground, due to the plane parallax, are measured as HGCA. We generalize a threshold of HGCA to separate lying states from standing and sitting states from view-invariant distributions of HGCA with respect to human states. We propose using human state simulation in which camera viewpoints are freely changed to capture 3D human models in various states. Hundreds of images are generated from the simulation as training data to build these distributions of HGCA. Finally, we perform temporal analysis of human state transition to claim falls. We test both our approaches on "multiple camera fall dataset" leading to competitive performance and lower computational cost with other methods tested on the same dataset [Rougier et al., 2007b, 2011b; Auvinet et al., 2011].

1.3.3 Bag-of-Video-Word approaches to fall detection

Recently, BoVW approaches to human action recognition have demonstrated good performance on a wide range of datasets which can be decomposed into two categories: heterogeneous and specific action datasets. Among heterogeneous action datasets, e.g., KTH [Schuldt et al., 2004], Weizmann [Blank et al., 2005], Hollywood 2 [Laptev et al., 2008], and UCF101 [Soomro et al., 2012], etc., only HMDB51 [Kuehne et al., 2011] dataset contains fall actions and few actions of daily living. In addition, the

subjects in the dataset are mostly young people rather than the elderly. Meanwhile, fall actions are omitted in some specific action datasets, exclusively dedicated to actions of daily living, i.e., URADL [Messing et al., 2009], TUM Kitchen [Tenorth et al., 2009], and MPII Cooking [Rohrbach et al., 2012] datasets. It means that fall actions are treated separately from the other actions of daily living in the context of dataset creation. It is subjective to draw similar good performance of BoVW approaches to fall detection based on its results on these datasets reported in the literature. To the best of our knowledge, there is no research work evaluating BoVW approaches to fall detection on a dataset, exclusively dedicated to fall detection. That is, the dataset must contain both fall actions and a variety of other actions of daily living, such as sit up, stand up, sit down, lie down, walk, carry objects, do housework, take off cloth, and put on cloth, etc. Hence, in this dissertation, we carry out an empirical study of evaluating the effectiveness of BoVW approaches to fall detection on "multiple cameras fall dataset." We use similar evaluation protocol, proposed in [Wang et al., 2009], with STIP and HOG/HOF descriptors [Laptev, 2005] and nonlinear Chi-Square kernel SVM classifier. In comparison with our proposed and state-of-the-art methods, tested on the same dataset, BoVW approach produces comparable accuracy recognition but are more computationally expensive.

1.4 Dissertation organization

The rest of our dissertation is structured as the following. We will provide the background on accidental falls and a review of vision-based fall detection methods in chapter 2. In chapter 3, we describe the combination of heights and occupied areas for fall detection. Chapter 4 presents the method of fall detection based on human-ground contact areas. The empirical study of accessing the effectiveness of BoVW approaches to fall detection is carried out in chapter 5. Finally, chapter 6 concludes the dissertation and delineate future directions to fall detection.

Chapter 2

Background on fall detection

In this chapter, we investigate the definition and characteristics of a fall. Once they are clearly defined, it helps understand the existing algorithms and aids to design novel effective ones [Yu, 2008]. Subsequently, we describe the benchmark dataset and performance evaluation criteria used in our work. The summarization of related works to provide the context of our proposed approaches comes in the last section of this chapter.

2.1 Fall definition

The fall of a person can be defined as a rapid change from upright/sitting postures to almost lengthened/lying ones [Noury et al., 2008] subsequently followed by a relatively immobile duration caused by shock impacts of the fall. According to this definition, a fall can be broken into four phases, that is, prefall phase, critical phase, post-fall phase and recovery phase [Noury et al., 2008].

In the *prefall phase*, the elderly perform daily activities usually in slow manners, characterized by slow motion. However, some sudden and quick movements that may sporadically happen by confounding activities such as sitting down brutally.

The human body starts to fall or in other words, to change postures suddenly and quickly during the *critical phase*. This phase ends when the human body hits the ground or some objects and usually lasts between 300 and 500 ms.

In the *post-fall phase*, people stay relatively immobile on the ground in lying states. The timing of this phase is much longer than the critical phase, depending on the physical pain and shock impacts of the fall. Subsequently, people try to stand up either by their own efforts or by the supports of other people during the *recovery phase*.

Some people argue that the post-fall phase is not necessary to exist or only happen in a very short time. It means that fallen people experience directly to the recovery phase. Such falls do not threaten seriously fallen people's lives then are not considered in our work.

Although fall events and confounding events have sudden and quick movements in common, characteristics of a fall after the critical phase are quite unique. This fall definition is critical in discriminating fall from usual events, particularly the confounding ones. Various methods take this definition into consideration to reliably infer fall events, as shown in our summarization of related works in section 2.3. Our proposed approaches also rely on this definition for fall inference.

2.2 Benchmark dataset and evaluation criteria

An increasing number of publications have been dedicating to fall detection by using not only vision sensors but also the other kinds of sensors. It is an urgent problem of how to evaluate fairly their performances. In computer vision, it is recommended to use common and publicly available datasets for benchmarking developed algorithms because of two major merits. Firstly, it saves time and resources for collecting new samples for experiments [Chaquet et al., 2013]. Finally and more importantly, it facilitates fair comparisons of different approaches and provides in-depth understanding of (in)abilities of the different approaches [Poppe, 2010]. Therefore in this work, we use a common and public benchmark dataset, namely, "multiple camera fall dataset", recently released by Université de Montreal [Auvinet et al., 2010] in experiments. Furthermore in the field of fall detection, we also need to evaluate fairly between vision-based methods and wearable device-based methods as well as ambient

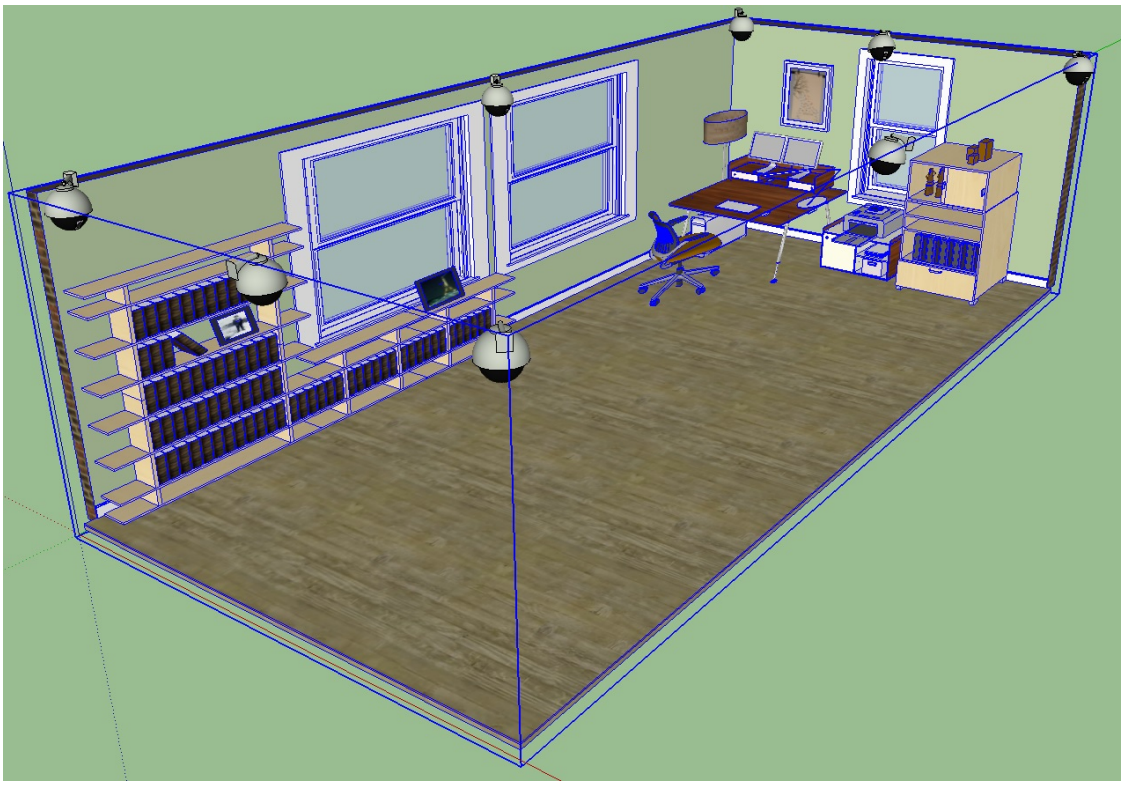


Figure 2-1: The Layout of camera network in "multiple camera fall dataset"

device-based methods. [Noury et al., 2007] proposed a common performance evaluation criteria to address this urgent need. In this section, we describe the common benchmark dataset and the performance evaluation criteria used in our work.

2.2.1 Multiple camera fall dataset

In regard to making a publicly available dataset for the scientific community, real fall situations seem to be inapplicable due to the issue of privacy protection. More importantly, gathering real falls of the real elderly in home environments is a daunting task. Rather than that, simulated falls performed by a young actor, instead of the real elderly, are favorable and are commonly adopted in the literature. However, they did not make their own collected samples publicly, see Table 2.2. Consequently, the performances of their approaches cannot be accessed exactly and fairly because the challenges of their dataset are not available. Recently, a research group at Université de Montreal designed a multiview benchmark dataset and made it publicly for research

purposes [Auvinet et al., 2010]. To the best of our knowledge, [Auvinet et al., 2010] and another one in [Shoaib et al., 2011b] are the only two publicly available dataset for vision-based fall detection methods up to date. However, we only use the dataset in [Auvinet et al., 2010] in experiments since the one in [Shoaib et al., 2011b] is a single-view dataset.

In this dataset, simulated falls are performed by a young actor who is an experienced clinician in the area of elderly care. Their laboratory was rearranged to look like a home environment with sofa, chair, and table, etc. The actor performed not only various kinds of falls but also other usual activities, some of which are considered as confounding events or like-fall events. All activities were recorded simultaneously by eight inexpensive IP cameras with lenses to capture the whole room as widely as possible. Figure 2-1 shows the layout of eight-camera arrangement. They provide in the dataset not only video samples, but also camera calibration data, information of camera network synchronization, and more importantly, event annotation, etc. [Auvinet et al., 2010]. In the following, we summarize the challenges of the dataset.

Challenges of the dataset

The dataset was designed to include 24 scenarios, twenty two of which are short video samples. The short scenarios aim at demonstrating various kinds of falls, for example, falling forward, falling backward, falling due to balance loss, falling during stand-to-sit transfers, falling during sit-to-stand transfer, and falling to furniture. On the contrary, the two last scenarios which are several-minute length concern more about usual activities, especially the confounding events such as lying down on sofa, sitting down brutally, crouching on the ground, doing housework, carrying objects, rearranging furniture, and putting on and taking off a coat, etc., although some confounding events are included in some short scenarios. All scenarios are summed up in Table 2.1 along with their challenges. Some frame examples of the dataset are demonstrated in Fig. 2-2. In total, the dataset contains 24 fall events[1] and 24

[1]Although in the dataset annotation [Auvinet et al., 2010], there are 25 falls but one happened on the recovery phase. It is argued that if the first fall had been detected then it was not necessary to detect the second one which happened in the recovery phase.

confounding events (11 crouching events, 9 sitting events, and 4 lying events)

The other challenges include high video compression (MPEG4), cluttered and texture background, shadow and reflection, illumination variations and occlusion by furniture. The dataset was shot under typical indoor lighting conditions that are with ambient lights and without sudden changes of illumination. Reflection on the ground may pose challenges for extracting silhouettes by background subtraction algorithms. Occlusion by furniture also happens very often that may deteriorate the quality of extracted features. Although there is only one actor performed everything in the dataset, his appearance changes between scenarios and even within a scenario by putting on and taking off a coat.

Table 2.1: The summation of all scenarios in the dataset along with their challenges. We take the viewpoint of camera 2 to make this table. Therefore, the table claiming no occlusion from this viewpoint does not mean no occlusion from other viewpoints.

No.	Types of falls	Confounding events	Occlusion	Mattress usage
1	Falling backward	Putting a coat on	None	Yes
2	Loss of balance	None	None	Yes
3	Falling forward	None	None	Yes
4	Loss of balance Falling when recovery	None	None	Yes No
5	Falling forward	None	None	Yes
6	Falling backward	None	None	Yes
7	Loss of balance	None	None	Yes
8	Loss of balance	None	None	Yes
9	Sit-to-stand transfer	Sitting down	None	Yes
10	Stand-to-sit transfer	None	None	Yes
11	Loss of balance	None	None	Yes
12	Loss of balance	None	None	Yes
13	Sit-to-Stand transfer	Sitting down brutally Lying down on sofa	None	Yes

Continued on next page

Table 2.1 – *Continued from previous page*

No.	Types of falls	Confounding events	Occlusion	Mattress usage
		Lying to Sitting		
14	Sit-to-Stand transfer	Sitting down brutally Lying down on sofa Lying to Sitting	None	Yes
15	Loss of balance	Sitting down brutally Standing up	None	No
16	Falling to table	Kneeling Crouching Sitting down brutally Standing up	None	No
17	Falling to table	Kneeling Crouching Sitting down brutally Standing up	None	No
18	Falling forward	Sitting down Standing	None	No
19	Falling to sofa	None	Yes	No
20	Falling to sofa	None	None	No
21	Sit-to-stand transfer	Sitting down	None	No
22	Sit-to-stand transfer	Sitting down	Yes	No
	Falling to a chair	taking a coat off Crouching Carrying objects Putting objects down Standing up Crouching Carrying objects	Yes	No

Continued on next page

Table 2.1 – *Continued from previous page*

No.	Types of falls	Confounding events	Occlusion	Mattress usage
23	Sit-to-Stand transfer	Displacing objects Sitting down brutally Lying down Standing up Sitting down brutally Lying down Standing up Crouching Sitting down Standing up Carrying objects		
24		Displacing a chair Doing housework Kneeling Dropping sweeper Kneeling Kneeling Crouching Carrying a coat Dropping the coat Bending to take coat Kneeling Dropping sweeper	None	No

(a) Falling forward (b) Falling backward (c) Loss of balance

(d) Falling to sofa (e) Falling to a table (f) Falling during stand-to-sit transfer

(g) Falling during Sit-To-Stand Transfer (h) Sitting down brutally (i) Crouching on the ground

(j) Sitting on a sofa (k) Lying on a sofa (l) Kneeling on the ground

(m) Carrying an object (n) Doing housework (o) Take off a coat

Figure 2-2: Frame examples in "multiple camera fall dataset"

(a) Standing (b) Sitting (c) Falling

Figure 2-3: Image samples of our In-House video sample

2.2.2 Our In-House sample

We also record a video sample containing some actions of daily living, i.e., walking and sitting, as well as falling on the ground in a classroom of our campus. The simulated actions were recorded simultaneously by two cameras [2]. These two views are obliquely configured relatively orthogonal for testing both of our proposed approaches in this dissertation. The video sample is two and a half minutes in length, in which the person sits on a chair and falls on the ground twice (for each one). The test is to further confirm the validity of our approaches since the camera settings and environments are different from that in the dataset. Some image samples of our in-house video sample are shown in Fig. 2-3.

2.2.3 Performance evaluation criteria

Since the results of fall detection methods are binary such as "detected" or "not detected", regardless of types of sensors used, Noury *et al.* [Noury et al., 2007, 2008] proposed the following performance evaluation criteria.

There are 4 possible situations.

1. True Positive (TP): the number of falls correctly detected, that is, after a fall had happened, the methods detected it successfully.

2. False Positive (FP): the number of normal activities detected as falls, that is,

[2] one is Nikon camera P500 and the other is Canon HD recorder

the methods claimed a fall but it had not occurred.

3. True Negative (TN): the number of normal activities not detected as falls, that is, an usual activity occurred and the methods did not produce any fall event.

4. False Negative (FN): the number of falls not detected correctly, that is, a fall had happened but the methods did not claim it as a fall.

We compute 2 following criteria to evaluate the response of these 4 situations.

1. Sensitivity

$$Se = \frac{TP}{TP + FN} \tag{2.1}$$

2. Specificity

$$Sp = \frac{TN}{TN + FP} \tag{2.2}$$

High sensitivity means that most fall incidents are correctly detected. Similarly, high specificity implies that most normal activities are not detected as falls. A good fall detection methods must achieve high values of both sensitivity and specificity.

In implementation, we use the event annotation, provided in the dataset. In each video sequence, the actual time occurring a fall (denoted by t_{fall}) is manually annotated. This time is defined by the first moment of human body hitting the ground after a fall. A fall is detected after t_{fall}, resulting in TP. A fall is not detected after t_{fall}, resulting in FN [Auvinet et al., 2010].

2.3 Related works

In this section, existing fall detection methods are reviewed to delineate the context of our proposed approaches. However, we only take vision-based methods into consideration. Making a survey of wearable device-based methods and ambient device-based

methods goes beyond the scope of our work. In this regard, we refer to some recent comprehensive reviews of fall detectors [Yu, 2008; Ward et al., 2012; Mubashir et al., 2013; Spasova and Iliev, 2014].

According to the common pipeline of fall detection methods in Fig. 1-2, feature extraction, feature classification (or human states/postures classification) and fall inference are the most important parts. Therefore for each method reviewed in this section, we highlight the extracted features along with their discriminative powers and how to classify them and make decisions. We classify fall detection methods into three categories based on the number of cameras and types of camera employed. The first two categories using regular RGB cameras are broken into single-view and multiple-view approaches. The last category uses RGB-Depth cameras among which Kinect is the most prevalent. It is noted that all cameras must be stationary.

2.3.1 Single-view approaches

In this section, we focus on monocular vision-based methods of fall detection. A common class of approaches are based on dimension variations of 2D human body silhouettes. The early work of [Anderson et al., 2006] analyzed the sizes of silhouettes. The width-to-height ratios or aspect ratios of humans and off-diagonal term from covariance matrix are taken as adequate features for training Hidden Markov Models (HMMs) to recognize falls. The aspect ratios of humans in standing and lying states are large and small, respectively. However, this observation may not be true in consideration of human body upper limb activities. To eliminate this effect, [Liu et al., 2010] used a statistical scheme to remove peaks in vertical histograms of silhouette images. They proposed k-Nearest Neighbor (kNN)-based posture classifier working with a feature space composing of the aspect ratios and the difference between height and width of silhouettes. By taking the different pace between falls and like-fall actions into account, critical time difference is used to claim a posture change as a fall. [Huang et al., 2008] introduced the combination of aspect ratio and silhouette size variations to discriminate fall and non-fall states by a linear classifier. Surrounding and personal information such as weight, height, health history, being in toilet, and

being in dining room, etc. are also integrated into the linear classifier by modifying its weights. Both three methods do not take the occlusion by furniture into consideration. They merely reported the experiments with cameras placed sideways since the observations of their designed features based on size variations of 2D silhouettes only make sense in this camera setting. In practice of indoor surveillance, the cameras are preferred to be in oblique settings for wider views and occlusion avoidance.

Several works employ context information for aiding fall inference by dividing the environment into activity and inactivity zones. [Lee and Mihailidis, 2005] labeled furniture areas in the images captured from a top-view camera as inactivity zones, i.e., chair, sofa, and bed, etc. The system treats lying in activity zones as a fall but consider ones in inactivity zones as acceptable. Position and speed of centroids, perimeter and Feret diameter of blobs are extracted and thresholded to distinguish standing and lying postures. In the similar experimental scenarios, [Charif and McKenna, 2004] argued that there are a few places in a room in which people are relatively inactive most of the time for relaxing activities, i.e., watching television, reading newspaper, and drinking tea, etc. They are tracked and checked whether they are inactive in a known inactivity zone. Their immobility outside known inactivity zones is likely caused by fall occurrences. However, both methods expose several limitations. Firstly, the speed estimation of 2D silhouettes is highly sensitive to cluttered background and a variety of human daily activities. Secondly, their adaptation to environment changes is poor. Finally, using top-view cameras seems to be inappropriate for the problem of fall detection since crucial clues from the vertical motion of human body to recognize a fall are not available. Recently, by using an obliquely-placed camera in a real home environment, [Shoaib et al., 2010, 2011a] presented a context model to learn the head and floor planes as well as (in)activity zones from the foregrounds of moving person in an unsupervised manner. Distance measures between detected heads and reference heads, provided by the context model, are adopted as a discriminative feature to distinguish walking and sitting from lying actions by thresholding. However, confounding events such as crouching down on the ground and lying down on the ground are classified as falls since these actions take place in activity zones, rather than in

inactivity ones, i.e. on a sofa or a bed, etc. Moreover, a major disadvantage of these approaches is that fall occurrences in inactivity zones are not taken into account.

Since falls are associated with quick movements of human body, motion features are employed to detect large motion events, before extracting other features for fall inference. Motion History Image (MHI) is adopted to quantify the motion of human body [Rougier et al., 2007a] to detect large-motion events. Upon such events, silhouettes are approximated by ellipse models whose orientation angle and ratios of major to minor axes, so-called axis ratio, are extracted as features. Hand-designed thresholds are applied to these features to detect posture changes. Falls are confirmed if the posture change is followed by a sufficient motionless duration. It is reported to run in 10 fps with video stream's resolution of 320x240. Similarly, [Liao et al., 2012] introduce Integrated SpatioTemporal Energy map (ISTE) to quantize the body motion. ISTE is demonstrated to deal with low frame-rate video streams better than MHI does. They quantified the body motion into No Motion, Regular Motion, and Large Motion which are combined with features of orientation angle, displacement, and axis ratio in Bayesian Belief Network to reason not only fall but also slip-only events. [Chen et al., 2010] present an ingenious combination of distance map of two sampling human skeletons and variation analysis of ellipse human models for fall inference. The suspicious incidents are further verified by checking inactive states of the person in a period of time.

Human shape is studied for fall detection since it is believed that human shape changes progressively during usual activities but drastically during falls. Log-polar histograms are used to represent silhouettes' shape [Rougier et al., 2007b, 2011b]. Full Procrustes distance [Rougier et al., 2007b] and mean matching cost [Rougier et al., 2011b] are extracted as good features for fall detection. In principle, both full Procrustes distance and mean matching cost should be high during falls and low right after that. In their preliminary work [Rougier et al., 2007b], falls are detected by thresholding the full Procrustes distance and checking the inactive states in a period. In their extended work [Rougier et al., 2011b], GMM is employed to detect falls independently in four views and the detection results are fused to enhance the

performance. This method is reported to work with the frame rate of 5 fps due to the expense of high computational cost. [Htike et al., 2011] utilize chord distribution histograms of silhouettes' shape as view-invariant representation of various 2D poses. The proposed method performs human pose recognition then detects falls by fuzzy HMM. Real-time performance is claimed in the paper but without giving specific frame rate. [Khan and Sohn, 2011] describe silhouettes' shape by R-Transform to perform posture recognition by Kernel Discriminant Analysis (KDA) and event detection by HMM. However, their experiments are setup with sideway cameras. [Yu et al., 2012] perform posture classification by a directed acyclic graph SVM working with orientation angle, ratio of ellipse axes, and projection histograms along the axes of the ellipse as input features. Lying or bending on the ground regions in a period are considered as falls.

To take advantages of 3D information in detecting falls, a promising solution is to make use of single calibrated cameras. [Rougier et al., 2006] describe a particle filters-based method of tracking 3D head, with manual initialization, to extract head velocity for fall inference. Their solution is sensitive to confounding actions like sitting down brutally. [Cucchiara et al., 2005] trained probabilistic projection maps for posture classification, i.e., standing, sitting, lying, and crouching. They suggest using a tracking algorithm to handle occlusion with a state transition graph for reliable event classification results. Their later work [Cucchiara et al., 2007] extends [Cucchiara et al., 2005] to cope with multiple rooms by using multiple cameras whose fields of view are partially overlapped. Camera hand-off to identify the same person across various rooms is treated by warping human appearance between views based on homography. The warping also helps alleviate the problem of occlusion by furniture. A HMM is trained for obtaining more robust recognition results. Although multiple cameras are used, the final decision is made independently by only one camera which is observing the person.

In sum, 2D spatial features extracted from a single camera, seem to be insufficiently discriminative to distinguish fall from usual events. Falls in parallel to the optical axis of cameras also pose difficulty for single-view approaches. Moreover,

several assumptions about camera viewpoints, occlusion, (in)activity zones and high computational cost, seem to make single-view approaches unfeasible. Meanwhile, 3D spatial features combined with temporal structures of actions seem to be more discriminative than 2D spatial ones to discern falls. In the next section, we review multiview approaches that are capable of extracting 3D spatial features.

2.3.2 Multivew Approaches

In this section, we consider methods of using more than one camera whose fields of view are partially overlapped. [Thome et al., 2008] applied the metric image rectification to derive the 3D angle between vertical line and principal axis of human ellipse models. Decisions made independently by multiple cameras are fused in a fuzzy context to classify human postures. Layer HMM is hand designed to make event inference. Two uncalibrated and perpendicular cameras are set up in [Hazelhoff et al., 2008]. Principal Component Analysis (PCA) is applied to determine the direction of main axis of the human body and ratios of variances in x and y directions. Fall events are inferred by a multi-frame Gaussian classifier and verified by head tracking based on skin-color information. Its reported frame rates are about 15 fps and 5 fps with 320x240 and 640x480 video resolutions, respectively.

[Anderson et al., 2009] introduced a framework of fall detection in the light of constructing voxel person. A hierarchy of fuzzy logic is proposed in this research for human state classification (the first level) and for event inference (the second level). The linguistic aspect of fuzzy logic makes this framework flexible, allowing for user customization based on their knowledge of cognition and physical ability. The two studies of [Zambanini et al., 2010] and [Zweng et al., 2010] are inspired by [Anderson et al., 2009]. [Zambanini et al., 2010] emphasize on the low-cost computation by employing low-cost features, i.e., aspect ratios, orientation, axis ratios and motion speed extracted from the voxel space. Despite using same fuzzy logic-based posture estimation in [Anderson et al., 2009], a less sophisticated reasoning mechanism based on computing fall confident values is adopted to realize the real-time performance but without giving specific frame rate. [Zweng et al., 2010] also utilize same features,

fuzzy logic-based posture estimation, and fall confident value computation with [Zambanini et al., 2010]. However, the features are extracted directly from 2D images, in turn, not view-invariant. Therefore, decisions made independently from each view are fused by a statistical behavior model, so-called accumulated hit-map. By the same methodology of reconstructing voxel person, [Yu et al., 2011] also compute the differences of centroid positions and orientation of a voxel person to classify falls against other normal events by using one class SVM. [Auvinet et al., 2011, 2008] discussed a method of reconstructing 3D human shape from a network of cameras. They proposed the idea of Vertical Volume Distribution Ratio since the volumes of standing and lying-down people are vertically distributed significantly differently. The method is able to handle occlusion since the 3D reconstructed human shape is contributed from multiple cameras.

In summary, on the one hand, the main advantage of multiview approaches to fall detection is the capability of extracting 3D spatial features, i.e., voxel person, and 3D silhouettes, etc. They are highly discriminative in classifying postures or states, in turn, leading to better fall detection results. In addition, features obtained from multiple views are more reliable than those obtained from single views due to occlusion that is frequently happened in indoor environments. People may be occluded in one view but are likely visible in other ones. In other words, multiview approaches can deal with the problem of occlusion better than single-view approaches. On the other hand, multiview approaches exposes several limitations. Firstly, adding cameras make the methods more complex and require more computing resources. However, the evolution of computing devices such as GPU may ease this difficulty of multiview approaches. Secondly, adding more cameras make the monitored window narrower. Otherwise, lens must be used to capture the home environment as widely as possible, leading to highly distorted images that also pose some difficulty. Thirdly, multiview approaches must require performing synchronization, calibration and registration. As we mentioned in section. 1.2, maintaining the calibration and registration of camera networks during operation is a daunting task [Javed and Shah, 2008]. The performance of multiview approaches would be deteriorated in the case of changing

environments and changing viewpoint unexpectedly without doing calibration and registration again.

2.3.3 RGB-Depth camera based approaches

In contrast to multiview approaches in which 3D information is computed from color images captured from multiple cameras, RGB-Depth cameras are able to offer both color images and the depth map of the scenes. In addition, the introduction of the most prevalent and low-cost depth cameras, Microsoft Kinect, leads to a promising solution to fall detection. In this section, we discuss some recent methods, employing RGB-Depth cameras.

[Rougier et al., 2011a] present a method of efficiently extracting human centroid height relative to the ground from the depth map sequences, provided by a Kinect. They argue that most falls end on the ground or near the ground. They also propose using the 3D body velocity computed right before the body occluded by the furniture in order to minimize the miss detection in the case of occurring occlusion by furniture during either critical phase or post-fall phase. Fall events are detected by thresholding these features by hand-designed thresholds. [Planinc and Kampel, 2012a] estimate the major orientation of the human body in 3D space by using the skeleton. A fall is claimed if the major orientation of the person is parallel to the ground and the height of the spine is near the ground floor. Apparently, confounding events are not taken into consideration. In their later work [Planinc and Kampel, 2012b], same features are used but the final decision is refined by using fuzzy logic like in [Anderson et al., 2009; Zambanini et al., 2010; Zweng et al., 2010]. The work of [Zhang et al., 2012a] selects 8 tracked body joints on head and torso by using Microsoft Kinect SDK for calculating kinematic features and human height. They design a hierarchy SVM classifier to recognize 5 activities, i.e., fall from chair, fall from standing, stand on the ground, sit on chair, and sit on floor. In [Zhang et al., 2012b], five features i.e., duration, total head drop, maximum speed, smallest head height, and fraction of frames where head drops, are determined from the depth map sequences and combined in a Bayesian framework to make event decision. [Mastorakis and Makris, 2012]

determine a fall according to velocity, measured by the contraction or expansion of the 3D bounding box. By explicitly using 3D bounding box, this method requires no prior knowledge about environments, for example, the ground floor in all above methods [Rougier et al., 2011a; Planinc and Kampel, 2012a,b; Zhang et al., 2012a,b]. Three Dimensional Motion History Images (3D-MHI) are proposed to compute Hu moments for fall classification and confirmation by using SVM [Dubey et al., 2012].

In conclusion, Kinect has several advantages over RGB cameras. Firstly, Kinect is able to work with and without ambient lights (at night) and is also insensitive to both steady and sudden illumination changes. Secondly, it is capable of operating in real time with 30 fps [Planinc and Kampel, 2012a; Mastorakis and Makris, 2012]. Finally and more importantly, it provides fast 3D information.

However, Kinect also exhibits some limitations. Firstly, it is sensitive to external infrared source, i.e., the sun light. Secondly, the simultaneous use of multiple Kinects to deal with occlusion is quite difficult. The depth map is degraded when the fields of view of these Kinects are partially overlapped. It is because the structured light sources of the Kinects interfere each other to produce so-called *crosstalk* phenomenon. Recently by using external mechanical mechanism to vibrate structured light of Kinects, good quality of depth map is attained without compromising the frame rate [Butler et al., 2012]. Finally, Kinect is unable to sense depth information beyond 4 meters.

Taking the trade-off into consideration, we prefer to multiview approaches since RGB cameras can keep a weather eye on larger areas than Kinect does. Modern computing devices also help realize multiview approaches in real-time. Adaptive background subtraction techniques [Stauffer and Grimson, 2000; KaewTraKuPong and Bowden, 2001; Zivkovic, 2004] can handle well steady illumination changes. We argue that sudden illumination changes by changing the ambient light only happen in few frames and do not affect seriously the fall detection performance. Moving cast shadow can be removed effectively by using moving cast shadow removal techniques like in [Cucchiara et al., 2001].

2.3.4 The context of our proposed approaches

Our proposed approaches make use of 3D spatial features, i.e., the combination of heights and occupied areas, extracted from 3D cuboids [Hung and Saito, 2012, 2013] and HGCA [Hung et al., 2013] to classify human states into standing, sitting, and lying. Falls are inferred by analyzing human state transitions. Although the idea of approximating the person of interest by 3D cuboids or 3D bounding boxes in our work [Hung and Saito, 2012, 2013] is similar in [Mastorakis and Makris, 2012], our work and theirs were proposed concurrently. In addition, we are considering different features, extracted from different camera settings, that is, two cameras whose fields of view are relatively orthogonal, rather than with Kinect. In this section, we analyze qualitatively our approaches against the backdrop of above related works.

Firstly, our approaches work well with oblique camera settings which are favorable among vision-based fall detection techniques. Secondly, since 3D spatial features are highly discriminative in classifying human states, our approaches can distinguish falls in any direction, particularly the ones in parallel to cameras' optical axis. Falls and confounding events can also be recognized effectively. For example, although a fall and a sit down brutally have fast body movements in common, people take more areas after falling than after sitting down. Although a fall and a crouch on the ground lead to high values of 3D-area-based features, the fall occurs quickly in contrast to a slow manner of a crouch by the elderly. Moreover, the human state classification is performed in our work by using machine learning techniques, rather than by hand-designed thresholds in [Lee and Mihailidis, 2005; Anderson et al., 2006; Huang et al., 2008; Charif and McKenna, 2004; Rougier et al., 2011a].

Thirdly, in contrast to methods using context information [Lee and Mihailidis, 2005; Charif and McKenna, 2004; Shoaib et al., 2010, 2011a], we do not make any assumption of environments. That is we take falls in inactivity zones into consideration since such falls happen very often with the elderly. They are prone to suffering from dizziness and syncope when changing states from resting to active, and vice versa.

Fourthly, we realize that to recognize a fall, it is not necessary to reconstruct

voxel person [Anderson et al., 2009] or 3D silhouettes in 3D world [Auvinet et al., 2011] which are expensive computation. Good measures of our 3D spatial features are capable of producing competitive performance and requiring lower computational cost.

Finally, our approaches are able to adapt well with the environment changes since we do not use camera calibration, manual registration or complicated site models. As we mentioned in Sect. 1.2, maintaining calibration and registration, etc., of a camera network during operation is a daunting task. But this task is a must otherwise the performance will be deteriorated. In [Hung and Saito, 2012, 2013], we employ LET in feature normalization but LET can be automatically initialized and updated during operation [Hung et al., 2010, 2012]. The LET update procedure helps our approach [Hung and Saito, 2012, 2013] adapt to unexpected and sudden environment changes. If camera viewpoints are changed unexpectedly during operation, the obtained foregrounds will cover major parts of the viewing window. Detecting such case is straightforward for re-running the LET initialization. In our later work [Hung et al., 2013], we only need to calibrate homography of the ground between views by using the four-point algorithm [Hartley and Zisserman, 2004]. That is, we need four landmark points on the ground which are visible in all views. One solution is to design a projector or colorful light emitting sources to create virtual landmark points on the ground for calibration when required. The homography re-calibration can be done with least human intervention. However in our implementation, four-point correspondences are selected manually.

In summary, we recap all reviewed methods and ours on Table 2.2, in terms of features, classification, event inference, background subtraction algorithm, capability of dealing with challenges, real-time performance, datasets and accuracy performance. The table is adapted from their reported results.

In the next chapters, we will describe our proposed approaches.

Table 2.2: The summarization of all reviewed methods and ours in terms of used features, classification, event inference, background subtraction algorithm, capability of dealing with challenges i.e., sensitive events, viewpoints, human movements, occlusion, lighting, real-time performance, datasets and accuracy performance. This table is adapted from their reported results. Some notations used in this table include OH - Occlusion Handling, SD - Self-collected Dataset, F - Fall, NF - Non-Fall, RT - Real-Time, L1RO - Leave 1 Record Out, SE - SEnsitivity, SP - SPecificity.

Method	Features	Classifier	Fall inference	No. of cam	Background Subtraction	Sensitive events	View-points	Sensitive human movement	Occlusion	Lighting	Real-time	Data-set	Performance (SE & SP)
[Anderson et al., 2006]	Aspect ratio and off-diagonal term of covariance matrix		HMM	1	[Chen et al., 2006]	Parallel to optical axis	sideways	Upper limb activity	No OH	Shadow removal	Not reported	SD	Not reported
[Liu et al., 2010]	Aspect ratio	kNN	Critical time difference	1	Inter-frame Difference	Parallel to optical axis	Sideways		No OH	Without shadow removal	Not reported	SD (45F, 45NF)	82 % & 87 %
[Huang et al., 2008]	Aspect ratio, silhouette size variation	Linear classifier	Linear classifier	1	[Wren et al., 1997]	Parallel to optical axis, sit down brutally	Sideways	Upper limb activity	No OH	Without shadow removal	Not reported	SD (20F, 80NF)	90% & 96%
[Lee and Mihailidis, 2005]	Position, 2D speed of Centroid, Perimeter, Feret Diameter of Blobs	Threshold	Lying in activity zones is caused by falls	1	[Wren et al., 1997]	Falls in inactivity zones	Not consider	Overhead	No OH	Without shadow removal	Not reported	SD (126 F 189 NF)	77 % & 95 %
[Charif and McKenna, 2004]	2D speed of Centroid	Threshold	inactive in activity zone is caused by fall	1	[McKenna et al., 2000]	Falls in inactivity zones	Overhead		No OH	Shadow removal	Not reported		

Continued on next page

Table 2.2 – *Continued from previous page*

Method	Features	Classifier	Fall inference	No. of cam	Background Subtraction	Sensitive events	View-points	Sensitive human movement	Occlusion	Lighting	Real-time	Data-set	Performance (SE & SP)
[Shoaib et al., 2010]	Head distance	Threshold	Evident accumulation mechanism	1	[Stauffer and Grimson, 2000]	Falls in inactivity zones, Crouch, Lie on the ground	Oblique		No OH	Shadow removal	Not reported	SD	SE = 96 %
[Rougier et al., 2007a]	MHI, orientation, axis ratio		Lack of motion after falls	1	[Kim et al., 2005]	Sit down brutally	Oblique		No OH	Shadow removal	10 fps (320x240)	SD (17F, 24NF)	88% & 88%
[Liao et al., 2012]	ISTE, orientation, displacement, axis ratio		Bayesian belief network	1	[Kim et al., 2005]	Sit down brutally	Oblique		No OH	Shadow removal	Not reported	SD	95% & 87%
[Chen et al., 2010]	Skeleton, orientation, axis ratio	Threshold	Lack of motion after falls	1	Not reported	Sit down brutally	Not reported		OH by tracking	Not reported	Without shadow removal	SD (22F, 32NF)	91% & 94%
[Rougier et al., 2007b]	Full Procrustes distance	Threshold	Threshold	1	[Kim et al., 2005]		Oblique		No OH	Without shadow removal	5 fps	[Auvinet et al., 2010]	95% & 96%
[Htike et al., 2011]	Chord distribution histogram of silhouette		Fuzzy HMM	1	[Haritaoglu et al., 2000]		Oblique	Carry object	No OH	Without shadow removal	RT	[Zambanini et al., 2010]	100% & 86% (L1RO)
[Khan and Sohn, 2011]	R Transform	KDA	HMM	1	[Elgammal et al., 2000]		Sideways	Carry object	No OH	Without Shadow	Not reported	SD	Not reported
[Yu et al., 2012]	Orientation, axis ratio, projection histogram	DAG-SVM	Lie or Bend on the ground are falls	1	[Kim et al., 2005]	Sit, crouch on the ground	Oblique		No OH	Handling sudden change	Not reported	SD (240F, 240NF)	97% & 99%
[Rougier et al., 2006]	Head 3D velocity		Threshold	1	Not use	Sit down brutally	Oblique		No occlusion	Not affected	Not reported	SD (9F, 10NF)	67% & 90%
[Cucchiara et al., 2005]	Probabilistic projection maps	Visual object based-classifier	State transition analysis	1	[Cucchiara et al., 2002]		Oblique	Carry object	OH by tracking	With shadow removal	15 fps	SD	SE = 100%

Continued on next page

Table 2.2 – *Continued from previous page*

Method	Features	Classifier	Fall inference	No. of cam	Background Subtraction	Sensitive events	View-points	Sensitive human movement	Occlusion	Lighting	Real-time	Data-set	Performance (SE & SP)
[Thome et al., 2008]	3D orientation angle	Threshold	Layered HMM	2	[Stauffer and Grimson, 2000]		Oblique		OH by tracking	Without shadow	27.8 fps (320x24)	SD (50F, 50NF)	98% & 100%
[Hazelhoff et al., 2008]	Orientation, ratio of variances in x, y		Multi-frame Gaussian classifier, head-based posfall verification	2	Not reported		Oblique	Carry object	No OH	Without shadow removal	15 fps (320x240)	SD	SE = 85 %
[Anderson et al., 2009]	Voxel person	Fuzzy logic	Fuzzy logic	2	[Luke et al., 2008]		Oblique		No OH	Shadow removal	Not real-time	SD (14F, 32NF)	100% & 94%
[Zambanini et al., 2010]	Aspect ratio, orientation, axis ratio	Fuzzy logic	Fall confidence	4	Not reported	Sit on a chair	Oblique		OH by multi-view	Without shadow removal	Real-time	SD (43F, 30 NF)	95% & 80%
[Zweng et al., 2010]	Aspect ratio, orientation, axis ratio	Fuzzy logic	Fall confidence, accumulated hitmap	4	[Wren et al., 1997]		Oblique	Stationary in a long time	OH by multi-view	with shadow removal	Not reported	[Zambanini et al., 2010] but 49F, 24NF	98% & 67%
[Yu et al., 2011]	Differences of centroid and orientation of a voxel person		One-class SVM	4	[Kim et al., 2005]		Oblique		OH by multi-view	Without shadow removal	Real-time	SD (29F, 29NF)	100% & 100%
[Auvinet et al., 2011, 2008]	3D silhouette vertical distribution	Threshold	Threshold	3 - 8	[Piccardi, 2004]		Oblique		Occlusion resistance	Without shadow removal	16 fps (3 views + GPU	[Auvinet et al., 2010]	81% & 100%
[Rougier et al., 2011b]	Mean matching cost, Full Procrustes distance		GMM and majority vote	4	[Kim et al., 2005]		Oblique		No OH	Without shadow	5 fps	[Auvinet et al., 2010]	95% & 96%
[Rougier et al., 2011a]	3D centroid height, velocity	threshold		1 Kinect	Not use		Oblique		OH by 3D velocity	insensitive	Real-time	SD (25F, 24NF)	96% & 100%
[Planinc and Kampel, 2012a]	Orientation, velocity	threshold		1 Kinect	Not use	Lie, crouch on the ground	Oblique		No OH	Insensitive	30 fps	SD (40F, 32NF)	93% & 100%

Continued on next page

Table 2.2 – *Continued from previous page*

Method	Features	Classifier	Fall inference	No. of cam	Background Subtraction	Sensitive events	View-points	Sensitive human movement	Occlusion	Lighting	Real-time	Data-set	Performance (SE & SP)
[Planinc and Kampel, 2012b]	Orientation, velocity	fuzzy logic	Threshold	1 Kinect	Not use	Lie, crouch on the ground	Oblique		No OH	Insensitive	Not reported	SD (40F, 32NF)	99% & 100%
[Zhang et al., 2012a]	Structure difference cost, height, histogram of aspect ratio	Hierarchy SVM		1 Kinect	Inter-frame Difference		Oblique		No OH	Insensitive	Not reported	SD	Not reported
[Zhang et al., 2012b]	5 kinds of features		Bayesian Inference	1 Kinect	Not use		Oblique		No OH	Insensitive	Real-time	SD (26F, 40NF)	100% & 95%
[Mastorakis and Makris, 2012]	Velocity, 3D bounding box	Threshold	Inactive verification	1 Kinect	Not use		Olique		No OH	Insensitive	30 fps	SD (48F, 112NF)	100% & 100%
[Dubey et al., 2012]	Hu-moment, 3D-MHI		SVM	1 Kinect	Not use		Oblique		No OH	Insensitive	Not reported	SD (39F, 999NF)	97% & 97%
Our method [Hung and Saito, 2012, 2013]	Heights, occupied areas	SVM	State transition analysis	2	[KaewTraKuPong and Bowden, 2001]		Oblique		OH by multi-view	Without shadow removal	25 fps (320x240)	[Auvinet et al., 2010]	96% & 100%
Our method [Hung et al., 2013]	HGCA	Threshold	State transition analysis	2 or more	[KaewTraKuPong and Bowden, 2001]		Oblique		OH by multi-view	Without shadow removal	25 fps (320x240)	[Auvinet et al., 2010]	96% & 96%

Chapter 3

Fall detection based on heights and occupied areas

3.1 Approach overview

Our key idea in this method is to approximate the person of interest by 3D cuboids from which we extract 3D spatial features for fall inference, i.e., the combination of heights and occupied areas. The occupied areas are defined as the bottom areas of the 3D cuboids. We realize that people in lying states occupy larger areas than those in standing and sitting states. The heights of people in standing states are also greater than that of people in sitting and lying states. Therefore, the combination of heights and occupied areas is highly discriminative in distinguishing human states into standing, sitting and lying. Intuitively, these three human states fall into three separable regions of the proposed feature space, composing of heights and occupied areas.

To facilitate the low-cost computation of 3D cuboid approximation, we configure two cameras whose fields of view are relatively orthogonal as shown in Fig. 3-1. 2D bounding boxes of the person of interest extracted from two cameras serve as two orthographic projections of the 3D cuboids. As a result, the reconstruction of 3D cuboids is straightforward. The occupied areas are defined as the bottom areas of the 3D cuboids and are determined by dot product of the widths of two

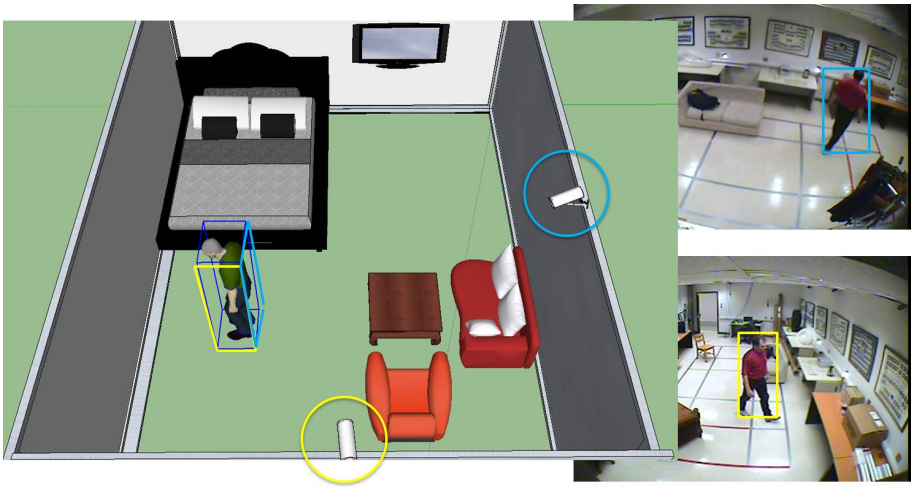


Figure 3-1: Two cameras whose fields of view are relatively orthogonal. It is straightforward to observe that 2D bounding boxes extracted from two cameras serve as 2D orthographic projections of the 3D cuboids of the person of interest.

2D bounding boxes. However, the reconstructed 3D cuboids are not view-invariant across the viewing windows due to the camera perspective. Hence, we suggest using Local Empirical Templates (LET) which were originally proposed for counting people [Hung et al., 2010, 2012], to normalize the reconstructed 3D cuboids. Similarly, we also divide the image (or the scene) into local image patches (or grid cells) and define LET as the sizes of a standing person in local image patches. LET are used in our work because of two following attractive properties. Firstly, LET in unknown scenes can be easily extracted by an automatic manner. Even though LET extraction is considered as the prerequisite initialization of our approach, we can perform it in an automated way. In cases of unexpected changes in viewpoints, the initialization can be redone or LET can be updated automatically without engineering intervention. Secondly, by its nature, LET hold the perspective information that can be used to normalize 3D cuboids, in turn, make them become view-invariant across the viewing windows. The normalization is not only to cancel the camera perspective but also to take the features of standing people as the baselines, making the feature space composing of normalized heights and normalized occupied areas separable for three human states, i.e., standing, sitting, and lying. In the last step, we perform time-series analysis of human state transition, which is inspired by the state transition

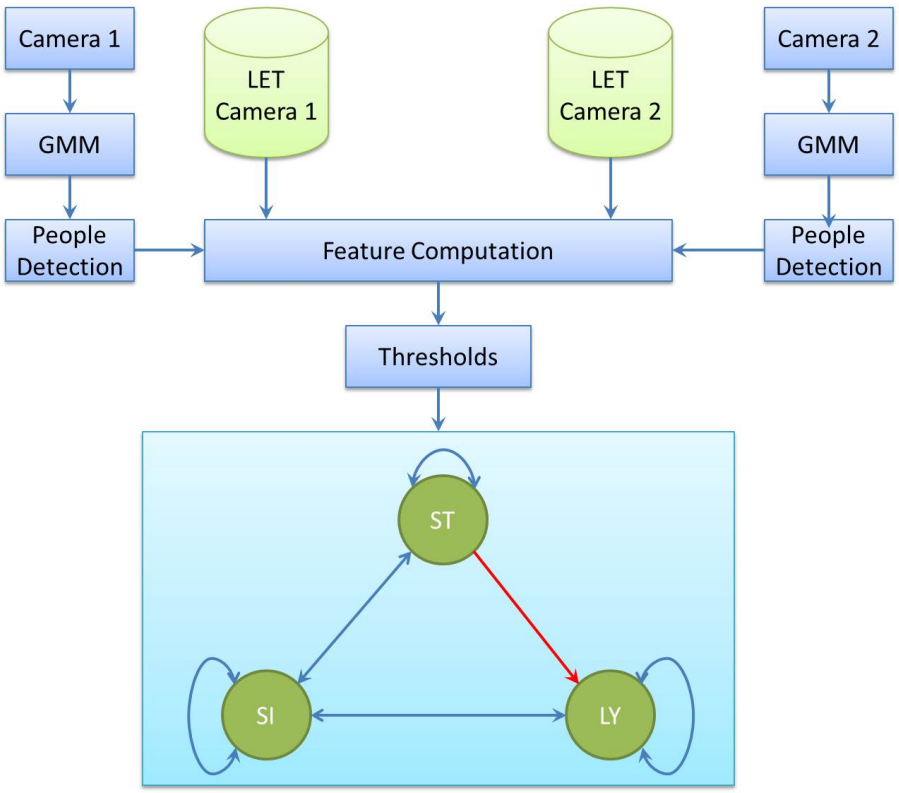


Figure 3-2: The flowchart of our proposed method

graph in [Cucchiara et al., 2005].

Fig. 3-2 shows the flowchart of our proposed approach. The two cameras are in oblique viewpoint settings. The video sequences are processed by Gaussian Mixture Models (GMM) [KaewTraKuPong and Bowden, 2001] to segment foregrounds for detecting the person of interest. In the next sections, we describe the key modules of our proposed approach in detail.

3.2 Local empirical templates

In this section, we provide more insights into LET, particularly how to obtain LET in unknown scenes. LET are the foregrounds induced by a single person (in the standing posture) in the local image patches. They are clustered upon their features of similar silhouettes along trajectories because people are different in sizes and exhibit various activities of upper limbs during walking. Each image patch has the most appropriate

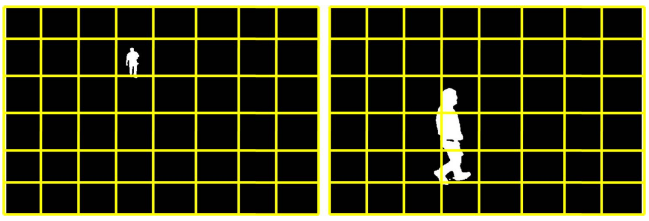


Figure 3-3: Local Empirical Templates. Sizes of the grid in this figure are for demonstration purpose. In practice, the number of cells is determined based on the viewpoints and image resolution.

LET, resulting in tens or hundreds of empirical templates across the viewing window, depending on the image resolution. Roughly speaking, the local empirical templates can outline the foregrounds typically made by a single person at local patches. By its nature, LET hold the perspective information of the scenes. Given a scene, the empirical templates most appropriate for the scene should be determined when a single person is spotted in the scene.

We divide the image of the scene into many cells or image patches, as shown in Fig. 3-3. There is one LET reflecting the typical sizes of standing people in each cell. Our observations in Fig. 3-3 are that the size of the person in the left image is small since he is far from the camera. Meanwhile, his size in the right image is greater since he is close to the camera. Evidently, LET hold the perspective information of the scene.

Suppose that the scene is divided into $M \times N$ cells so that the sizes of people are nearly constant in each cell. The number of cells depends on the resolution and the viewpoints of cameras. It is reasonable to observe that one LET does not appear fully in one cell but expand in several cells as shown in Fig. 3-3. Thus, we define the LET for the cell (i, j) as the following.

$$T(i,j) = \{W_T(i,j), H_T(i,j)\} \tag{3.1}$$
$$i \in [1, M], j \in [1, N]$$

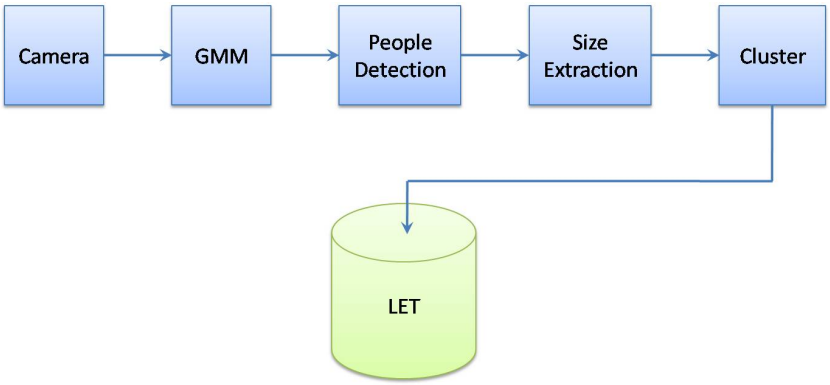


Figure 3-4: The flowchart of LET extraction process

where, $T(i,j)$ the LET whose head appears in the cell (i,j), $W_T(i,j)$ and $H_T(i,j)$ width and height of the LET, respectively.

The fall detection method is dedicated to a specific elderly person. LET should be determined by capturing the sizes of the person of interest to improve the accuracy of the method. By using LET, it is straightforward to customize our approach to various elderly people. The privacy of users is also protected because LET are merely the foregrounds. The customization can be done easily without engineering intervention since LET can be extracted automatically for unknown scenes. In contrast to [Hung et al., 2010, 2012] in which not only single people but also a group of people with occlusion each other are taken into consideration, we are dealing with a single person of interest. The automated LET extraction is more straightforward than in [Hung et al., 2010, 2012]. We capture the foregrounds and trajectories of the person moving around the scene. The sizes of foregrounds are extracted and kept in each cell's buffer for clustering to generate an appropriate LET for the cell. Since the cameras are in oblique settings, LET in the cell (i,j) and in its neighborhood cells should not be much different. Otherwise, the foreground extraction seem to be erroneous due to noise or occlusion by furniture. In such cases, we perform interpolation to get better results. The flowchart of LET extraction process is shown in Fig. 3-4.

In cases of occurring unexpected changes in camera viewpoints, the initialization can be redone automatically without engineering intervention. However, we argue that the viewpoint only changes a little due to earthquake or human factors. As a

result, LET do not change much, then we can use the following formulae to update LET.

$$H_T^{new}(i,j) = (1-\alpha)H_T^{old}(i,j) + \alpha H(i,j) \tag{3.2}$$

$$W_T^{new}(i,j) = (1-\alpha)W_T^{old}(i,j) + \alpha W(i,j) \tag{3.3}$$

where α the learning rate, $\{W_T^{new}(i,j), H_T^{new}(i,j)\}$ the updated LET of the cell (i,j), $\{W_T^{old}(i,j), H_T^{old}(i,j)\}$ the LET of the cell (i,j) before updated, $\{W(i,j), H(i,j)\}$ the size of person observed at cell (i,j).

3.3 People detection

Foreground, segmented by GMM [KaewTraKuPong and Bowden, 2001], is enhanced by applying morphological operators such as open and close to eliminate pepper noise before being labeled by connected component algorithms (CCA). Isolated foreground regions labeled by CCA are so-called blobs. After these preprocessing steps, a pool of N blobs $\{B_1, B_2, ..., B_N\}$ is created for the algorithm of people detection.

We search in the pool of blobs to find a head candidate and then group blobs in the neighborhood of the head candidate to form a person. The common labeling order of CCA is from top to bottom and subsequently from left to right of images. People are supposed to be in upright poses. Consequently, the blob with smallest label is likely the head candidate. LET of the cell in which the head candidate appears provides the tentative size of detected person $\{W_T, H_T\}$ or the tentative area in which the person appears. All blobs whose centroids satisfy the spatial constraint posed by LET likely belong to the person. They are grouped together for accumulating their densities and extracting the boundaries. We take the ratio of the total density to the size of the appropriate LET by the following formula in [Hung et al., 2010].

$$D = \frac{Total\_Density}{W_T \times H_T} \tag{3.4}$$

We confirm a detection if the density ratio exceeds a particular threshold. Please refer to Table 1 in the study of [Hung et al., 2010] for selecting the threshold of 0.3. We update the pool of blobs by removing blobs of detected people. In the next search, the head candidate is associated with the blob with the smallest label remaining in the pool. The process of searching for head candidates and grouping blobs in the neighborhood of head candidates is continued until there is no blob remaining in the pool. The algorithm of people detection is summarized as pseudo code in Fig. 3-5.

3.4 Features computation

Since we are using two cameras whose fields of view are relatively orthogonal as shown in Fig. 3-1, the 2D bounding boxes extracted from the two cameras serve as two orthographic projections of the 3D cuboids. Thus, occupied areas which are defined as the bottom areas of the 3D cuboids, can be determined by the dot product of the widths of the two 2D bounding boxes. Suppose that the person appears in the cell (m, n) in the first view with the size of $\{W_1(m, n), H_1(m, n)\}$. We also observe this person in the cell (p, q) in the second view with the size of $\{W_2(p, q), H_2(p, q)\}$. The occupied area is estimated as the following.

$$OA(m, n, p, q) = W_1(m, n) \times W_2(p, q) \tag{3.5}$$

However, the estimated occupied areas by Eq. 3.5 vary across the viewing window because of the camera perspective. We normalize it by using an appropriate LET to make the feature view-invariant across the viewing window. We extract the LET $T_1(m, n) = \{W_{T1}(m, n), H_{T1}(m, n)\}$ in the cell (m, n) in the first view and $T_2(p, q) = \{W_{T2}(p, q), H_{T2}(p, q)\}$ in the cell (p, q) in the second view. The occupied area of LET can be estimated by the following formula.

$$OA_{LET}(m, n, p, q) = W_{T1}(m, n) \times W_{T2}(p, q) \tag{3.6}$$

We take the ratio of the occupied area of detected person to that of an appro-

```
LOOP
    IF N > 0
        Head_Candidate ← Blob with smallest index = B_si
        Density ← Density(B_si)
        P ← Position(B_si) = (m, n)
        LET ← T(m, n) = {W_T, H_T}
        Spatial_Constraint ← (m, n, W_T, H_T)
        IF N > 1
            sum ← 0
            LOOP in N blobs
                IF B_i satisfies Spatial_Constraint
                    Select B_i for grouping
                    Update the boundaries of detected person
                    Density ← Density + Density(B_i)
                    Remove B_i from the pool of blobs
                    sum ← sum + 1
                END
            END LOOP
            N ← N − sum
            Update the pool of blobs
        END
        Take density ratio D by Eq. 3.4
        IF D > Threshold
            Confirm 'A person is detected'
            Mark a rectangular box for detected person
        END
    ELSE
        Exit LOOP
    END
END
```

Figure 3-5: The algorithm of detecting people from the pool of blobs

priate LET for perspective normalization, leading to a promising feature, so-called normalized occupied area NOA.

$$NOA = \frac{OA(m,n,p,q)}{OA_{LET}(m,n,p,q)} = \frac{W_1(m,n) \times W_2(p,q)}{W_{T1}(m,n) \times W_{T2}(p,q)} \tag{3.7}$$

It is noted that LET are defined as the sizes of a standing person appearing in the vicinity of detected person. The normalization in Eq. 3.7 not only cancels the perspective but also takes the features of standing people as the baselines. In other words, the normalization measures the distance between the features of detected people and the appropriate LET (in standing states). NOA is both lower and upper bounded and does not depend on the cell index. The cell-index notation of NOA in Eq. 3.7 are removed for simplicity. NOA is also highly relevant to the three typical states because of the feature-state relationship. A person lying on the ground occupies a larger area than standing and sitting. The occupied area in sitting state is, in general, larger than that in standing states.

$$NOA_{Standing} < NOA_{Sitting} < NOA_{Lying} \tag{3.8}$$

However in practice, poor foreground segmentation, occlusion, and human body upper limb activities, might cause the estimation of NOA in standing and sitting states by Eq. 3.7 to be quite similar. Fortunately, the humans' heights are significantly different and can be used to discriminate standing states from sitting and lying states. In contrast to the estimation of 2D bounding boxes' widths, the estimated heights are prone to be erroneous caused by occlusion by furniture. However, we argue that people are occluded in one view but likely visible in the other one. The heights of two 2D bounding boxes should be fused for an enhanced result. In our work, we simply use maximum rule in the fusion. We also normalize the heights by an appropriate LET for perspective cancellation. The estimation of normalized heights is summed up by the following formulae.

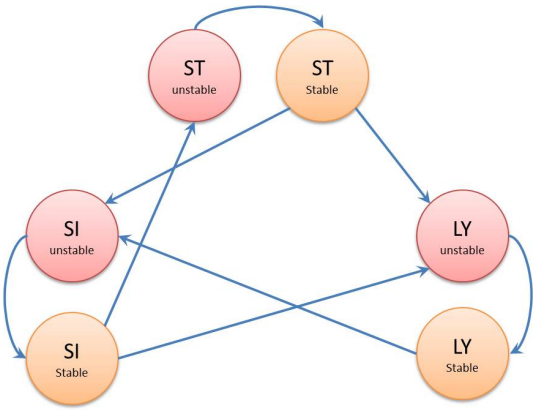


Figure 3-6: Time-series human state transition

$$
\begin{aligned}
NH_1 &= \frac{H_1(m,n)}{H_{T1}(m,n)} \\
NH_2 &= \frac{H_2(p,q)}{H_{T2}(p,q)} \\
NH &= Max(NH1, NH2)
\end{aligned} \tag{3.9}
$$

In summary, we have the feature space composing of normalized heights and normalized occupied areas that is separable for three typical states of humans. This property is discussed and demonstrated in Section 3.6.1. In the next section, we present how fall events are discriminated from usual ones, given a sequence of human states.

3.5 Fall inference

It is impossible to recognize human actions in a single frame or few frames since actions have temporal structures. Hence to make the fall event inference, we eye on the states of the elderly person in a period of time. In this paper, three typical states Standing (ST), Sitting (SI) and Lying (LY) are taken into consideration. A time-series analysis of human state transition shown in Fig. 3-6 that is inspired by the state transition graph in the study of [Cucchiara et al., 2005] is adopted. Table

3.1 sums up all actions, which can be inferred from the time-series analysis of human state transition. In general, all state transitions are allowed. However, for the specific application dedicated to the elderly, the direct transition from LY to ST states is quite improbable. The elderly often make the transitions in a gentle way from LY to SI and then to ST states.

Suppose that we keep states of the monitored elderly person in N frames for making event inference in a probabilistic manner. The *instant state* classified in each frame is not reliable for detecting state transitions. Therefore, we suggest using *stable states* and *unstable states*, instead of *instant states*. Only one out of three states, i.e., ST, SI, and LY, appearing in the window of N frames with the highest probability, is the stable state. The others are defined as unstable states.

$$Stable\_State = argmax_x\{P(x); ST, SI, LY\} \tag{3.10}$$

where, $P(x)$ the probability of observing the state x in the window of N frames, evaluated by frequentists paradigm, with $x \in \{ST, SI, LY\}$. Direct transitions between two stable states are not allowed. A state transition must undergo an unstable state before reaching its corresponding stable state, as illustrated in Fig. 3-6. When a state transition is in progress, the probability of observing the current stable state gradually decreases. Meanwhile, the probability of observing one of the other unstable states slightly increases. The state transition is confirmed upon the generation of a new stable state by Eq. 3.10.

In this work, we are interested in detecting fall incidents rather than other events.

Table 3.1: Actions can be inferred from the time-series analysis of human state transition

Current States	Next States		
	ST	SI	LY
ST	Standing or Walking	Sitting down	**Falling**
SI	Standing up	Sitting	Lying down
LY	*NA*	Getting up	Lying

START
    Update the pool of N states
        Delete the oldest state
        Add the latest state
    $stable\_state = argmax_x\{P(x); ST, SI, LY\}$
    IF $(current\_stable\_state == ST)\&(stable\_state == LY)$
        A Fall probably happened
        $start\_counter \leftarrow true$
        $counter \leftarrow 0$
        $current\_stable\_state \leftarrow stable\_state$
    END
    IF $start\_counter == true$
        IF $(stable\_state == LY)\&(current\_stable\_state == LY)$
            $counter \leftarrow counter + 1$
        END
    END
    IF $counter > Threshold$
        Confirm the Fall
        $start\_counter \leftarrow false$
    END
    IF Other state transitions happen
        $current\_stable\_state \leftarrow stable\_state$
    END
END

Figure 3-7: The time-series analysis of human state transition

In consideration of the definition and characteristics of a fall as discussed in Section 2.1, a fall event can be inferred by a direct transition from standing to lying states and subsequently an observation of staying in the lying state in some moments. Therefore, we dedicate a special attention to the aftermath of such state transitions to confirm a fall by verifying the duration of staying in the lying state after the state transition happened. The time-series analysis of human state transition to make inference of fall incidents is summarized as pseudo code in Fig. 3-7.

3.6 Performance evaluation

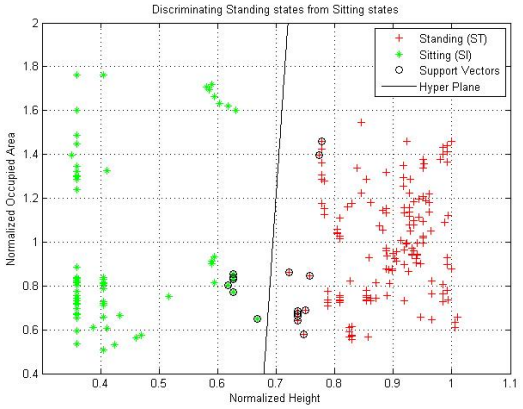
3.6.1 Linearly separable feature space

It is stated in Section 3.1 that the proposed feature space is separable for three typical states of humans. This section will discuss, demonstrate this statement, and find the decision boundaries for the state classification.

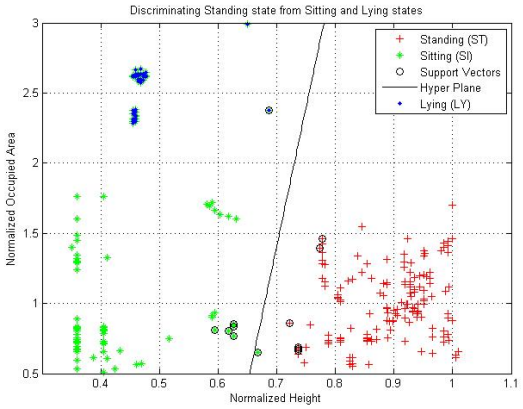
Since LET are defined as the sizes of standing people in local image patches, the normalization process takes the features of standing people as the baselines. It creates the distance measures between the features of detected people and the appropriate LET (in standing states). Thus, normalized heights, NH, of standing people should be approximate 1. For people in sitting and lying states, normalized heights are much smaller than 1. It is possible to distinguish standing states from sitting and lying states only based on the feature of normalized height. To discriminate lying states from sitting states, occupied area is a strong discriminative feature. Apparently, a person in lying states occupies a larger area than in sitting states. As a result, there exist two linear decision boundaries separating the feature space for three states of standing, sitting, and lying.

To demonstrate our discussion and to find the decision boundaries, we use the ninth scenario of the dataset for training purpose. In this scenario, the man approaches to the chair after entering the scene. He sits on the chair for a while and stands up before falling to the ground. The annotation of this scenario provides the state label in each frame. We calculate the feature vectors for every frame in combination with the corresponding state labels to create the training data.

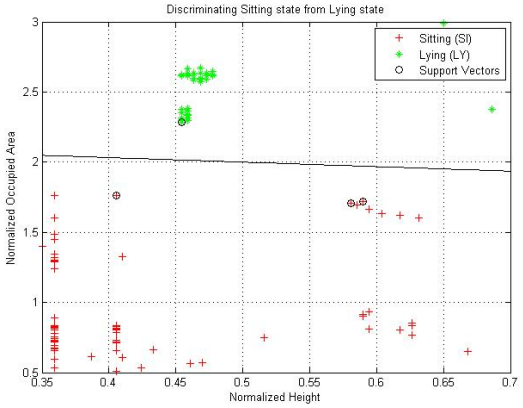
Both the training data sketched in the feature space in Fig. 3-8d and the above discussion show that it can be linearly separated. Therefore in this paper, two-class support vector machines (SVM) are adopted to find the decision boundaries for separating the three states. We make three following experiments in training SVM to find the decision boundaries. Firstly, standing states are separated from sitting states by a nearly vertical line in Fig. 3-8a. Secondly, we combine sitting and lying states as one class. The second class of SVM is the standing state. The decision boundary



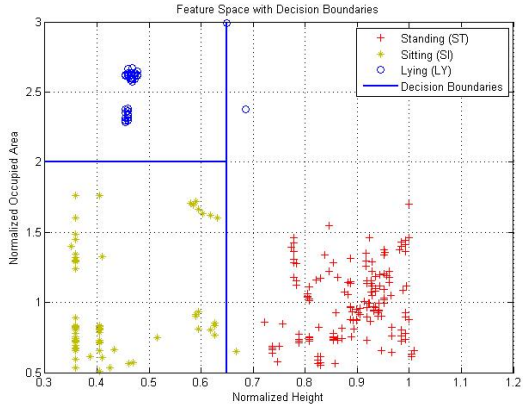
(a) Discriminating ST state from SI state



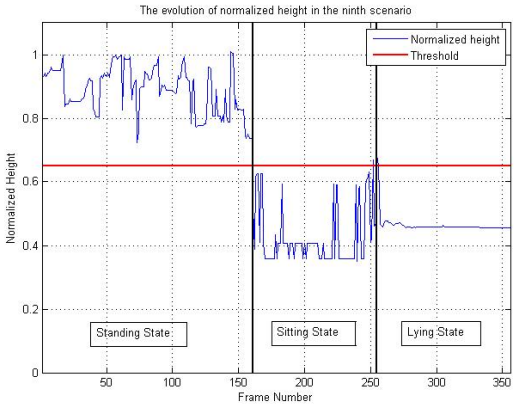
(b) Discriminating ST state from SI and LY states



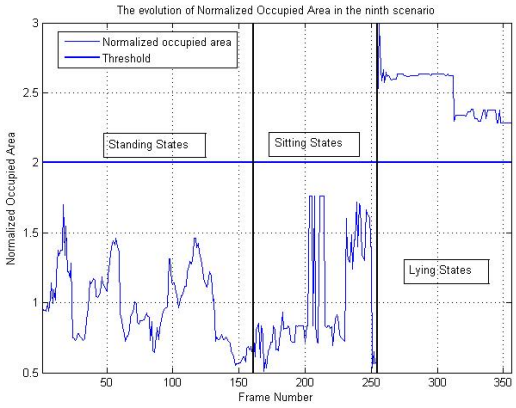
(c) Discriminating SI state from LY state



(d) Feature Space with modified decision boundaries



(e) Time-series evolution of normalized height



(f) Time-series evolution of normalized occupied area

Figure 3-8: Feature space of the ninth scenario with decision boundaries found by support vector machines

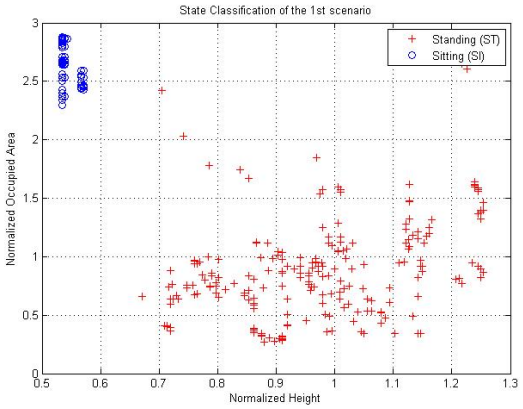
separating the two classes is given in Fig. 3-8b. Thirdly, the decision boundary for sitting and lying states is found in Fig. 3-8c as a nearly horizontal line.

The results of experiments in training SVM quite fit to our above discussion, except the one in Fig. 3-8b. However, it is clear to see some outliers in the training data of lying states, impairing the obtained decision boundary in Fig. 3-8b. The decision boundary in Fig. 3-8a indicates that normalized heights of people in sitting states cannot be greater than 0.7. This observation is also true for normalized heights of people in lying states. However, the decision boundary in Fig. 3-8b creates a region in which normalized heights of people in both sitting and lying states are well greater than 0.7. It is not reasonable in practice since the heights of people in sitting and lying states must be much smaller than in standing states. Therefore, the decision boundary for separating standing states from sitting and lying states in Fig. 3-8b should be a nearly vertical line, like the one in Fig. 3-8a. We make the modification for the obtained decision boundaries based on our prior knowledge of humans' heights, as shown in Fig. 3-8d. It leads to the generation of the thresholds for normalized heights and occupied areas, being 0.65 and 2, respectively. Fig. 3-8e and 3-8f show the time-series evolution of normalized heights and normalized occupied areas with obtained thresholds in the ninth scenario, respectively. In Fig. 3-9, we provide the visual results of state classification of the first and third scenarios in *multiple camera fall dataset* and the time-series evolution of each feature to further confirm the validity of the obtained thresholds.

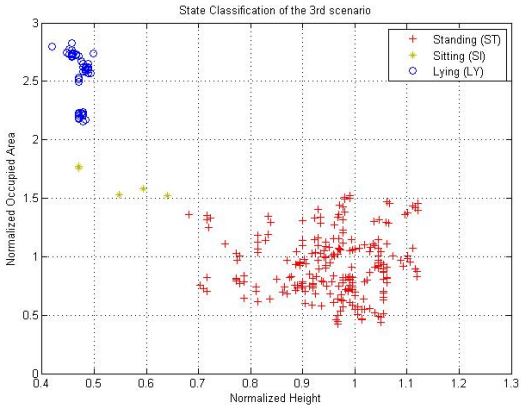
3.6.2 Performance evaluation and comparison

Our method detects 23 out of 24 fall incidents in the whole dataset. It only fails in the 22nd scenario in which the person is sitting on a chair and suddenly slips to the floor. Our method recognizes it as the lie-down event instead of a fall incident. No normal activity detected as a fall is reported in our experiments. The *sensitivity* and *specificity* are 95.8 % and 100 %, respectively.

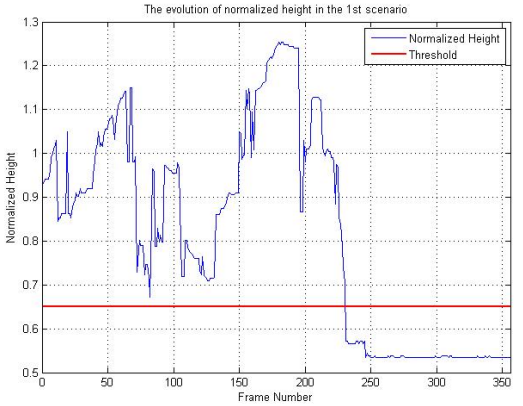
We compare the performance between our method and two state-of-the-art methods [Rougier et al., 2007b, 2011b; Auvinet et al., 2011], tested on the same dataset, in



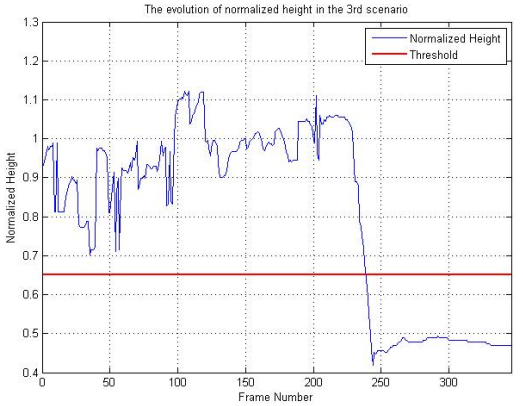
(a)



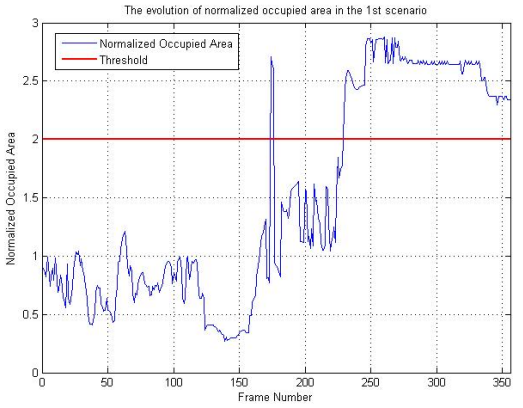
(b)



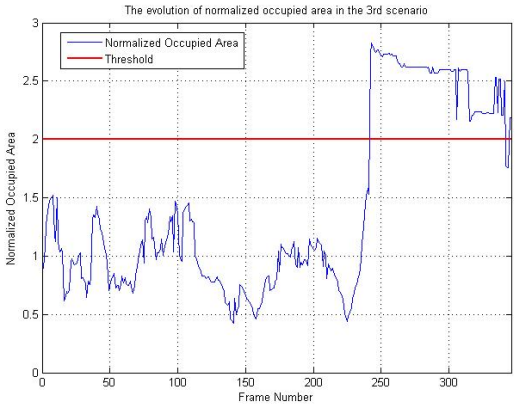
(c)



(d)



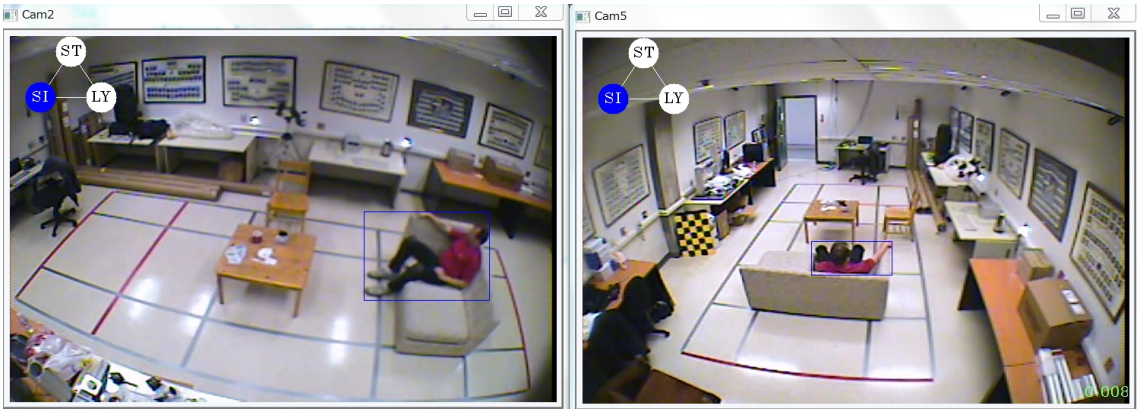
(e) The 1st scenario



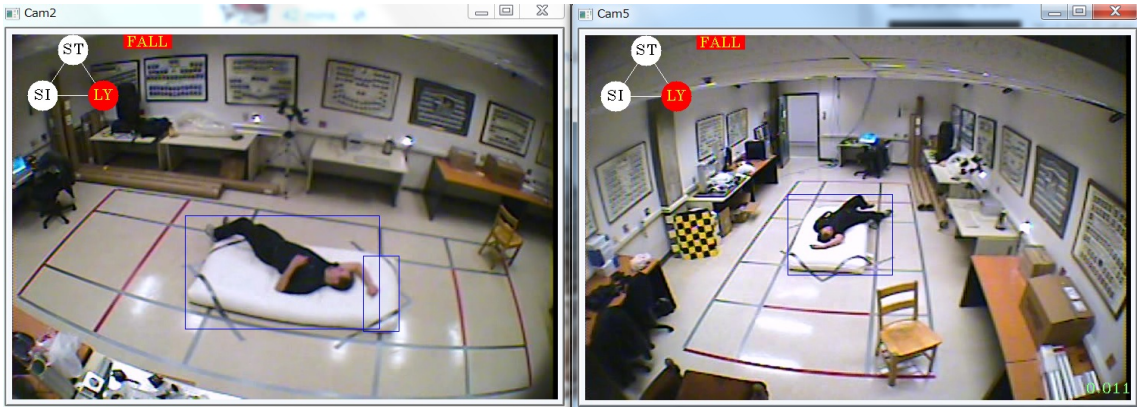
(f) The 3rd scenario

Figure 3-9: State Classification and the time-series evolution of normalized height and occupied area of the 1st and 3rd scenarios

(a) Standing state



(b) Sitting state



(c) Lying state

Figure 3-10: The results of human state classification in color image of both views. The Lying state in the last row is detected as a fall by the time-series analysis of human state transition.

Table 3.2. It is noted that the results of the method proposed by [Auvinet et al., 2011] are reported with a network of three cameras. The sensitivity can be boosted to 100 % if a network of more than four cameras is employed. However, both methods are high computational costs. In [Rougier et al., 2011b], the computational time for only shape matching is about 200 ms, resulting in an average frame rate of 5 fps. [Auvinet et al., 2011] presents three implementations of their method, i.e., CPU only, CPU for segmentation and GPU for projection, and GPU for both. The fastest performance by using GPU only, is 63 ms per one frame. It means the average frame rate is about 15 or 16 fps. The slowest performance by using CPU only, is 1140 ms per one frame. Average frame rate is less than 1 fps. Meanwhile, the early implementation of our approach reported in [Hung and Saito, 2013] is about 15 fps. However, in this implementation the input image sequences are resized to a half of the original, i.e., 360x240 to perform background subtraction. The resulting foreground images are resized to the original size of the input image sequences. We argue that the foregrounds do not change much by the scaling transformation since foreground images are binary. In this dissertation, we integrate the support of OpenMP to exploit more capability of CPU. Our approach is run at about 25 fps when the input image sequences are resized to a half and at around 15 fps without resizing the input image sequences[1]. However, the comparison is still quite unfair because of using different hardware and different background subtraction algorithms (see Table 2.2).

For a fairer comparison, we suggest measuring the processing time after having foreground images because all methods take them as input to extract features and perform recognition. The average processing time for feature extraction and fall inference of our approach is about 11 ms per one frame (running with original resolution of input image sequences. i.e., 720x480).

For our in-house video samples, our approach detects two fall actions and two sitting actions correctly. The visual result images are demonstrated in Fig. 3-11.

[1]The implementation in this dissertation is done by a notebook PC with chipset Intel core i7 3820QM, 16GB Ram

Table 3.2: Performance comparison between our method and two state-of-the-art methods [Rougier et al., 2007b, 2011b; Auvinet et al., 2011], tested on the same dataset.

	Sensitivity (Se)	Specificity (Sp)
Our method	**95.8 %**	**100 %**
[Auvinet et al., 2011]	80.6 %	100 %
[Rougier et al., 2007b, 2011b]	95.4 %	95.8 %

3.7 Discussions

In this section, false negative cases of this approach along with issues affecting its performance, i.e., lighting conditions, and occlusion by furniture are discusses.

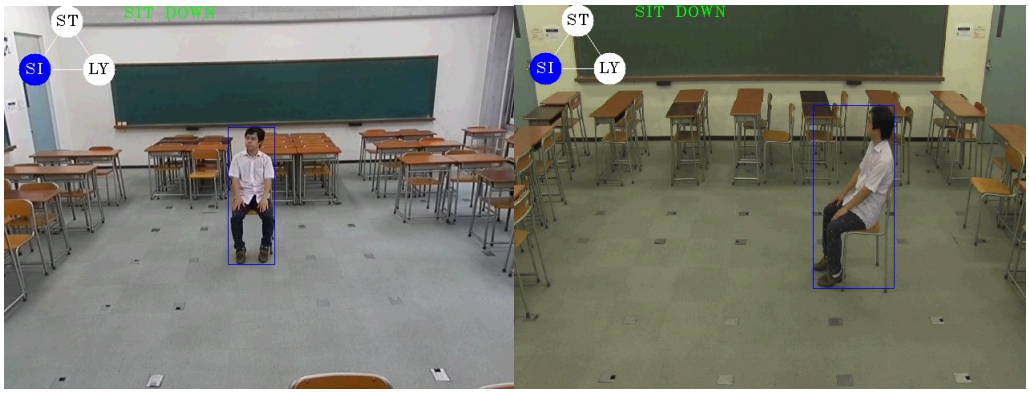
1. False negative cases

   Our proposed method only can detect a state transition from standing to lying as a fall. We argue that when the elderly want to lie down, for example, on a bed or a sofa, they will gradually sit down and then lie down. However, sit-to-stand-transfer fall in which the person changes states from sitting to lying is not considered or is detected as a normal transition by our method. This falling type happens quite often, in particular, when the person comes out of resting states in a bed or a sofa. Due to dizziness or syncope, the action of standing up is not completed or even not started. The person falls down on the ground rather than standing up. The reason of failure is by using height and occupied area, we can distinguish lying from standing and sitting states. But the information of where the person lie either on the ground after falls or on a sofa in normal situations is unknown. As a consequence, sit-to-stand-transfer falls are recognized as normal transitions like our result in the 22^{nd} scenario.

2. Lighting conditions

   Various background subtraction algorithms, i.e., basic motion detection [Benezeth et al., 2010], One-Gaussian model [Wren et al., 1997], Minimum, Maximum, and Maximum Interframe Difference [Haritaoglu et al., 2000], GMM [Stauffer and

(a) Standing (View 1) (b) Standing (View 2)



(c) Siting (View 1) (d) Siting (View 2)

(e) Falling (View 1) (f) Falling (View 2)

Figure 3-11: Visual results of our approach on the in-house video sample

Grimson, 2000; KaewTraKuPong and Bowden, 2001; Zivkovic, 2004], Kernel Density Estimation [Elgammal et al., 2000], and Codebook model [Kim et al., 2005] have been proposed in the literature. They are evaluated on a wide range of real, synthetic, and semisynthetic video sequences. Their performances are scored based on the robustness to various kinds of videos, i.e., noise-free static background, multimodal background, and noisy videos, the memory requirements, and the computational costs. The comparative study by [Benezeth et al., 2010] figures out that adaptive GMM not only produces good accuracy on both three types of videos but also requires less memory requirements and computational costs than other methods.

In our implementation, we do not use additional shadow removal algorithms along with the adaptive GMM [KaewTraKuPong and Bowden, 2001]. Small shadow and reflection still happen in our experiments but do not affect the performance of our approach. We argue that empirically there is almost no shadow and reflection for people sitting and lying on a sofa and even for those lying on the ground. Shadow and reflection happen and affect the accuracy of obtaining foreground of standing people. As a result, it may make the height of the person larger. Regardless of inaccurate estimated occupied area, the height of the person still well larger than the threshold of 0.65. The human state classification result is still correct, as the standing state.

3. Occlusion by furniture

   Occlusion by furniture[2] is very challenging because it happens frequently. It affects severely the accuracy of extracting features since the moving foreground of the person is not attained completely. We argue that the upper body parts are rarely occluded by the furniture. As a result, occupied area computed by Eq. 3.7 seems not to be affected by the occlusion. Although, the height of the person is severely shortened during occlusion, by using Eq. 3.9 the occlusion problem can be dealt by feature fusion. The person is occluded in one view but

[2] As stated in section 1.2, our proposed solutions target to single-user application. Thus, occlusion between human is not considered.

is likely visible in the other. In the experiments on a dataset containing limited challenges of the real world, using two cameras seems to sufficient. Eventually, up to four cameras[3] can be set up to overcome the occlusion problem at the expense of computational cost. Since the computational cost of adaptive GMM is much more than that of people detection, feature computation and event detection in our fall detection pipeline (see Fig. 3-2), the use of GPU can be easily realized the real-time performance.

3.8 Conclusions

We have presented a novel method of fall detection that plays as a central part of iPERS for aiding the elderly living alone. The novelty lies in the feature space composing of humans' heights and occupied areas to discriminate three typical states of humans, i.e. standing, sitting and lying. It is the fact that the heights of people in standing states are greater than in sitting and lying states. Moreover, People in lying states occupy a larger area than in sitting and standing states. Therefore, the proposed feature space is linearly separable for these three states. Fall incidents can be inferred from the time-series analysis of human state transition.

In implementation, we propose using two orthogonal views: (1) to simplify feature computation, and (2) to improve the reliability of computing the feature vector based on sizes of silhouettes in the presence of occlusion. People are partially occluded in one view but visible in the other one. The feature vector is normalized by the size of an appropriate LET to cancel the camera perspective and to realize the linear separability of the proposed feature space.

In performance evaluation, a good method of fall detection is associated with high *sensitivity* and *specificity*. We choose *multiple camera fall dataset* that only includes simulated falls by an experienced clinician in the healthcare for the elderly, to test our method for fair comparison with existing methods. The results of our method reach to 95.8 % of *sensitivity* and 100 % of *specificity*. It outperforms two state-of-

[3]Because of the relatively-orthogonal-views constraint

the-art methods [Rougier et al., 2007b, 2011b; Auvinet et al., 2011], tested on the same dataset. However, the comparison is merely based on the results tested on one dataset containing limited challenges of the real world. We need to further evaluate the proposed approach on the real falls of the real elderly in real home environments, in order to confirm the validity of this approach.

Chapter 4

Fall detection based on human-ground contact areas

In chapter 3, we describe an approach to fall detection based on 3D spatial features, extracted from 3D cuboids of the person of interest, that is the combination of heights and occupied areas. This method can distinguish lying from standing and sitting states. But the information of lying states such as the person lying either on the ground after falls or on a sofa in resting states is unknown. As a consequence, this method only can detect a state transition from standing to lying as a fall. Sit-to-stand-transfer falling type in which the person changes from sitting to lying states is not considered. Therefore in this chapter, we propose another 3D spatial feature to overcome the limitation of our solution in chapter 3. We argue that people always make a little contact with the ground during usual activities, mainly by the feet but often lie completely on the ground after suffering from accidental falls. We come up with another good 3D spatial feature for fall detection, so-called Human-Ground Contact Areas (HGCA). In this chapter, we describe how to estimate HGCA, classify HGCA into typical human states, and recognize falls based on given sequences of human states and HGCA.

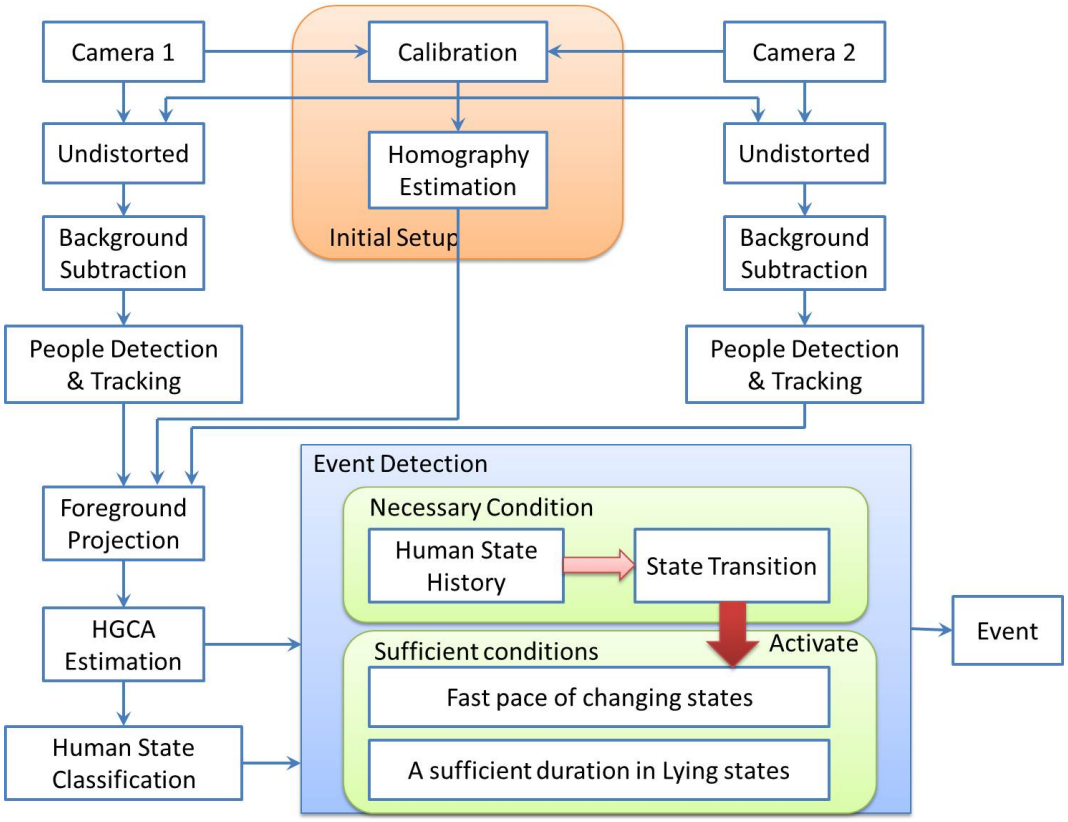


Figure 4-1: The flowchart of our proposed method

4.1 Approach overview

In multiple-view geometry [Hartley and Zisserman, 2004], any two images of a same planar surface (assuming a pinhole camera model) are related by a planar projective transformation, so-called homography. This geometrical relation was successfully applied to tracking people in multiple views [Khan and Shah, 2006]. In our work, we also use this geometrical relation to develop a low-cost and effective scheme of estimating HGCA for fall detection. Fig. 4-1 illustrates the flowchart of our proposed approach by using a pair of cameras. It is straightforward to extend our approach to a network containing more than two cameras by fusing detection results from pairs of cameras.

The homography of the ground plane between different views implies that only foreground pixels of people in contact with the ground plane are consistently projected

to their foreground regions in different views. Therefore, we project the foreground of the person of interest from one view to another by using the planar homography. There exist overlaps between the projected foregrounds and the foreground of the person in the latter view which indicate the contact areas between people and the ground. We measure the overlaps as HGCA. As our above argument, HGCA has close a close relationship with typical human states, i.e., standing, sitting, and lying. We generalize a threshold of HGCA to separate lying states from the others from view-invariant distributions of HGCA with respect to human states. We propose using human state simulation in which camera viewpoints are freely changed to capture 3D human models in various states. Hundreds of images are generated from the simulation as training data to build these distributions.

In indoor surveillance, optical lenses are used to capture the monitored space as widely as possible. Consequently, the images are highly distorted. It is necessary to undistort images since the homography is only held under a pinhole camera model [Hartley and Zisserman, 2004]. The cameras are calibrated to estimate intrinsic parameters including the focal length $\mathbf{f} = (f_x, f_y)$, the optical center $\mathbf{c} = (c_x, c_y)$, the skew coefficient α, and the distortion coefficients $\mathbf{k} = (k_1, k_2, k_3, p_1, p_2)$ for using in the image undistortion. In our implementation, these intrinsic parameters are obtained from the dataset [Auvinet et al., 2010].

The homography matrix H of the ground plane between the two cameras that will be presented in details in Section 4.2.1 is estimated by simply specifying four point correspondences [Hartley and Zisserman, 2004]. It is noted that the four point correspondences must be in undistorted images. These are the initial setup of our proposed method that can be done offline because of using stationary cameras in surveillance.

Foregrounds are segmented and enhanced by morphological operators before labeled by connected components algorithm, resulting in foreground blobs. These blobs are clustered to form foregrounds of people which are mapped between views by the homography of the ground to measure HGCA for event detection. In the next section, we introduce how to measure HGCA based on foreground projection by using

the homography.

4.2 HGCA Computation

4.2.1 Projecting foregrounds by using planar homography

To have the insight into the planar homography, let recall a proposition in multiple-view geometry.

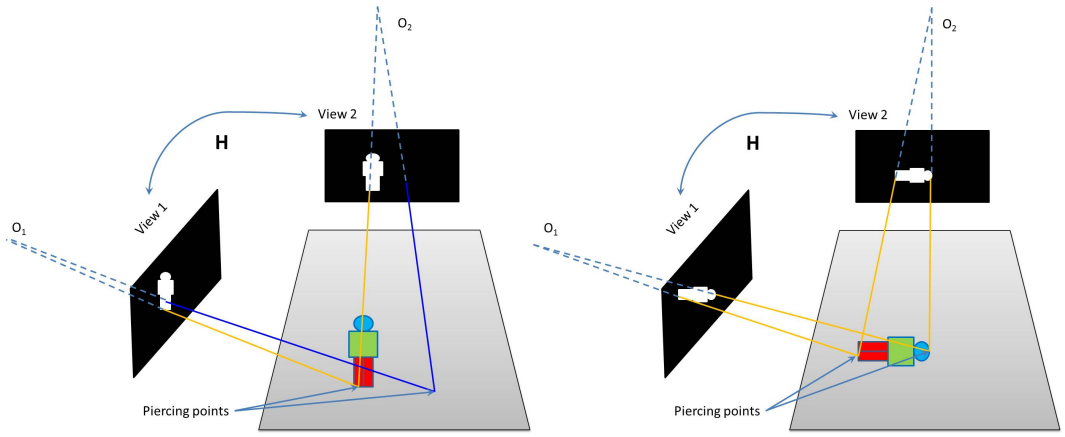
Proposition 1 [Faugeras and Lustman, 1988] Let p be a 3D point of a plane Π, the projections $p_1 = (x_1, y_1, 1)$ and $p_2 = (x_2, y_2, 1)$ of p (in homogeneous coordinates) on the two image planes are related by a homography H of the plane Π between the two views.

$$p_1 = Hp_2, \tag{4.1}$$

where H is a nonsingular 3×3 matrix. Let H_3 denote the third row of H. The point p_1 in the first image is mapped to p_m in the second one by the homography H [Khan and Shah, 2006].

$$p_m = (x_m, y_m, 1) = \frac{Hp_1}{H_3 p_1}. \tag{4.2}$$

The Proposition 1 implies that if the point p is on the plane Π, p_m and p_2 are coincident. But if the point p is not on the plane Π, there exists a misalignment between p_m and p_2, reflecting the plane parallax. These observations can be elaborated in Fig. 4-2. It is supposed that the ground plane is visible from the stationary cameras. People always make contact with the ground plane, mainly by their feet. Firstly, according to Proposition 1, only the projections of the feet on the image planes are related by the homography of the ground plane between the two views. This relationship is illustrated by the orange rays in Fig. 4-2a. Secondly, to clearly see the effect of the plane parallax, let consider the images of people's leg that is not on the ground plane. We assume that the projections of the leg on the image planes are known (inside the white regions in Fig. 4-2a). We project the ray from



(a) Standing people (b) Lying people

Figure 4-2: The illustration of using the planar homography of the ground (plane Π) between a pair of views for fall detection

the principal point of camera 1, O_1, to the leg on the image plane 1 until intersecting with the ground plane at so-called the piercing point. From this piercing point, we make another projection to the principal point of camera 2, O_2, intersecting the image plane 2 at a point outside the white region (see the blue ray in Fig. 4-2a). It is the mapped point of the leg on the image plane 1 into the image plane 2 by the homography of the ground plane but it is not coincident with the leg on the image plane 2.

We exploit these key observations to discriminate lying states from standing and sitting states for fall detection. If people are standing or sitting, making contact with the ground by feet, overlap regions between foreground in the second view and projected foreground happen at feet location. But when people lie on the ground, overlap regions cover almost whole body (see Fig. 4-2). The overlap regions between foregrounds are measured as HGCA.

4.2.2 HGCA Computation

Let Ψ_1 and Ψ_2 be sets of foreground in the first and second views, respectively. Let Ψ_m denote set of mapped foregrounds by homography H from the first to the second views and let X be human states, i.e., standing, sitting, kneeling, and lying, etc. We

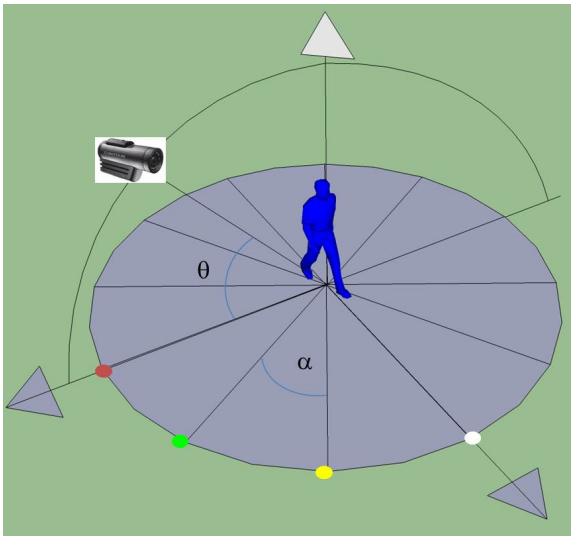


Figure 4-3: Simulation setup in Google Sketchup. Colorful dots are landmarks for homography calibration

suppose $\Psi_1 = \Psi_m = \{n\}$ pixels, and $\Psi_m \cap \Psi_2 = \{m \mid m \leq n\}$. HGCA is a function of X and is evaluated by

$$HGCA(X) = \frac{\Psi_m \cap \Psi_2}{\Psi_m} = \frac{m}{n}, \tag{4.3}$$

Equation 4.3 indicates how many percentages the projected foreground is overlapped by the foreground of the person in the second view. Therefore, the estimation of HGCA by Eq. 4.3 is view-invariant across the viewing window.

4.3 Human state simulation and classification

To generalize a threshold of HGCA to separate lying states from the others, training data is needed to build view-invariant distributions of HGCA with respect to human states. This section proposes a human state simulation by using Google Sketchup, as shown in Fig. 4-3, to generate training images. We consider three typical sitting poses, i.e. sitting on a chair and kneeling on the ground by one or two legs, and three lying poses, corresponding to three typical falls, i.e. falling forward, backward and sidewards. A virtual camera is positioned and freely changed on surface of a

hemisphere to capture images of 3D human models in various states and poses. The camera viewpoint is modeled in spherical coordinate system by

$$\begin{aligned} P_{camera} &= P(r, \alpha, \theta) \\ r &\approx const \\ \alpha &= [0, 180^o], \delta\alpha = 30^o \\ \theta &= [30^o, 75^o], \delta\theta = 15^o \end{aligned} \tag{4.4}$$

where, r the radial distance, α the azimuth angle, and θ the inclination angle. The angles are measured in degrees. The inclination angle greater than 75^o is not taken into account since the camera viewpoints are near the top view which is not appropriate for detecting falls. In addition, indoor surveillance cameras are often positioned obliquely near the ceiling. The spatial constraints make variations of inclination angles in the range of $[30, 75]$.

In simulation, both azimuth and inclination angles are changed by steps $\delta\alpha$ and $\delta\theta$ in Eq. 4.4 to generate 196 training images. Fig. 4-4 illustrates some generated images. Homography of the ground between these views are automatically calibrated by matching colorful dots on the ground. We project foregrounds between every pair of different views (α_1, θ_1) and (α_2, θ_2) in the training set with $\Delta\alpha = |\alpha_1 - \alpha_2| > 0$ or $\Delta\theta = |\theta_1 - \theta_2| > 0$ to estimate HGCA for building distributions. Our aim is to generalize a threshold of HGCA to separate lying states from the others regardless of viewpoint variations. It is noted that human models in simulation are stationary and cameras are moving. Both $\Delta\alpha$ and $\Delta\theta$ are determined by positions of cameras. But in reality, cameras are fixed and people are moving. When people travel in the fields of views of a fixed-camera pair, both azimuth and inclination angles $(\Delta\alpha, \Delta\theta)$ vary freely which are determined by positions of both people and cameras, not by the positions of cameras only.

We realize from the simulation that measuring HGCA by using Eq. 4.3 is usually inaccurate for a pair of viewpoints with $\Delta\alpha < 90^o$ as shown in Fig. 4-4 since overlap-

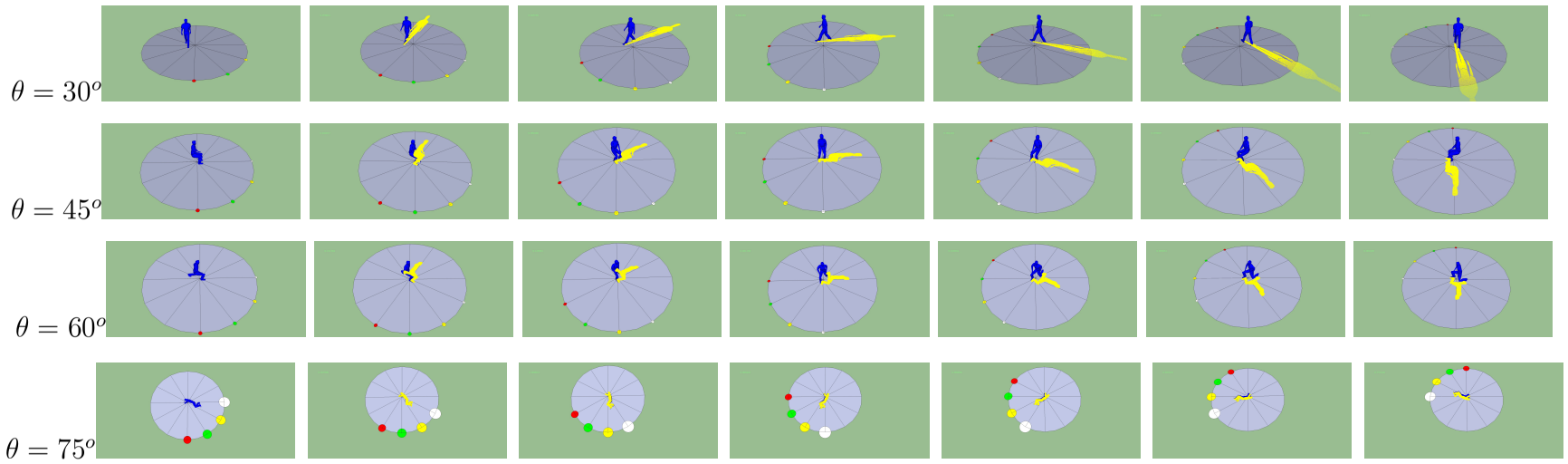
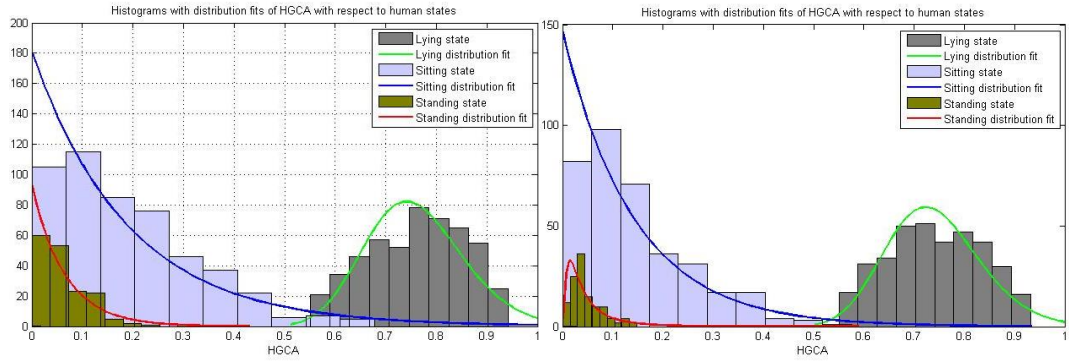


Figure 4-4: Some generated image samples from the simulation. We show standing, sitting, kneeling and lying people in rows. Foregrounds in first-column images are projected and overlaid by yellow foregrounds in images in other columns



(a) All viewpoints (b) All viewpoints except $\Delta\alpha = 30^o$

Figure 4-5: Histograms and distribution fits of HGCA with respect to human states. Exponential and normal distributions fits are for standing, sitting and lying states, respectively.

ping regions usually cover body parts which are not in contact with the ground. This phenomenon does not happen for a pair of viewpoints with $\Delta\alpha \geq 90^o$. Thus, using a pair of cameras positioned at least 90^o apart (in terms of azimuth angles) will lead to fewer situations occurring $\Delta\alpha < 90^o$ when people travel in their fields of views. The constraints of indoor spaces and camera placement make some viewpoints frequently happen and some rarely happen. Thus, viewpoints can be weighted unequally to adapt to specific contexts. Fig. 4-5b show distributions of HGCA excluding viewpoints of $\Delta\alpha = 30^o$. However, these distributions in Fig. 4-5a and 4-5b are quite similar. In this paper, we use the distributions in Fig. 4-5a in experiments for various camera settings.

To find a threshold to separate lying states from standing and sitting states, we build exponential distribution fits of HGCA with respect to standing and sitting states. We also build normal distribution fit of HGCA with respect to lying states. The threshold is found at the intersection of two distribution fits of HGCA with respect to sitting and lying states.

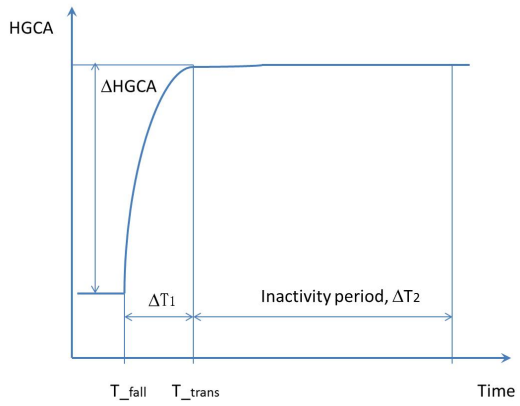


Figure 4-6: Typical fall characteristics based on HGCA

4.4 Framework for fall inference

The fall definition in Section 2.1 leads to a typical fall characteristics based on HGCA, as shown in Fig. 4-6. When a fall happens at T_{fall}, HGCA increases by $\Delta HGCA$ to make a state transition from Usual states composing of Standing, Sitting, and Kneeling to Lying states in a so-called falling period of ΔT_1. Subsequently, the fallen person is relatively immobile in a period of ΔT_2. In this section, we present a framework for fall inference based on the fall definition that can be broken into *necessary* and *sufficient conditions*.

Necessary condition is the change of human postures from upright to lengthened. In our framework, the change of human postures is described by a state transition from Usual states to Lying states. Such state transitions are caused not only by fall events but also by lying-down events (people lying on a sofa or a bed). Fall events are associated with a fast movement of human body. In contrast, lying-down events are performed in a leisure manner by the elderly. Therefore to claim a state transition as a fall event, we must verify the following *sufficient conditions*.

Sufficient conditions compose of a fast pace of changing states and an observation of Lying states in a sufficient duration after the state transition. Without satisfying both sufficient conditions, the state transition is likely not caused by a fall event. The proposed framework for fall inference is described in the module of Event Detection

in Fig. 4-1.

To this end, we keep both human states X and HGCA in the buffer of N frames. Upon a state transition at T_{trans}, we start verifying both sufficient conditions to whether claim a fall event. Fast pace of changing states is characterized by ΔT_1 and $\Delta HGCA$ which are evaluated by

$$\Delta HGCA = HGCA[T_{trans}] - HGCA[T_{fall}] \tag{4.5}$$
$$\Delta T_1 = T_{trans} - T_{fall}$$

where $\Delta HGCA$ the increment of HGCA in falling period. From distributions in Fig. 4-5, $\Delta HGCA$ should be at least 0.3 to prevent from trivial state transitions, likely caused by noise.

$$\Delta HGCA \geq 0.3 = Min_{\Delta HGCA} \tag{4.6}$$

Combine Eqs. 4.5 and 4.6,

$$HGCA[T_{fall}] \leq HGCA[T_{trans}] - Min_{\Delta HGCA} \tag{4.7}$$

T_{fall} and ΔT_1 are calculated by using Eqs. 4.5 and 4.7. A state transition is considered as fast if $\Delta T_1 \leq 1$ second [Noury et al., 2008]. Finally, extracting ΔT_2 is very straightforward to verify the second sufficient condition. In practice, 5 seconds are considered to be long enough for experiment conditions. Figure 4-7 shows the temporal evolution of HGCA of the scene 18 in our experiments.

4.5 Performance evaluation

We test our method with two pairs of cameras, i.e. cameras 2 and 7 (positioned 180^o apart), cameras 2 and 5 (positioned 90^o apart) and with three of them. Performance of three experiments are compared with that of other methods, tested on the same

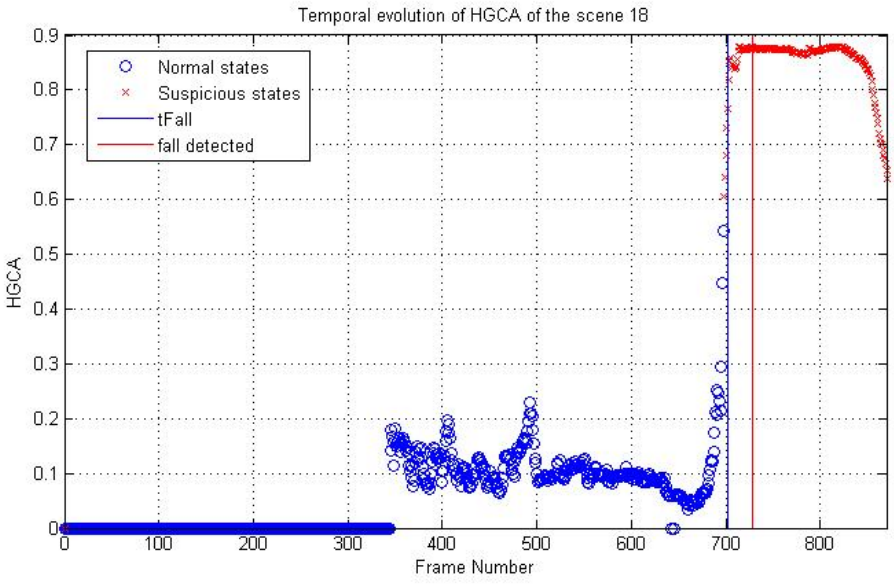


Figure 4-7: Temporal evolution of HGCA of the scene 18 in our experiments.

dataset, in Table 4.1. Occlusion by furniture frequently happens in indoor surveillance that make HGCA estimation sometimes inaccurate, degrading performance. Our approach fails to detect the fall events in scenes 15 and 22 by the pair of cameras 2 and 5 but correctly detects them by the pair of camera 2 and 7. The reason is that the sofa covers a part of the body when he lies on the ground after falling. Similarly, our approach fails to detect the fall event in the scene 19 by the pair of cameras 2 and 7 but correctly detects it by the pair of cameras 2 and 5, due to the same reason. Thus, we fuse the detection results by two pairs of cameras, simply by OR rule, producing better performance as shown in Table 4.1. The decision fusion makes the approach robust to furniture occlusion since we believe that people are occluded in one view but are likely visible from the other viewpoints. Some visual results of our proposed approach are shown in Fig. 4-8.

In implementation, we also use the strategy of resizing the input image sequences to a half of the original size and the support of OpenMP to exploit more capability of CPU. The average frame rate of our approach is about 25 fps[1]. If the input image

[1]The implementation in this dissertation is more optimized than the preliminary one reported in [Hung et al., 2013].

(a) Standing state

(b) Sitting state

(c) Lying state

Figure 4-8: The visual results of projecting foreground from one view to another by using homography of the ground between the two views for standing, sitting and lying state. The lying state in the last row is detected as a fall event by the temporal analysis of human state transition.

Table 4.1: Performance comparison between our method and three state-of-the-art methods [Auvinet et al., 2011; Hung and Saito, 2012; Hung et al., 2013; Rougier et al., 2007b, 2011b], tested on the same dataset. Results in [Auvinet et al., 2011] are with 3 cameras.

	Sensitivity	Specificity
Our method (cameras 2, 5 and 7)	95.8 %	96 %
Our method (cameras 2 and 7)	88 %	100 %
Our method (cameras 2 and 5)	88 %	96 %
[Hung and Saito, 2012, 2013]	95.8 %	100 %
[Rougier et al., 2007b, 2011b]	95.4 %	95.8 %
[Auvinet et al., 2011]	80.6 %	100 %

sequences are kept in the original size, our approach runs at about 15 fps. The processing time for feature extraction and fall inference of our approach is about 11 ms per one frame on average, running with the original image resolution (720x480).

For our in-house video sample, our approach detects two falling actions and two sitting actions correctly. The visual result images are demonstrated in Fig. 4-9.

In comparison with other methods tested on the same dataset, the fall in the scene 20 is not detected by [Rougier et al., 2007b] since the fall takes place in two steps. The person starts falling but leans on the sofa and then finishes the fall later. These two steps produce two consecutive peaks in the full Procrustes distance curve. The second peak makes the method failure since a sufficient period of immobility is not detected to confirm the fall. Occlusion by furniture also causes difficulty for this method because of using only one camera. However, both of our approaches can detect this fall event. When the person leans on the sofa, the occupied area and HGCA are still small. But when the person finishes the fall on the ground, these features change significantly to make a state transition for event detection.

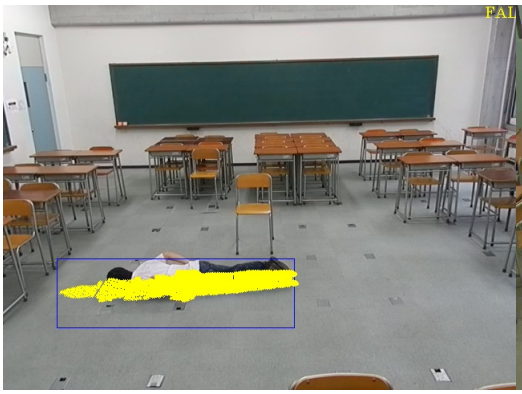
The method in [Auvinet et al., 2011] also cannot detect a fall on furniture so that a large part of the body remains above 40 cm. It means that their method fails to detect the falls in 15th and 16th scenarios. However, the argue that this limitation can be overcome by employing environmental information. Our approach in this chapter has the same limitation with [Auvinet et al., 2011] since the contact area between

(a) Standing (View 1)

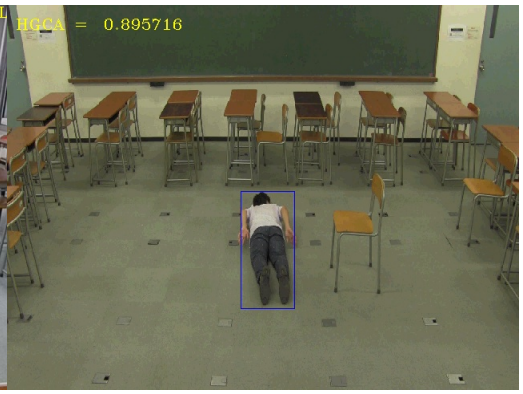
(b) Standing (View 2)

(c) Siting (View 1)

(d) Siting (View 2)



(e) Falling (View 1)



(f) Falling (View 2)

Figure 4-9: Visual results of our approach on the in-house video sample

the person and the ground is used. But our appoach in chapter 3 can detect the falls in both scenarios since the occupied areas change dramatically during the fall, regardless of where the person finishes the fall.

4.5.1 Performance discussions

In this section, false negative cases of this approach together with issues affecting its performance, i.e., lighting conditions and occlusion by furniture are discussed.

1. False negative cases

   Unlike the method in chapter 3, this method can detect sit-to-stand-transfer fall in the 22^{nd} scenario. We carry experiments on this scenario by using two pairs of cameras (the pair of 2 and 5, and the pair of 2 and 7), producing both true positive results. However, this method has several false negative cases. Firstly, it only can detect a fall ending on the ground since it relies on the measurement of contact area between the person and the ground. Secondly, it is sensitive to occlusion by furniture. The lying position on the ground after falling occluded by the furniture poses difficulty to this approach. The issue of occlusion will be discussed further in the following section.

2. Lighting conditions

   Like the method in chapter 3, we also do not use any shadow removal algorithm along with the adaptive GMM [KaewTraKuPong and Bowden, 2001]. Although the adaptive GMM can eliminate well the effect of shadow and reflection, small shadow and reflection still happen in the experiments and are included in the foreground of the person. We argue that shadow and reflection happens when people walking and kneeling on the ground. There is almost no shadow and reflection, when people sitting and lying on a sofa, and even lying on the ground. In the case of people lying on the ground, two sets of foregrounds Ψ_m and Ψ_2 are nearly completely overlapped, resulting in high scores of HGCA. Since shadow and reflection take place on the ground, it may contribute to the overlap

regions between two sets Ψ_m and Ψ_2 when people walking or kneeling on the ground, depending on positions of cameras and lighting direction. However, its contribution is not significant to make two sets Ψ_m and Ψ_2 nearly completely overlapped, consequently, not resulting in high scores of HGCA.

3. Occlusion by furniture

   This method is sensitive to the occlusion by furniture since it relies on the measurement of contact area between human and the ground. The fact of lower body parts being frequently occluded by the furniture makes the computation of HGCA inaccurate. In particular, the method fails to detect a fall event if after suffering the fall the lying position is partly or completely occluded by the furniture. The experiments reported in section 4.5 demonstrate the failure scenarios of our proposed approach. However, the occlusion problem can be overcome by using more than two cameras at the expense of computational cost. It is the fact that people are occluded in one view but likely visible in the other ones. That is, we have more chances to observe the person appearing fully for accurate feature computation. The final decision is enhanced by fusing decisions made independently by every pair of cameras. In section 4.5, we also present the results of fusing decisions made independently from two pairs of cameras (the pair of 2 and 5, and the pair of 2 and 7). All false negative results made by the two pairs of cameras due to occlusion are corrected.

4.6 Conclusions

In this chapter, we have presented a 3D spatial feature, Human-Ground Contact Areas, measuring only the contact area between the human and the ground floor. HGCA has a close relationship with human states because of the fact that people always make contact with the ground during usual activities, i.e., standing and sitting mainly by the feet. Meanwhile, people often lie completely on the ground after suffering from accidental falls. HGCA is highly discriminative to discern lying states

from other usual states, i.e., standing and sitting. We also introduce a low-cost scheme to estimate HGCA efficiently by using the foreground projection between two different views based on the homography of the ground. Fall inference is performed by using temporal analysis of human state transition, inspired by the fall definition. The performance of this approach is competitive with the state-of-the-art methods [Hung and Saito, 2013; Auvinet et al., 2011; Rougier et al., 2007b, 2011b]. However, the comparison is merely based on the results tested on one dataset containing limited challenges of the real world. We need to further access it in real situations with the real elderly, in order to confirm the validity of this approach.

Chapter 5

Bag of Video Word Approaches to Fall Detection

5.1 Introduction

Recently, Bag of Video Word (BoVW) approaches have produced good results in video-based human action recognition. In general, BoVW approaches employ local interest point detectors, i.e., Harris3D [Laptev, 2005], Cuboid [Dollár et al., 2005], Hessian [Willems et al., 2008], MoSIFT [Chen and Hauptmann, 2009] detectors, dense sampling, and dense trajectories [Wang et al., 2013] to locate salient points in space-time domain as interest points. Various descriptors like Cuboid [Dollár et al., 2005], HOG/HOF (Histogram of Oriented Gradient/Histogram of Optical Flow) [Laptev et al., 2008], HOG3D [Kläser et al., 2008], extended SURF (ESURF) [Willems et al., 2008], MoSIFT [Chen and Hauptmann, 2009] and Motion Boundary [Wang et al., 2013] descriptors are applied to describe shape and motion in the supported volumes of interest points. A set of D-dimension descriptors extracted from a training set are clustered to form a vocabulary, consisting of K video words. Given an unknown video clip containing an action of interest, all of its descriptors are quantized to the nearest video words in the vocabulary based on a certain distance measure, i.e., Euclidean distance. The histogram of occurrence video words is taken as a compact and holistic representation of the whole video clip, so-called the BoVW representation. Finally,

the BoVW representation is fed to classifiers like SVM to recognize action labels.

The advantages of BoVW approaches to human action recognition include the following.

1. The BoVW representation produces a fixed length vector irrespective of the length of video clips or the speed of performing the action. In other words, BoVW approaches somehow tolerate the timing of action.

2. BoVW approaches are able to handle camera motions since background subtraction is not made use of. Recognizing actions from mobile cameras and in movie sequences become tractable and more accurate than conventional approaches that require foreground images.

3. BoVW approaches also produce stunning performance when tested on KTH dataset, for example, 94.15 % in [Liu and Shah, 2008] and 96.33 % in [Gao et al., 2010] and absolute accuracy when tested on Weizmann dataset [Ikizler and Duygulu, 2007]. They are considered as the most basic common datasets in human action recognition.

However, BoVW approaches also expose several limitations and problems.

1. It is poor to localize actions in a long video clip. Conventionally, human action recognition algorithms are tested against short and manually segmented video clips containing only one action, performed once or repeatedly.

2. Orderless BoVW approaches ignore spatiotemporal structure between video words that is believed to be distinctive feature to discriminate actions.

3. It seems to be hard to choose an optimal size of vocabulary K.

4. The performance of BoVW approaches against viewpoint invariance and scale invariance is unclear. To the best of our knowledge, [Liu and Shah, 2008] and [Wang et al., 2013] are the first works applying BoVW related approach to ISXMAS, a multiview dataset of human action recognition.

In the literature, BoVW approaches have been reported to test on various common benchmark datasets, i.e., KTH [Dollár et al., 2005; Kläser et al., 2008; Laptev et al., 2008; Liu and Shah, 2008; Niebles et al., 2008; Nowozin et al., 2007; Schuldt et al., 2004], Weizmann [Ikizler and Duygulu, 2007; Kläser et al., 2008; Niebles and Li, 2007; Scovanner et al., 2007], IXMAS [Liu and Shah, 2008], UCF Sports[Wang et al., 2009], Hollywood [Wang et al., 2009], HMDB-51 [Wang et al., 2012] and very recently UCF101 [Soomro et al., 2012]. Among these datasets, KTH and Weizmann are often adopted for benchmarking BoVW approaches although they contain a few action classes and limited challenges and are considered as unrealistic datasets. UCF Sports dataset [Rodriguez et al., 2008] provides more action classes than KTH and Weizmann but focuses mainly on sports. Hollywood action datasets [Laptev et al., 2008] are very challenging, even now, containing actions segmented from movies. Since multiple cameras shot an action simultaneously and these image sequences are concatenated in various ways to produce an action clip in movies, camera viewpoint variations in Hollywood datasets are extremely challenging. Recently, a research group in University of Central Florida and a collaborated team between Karlsruhe Institute of Technology, MIT, and Brown University attempt to add more action classes to produce UCF-101 [Soomro et al., 2012] and HMDB-51 [Kuehne et al., 2011] datasets, respectively, mainly collected from YouTube and movies, for benchmarking action recognition algorithms.

Although fall action is included in HMDB-51, actions of daily living like carrying objects, rearranging furniture, changing cloths, and doing housework, etc. are missing. There are several datasets, devoted to actions/activities of daily living but fall actions are omitted, for instance, URADL dataset [Messing et al., 2009], MPII Cooking dataset [Rohrbach et al., 2012], and TUM Kitchen Dataset [Tenorth et al., 2009]. In addition, like-fall actions are also not included in all of them. Therefore in the context of fall detection, accessing performance of BoVW approaches to recognize fall actions based on the reported results tested on these datasets is quite subjective. To the best of our knowledge, there is no research work accessing the performance of BoVW approaches by carrying out experiments on a dataset, exclusively dedicated

to fall detection. Hence this chapter studies the effectiveness of BoVW approaches to discriminate fall actions from other actions of daily living by evaluating the standard BoVW approach and nonlinear SVM classifiers against the common "multiple camera fall dataset" [Auvinet et al., 2010]. Since the dataset is multiple-view, we are able to evaluate the view-invariance fall recognition performance of BoVW approach.

We continue this chapter by summarizing common datasets for human action recognition, particularly ones dedicated to actions of daily living in section 5.2. Our aim is to highlight the fact that fall actions are treated separately from other actions of daily living, in the context of dataset creation. Although BoVW approaches have demonstrated good performance on these common benchmark datasets, it is subjective to draw a similar good performance of BoVW approaches to fall detection. Section 5.3 recaps the common pipeline of BoVW approaches to human action recognition and its variants. Our experimental setup and results are reported in section 5.4. The conclusion of this chapter comes in the section 5.5.

5.2 Common human action recognition datasets

A variety of datasets have been created for benchmarking human action recognition algorithms which were comprehensively reviewed in [Chaquet et al., 2013]. Dataset creation plays a decisive role in both benchmarking performances and motivating the growing of algorithms. Chronological orders of the datasets reflect newly added challenges or increasing difficult levels with which the scientific community are dealing. This section recaps common human action recognition datasets chronologically, particularly ones devoted to actions of daily living. We aim at highlighting the fact that fall actions are often treated separately from other actions of daily living, in the context of dataset creation. Hence we need to study the effectiveness of BoVW approaches to fall detection by empirical experiments on a dataset consisting of both fall and actions of daily living, particularly the like-fall ones even though BoVW approaches have shown good performance on a wide range of common datasets.

5.2.1 Datasets of heterogeneous actions

1. KTH and Weizmann datasets

   The most common datasets which were first publicly introduced to the scientific community include KTH in 2004 [Schuldt et al., 2004] and Weizmann in 2005 [Blank et al., 2005]. Their creation marked a milestone in the development of automated methods of recognizing simple actions in videos. Although many state-of-the-art methods adopt them in evaluation, they are still considered as unrealistic datasets because of containing a few action classes and being collected in controlled environments. KTH takes 6 actions, i.e., walking, jogging, running, boxing, hand waving and hand clapping into account and Weizmann considers 10 actions of walking, running, jumping, galloping sideways, bending, one-hand waving, two-hand waving, jumping in place, jumping jack, and skipping. The actions were recorded by single static cameras with homogeneous background to foster the development of foreground-based algorithms. There is only one person acting in each short video sample whose length is just 3 or 4 seconds. Other common challenges in video-based human action recognition like illumination changes, human appearance variations, intraclass of actions variations, cluttered background, and camera motions, etc. are not taken into consideration even though the robustness set of Weizmann dataset includes such slight variations, like non-homogeneous background, different clothing, and intraclass of actions variations.

2. IXMAS

   In 2006, INRIA introduced IXMAS, a multiview dataset for view-invariant human action recognition [INRIA, 2006]. 13 actions e.g., check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw (over head), and throw (from bottom up) are included in IXMAS. The actions were performed 3 times by 11 actors in various positions and orientations with respect to 5 static cameras in laboratory environments. That is, it is also considered as unrealistic datasets. This dataset was designed to

encourage the research in multiview approaches to human action recognition even though it also provided alternative benchmarks for single view approaches by using one of its five views.

3. UCF101

   A research team in University of Central Florida have been producing a series of datasets for human action recognition in the wake of realizing restrictions of unrealistic datasets to the development of human action recognition algorithms. UCF sports action dataset was first introduced in 2008 [Rodriguez et al., 2008], focusing on a set of 9 actions in sports, i.e., driving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging, and walking. The samples were gathered from television channels like BBC and ESPN rather than hand-shot in controlled environments. The dataset creation was believed to inspire the development of human action recognition algorithms in consideration of realistic samples in unconstrained conditions.

   Following the success of UCF sports, UCF YouTube action or UCF11 dataset [Liu et al., 2009] was created one year later in 2009 by harvesting videos from YouTube, the most popular video-sharing website. As its name indicates, the dataset composes of 11 actions such as basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. Since videos in YouTube are believed to be captured under uncontrolled environments, the dataset is very challenging in terms of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, and intraclass of actions.

   Since then, the authors have been continuing to add more action classes in an attempt of delivering the most challenging and largest dataset to date, resulting in two extensions of UCF11 dataset which are UCF50 [Reddy and Shah, 2013] and UCF101 [Soomro et al., 2012]. The main motivation is to foster further study in action recognition by learning and discovering new realistic

action classes. As its names indicates, UCF50 and UCF101 datasets have 50 and 101 action classes. Specifically, UCF101 dataset contains 13320 videos from 101 action classes such as Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball Shooting, Basketball Dunk, Bench Press, Biking, Billiards Shot, Blow Dry Hair, Blowing Candles, Body Weight Squats, Bowling, Boxing Punching Bag, Boxing Speed Bag, Breaststroke, Brushing Teeth, Clean and Jerk, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Diving, Drumming, Fencing, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Haircut, Hammer Throw, Hammering, Handstand Pushups, Handstand Walking, Head Massage, High Jump, Horse Race, Horse Riding, Hula Hoop, Ice Dancing, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Knitting, Long Jump, Lunges, Military Parade, Mixing Batter, Mopping Floor, Nun chucks, Parallel Bars, Pizza Tossing, Playing Guitar, Playing Piano, Playing Tabla, Playing Violin, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Sitar, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rafting, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Shaving Beard, Shotput, Skate Boarding, Skiing, Skijet, Sky Diving, Soccer Juggling, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Swing, Table Tennis Shot, Tai Chi, Tennis Swing, Throw Discus, Trampoline Jumping, Typing, Uneven Bars, Volleyball Spiking, Walking with a dog, Wall Pushups, Writing On Board, and Yo Yo.

4. HMDB51 [Kuehne et al., 2011]

   The authors were inspired by indefatigable endeavors to create large-scale annotated static image datasets that have been recently revolutionizing the field of object categorization. HMDB51 was produced to meet the urgent need of benchmarking state-of-the-art human action recognition approaches. It was the largest dataset to its released date (2011) with 51 action classes distributed in about 7000 manually annotated clips, collected from movies and YouTube. The

51 action classes are divided into 5 groups

- General facial actions: smile, laugh, chew, and talk.
- Facial actions with object manipulation: smoke, eat, and drink.
- General Human Actions: cartwheel, clap hands, climb, climb stairs, dive, *fall on the floor*, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave.
- Human-Object Interactive Actions: brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, and throw.
- Human-Human Interactive Actions: fencing, hug, kick someone, kiss, punch, shake hands, and sword fight.

5.2.2 Datasets of activities of daily living

Besides datasets of heterogeneous actions, various research works concern about specific actions. Among existing datasets of specific actions, URADL [Messing et al., 2009], TUM Kitchen [Tenorth et al., 2009], and MPII Cooking [Rohrbach et al., 2012] datasets focus on a variety of actions of daily living.

1. URADL (University of Rochester Activities of Daily Living) dataset [Messing et al., 2009]

   URADL is a specific dataset dedicated to activities of daily living. it provides 10 activities of daily living, e.g., answer phone, dial phone, look up phone book, drink water, eat banana, eat snack, peel banana, chop banana, use silverware, and write on white board. Each performed 3 times by 5 different people. Although the dataset considers activities of daily living, it seems to focus on very specifically, for example, eating specific things like banana or snack. In this regard, the number of action classes is very limited.

2. TUM Kitchen dataset [Tenorth et al., 2009]

   TUM kitchen dataset introduced 10 activities of daily living, but restricted in kitchen spaces, for instance, reaching, reaching up, taking something, lowering an object, releasing grasp, opening a door, closing a door, opening a drawer, closing a drawer, and carrying. Challenges posed by this dataset include variation in performing activities, continuous motion (various actions are performed continuously), parallelism (actions are performed by either left or right hands), and body-size differences. Moreover, four different types of data are provided such as video data from 4 overhead static cameras, motion capture data, RFID tag, and magnetic sensor data.

3. MPII cooking activity dataset [Rohrbach et al., 2012]

   Similar to TUM kitchen dataset, MPII cooking activity dataset covers 65 cooking activities, continuously captured in a realistic setting. 12 different actors participated in a realistic recording process in which they prepared one to six out of 14 dishes, containing a varieties of cooking activities such as cut slices, take out from drawers, cut dice, take out from refrigerator, squeeze, peel, wash objects, and grate, etc. In total, 44 videos whose total length is over 8 hours were produced.

According to the above review of datasets, we come up with the following three observations. Firstly among datasets of heterogeneous actions, fall actions are included only in HMDB51 dataset (fall-on-the-floor category). Although HMDB51 dataset also contains some actions of daily living such as stand up, sit down, and walk, etc., the number of action classes is very limited. Moreover, the subjects in this dataset are mostly young people rather than the elderly. Secondly, three datasets, URADL, TUM Kitchen, and MPII cooking activity datasets address various actions of daily living, particularly ones in kitchen spaces, but omit fall actions. Finally, BoVW approaches were tested on these datasets and have demonstrated good recognition accuracy. However, in the context of fall detection, it is subjective to draw a similar good performance of BoVW approaches based on such reported results in

the literature. To the best of our knowledge, there is no empirical study accessing performance of BoVW approaches to discriminating fall actions from other actions of daily living, i.e., stand up, sit down, lie down, do housework, take off cloths, put on cloth, and carry objects, etc. Hence, in this chapter, we evaluate the effectiveness of BoVW approaches to fall detection and compare its performance with state-of-the-art methods, tested on "multiple camera fall dataset", in terms of recognition accuracy and computational cost.

5.3 Bag of Video Word Approaches to Fall Detection

In this section, we describe the standard BoVW approach and nonlinear SVM classifiers for recognizing fall actions. Fig. 5-1 shows the common pipeline of BoVW approaches to human action recognition. Variants of BoVW approaches in the literature differ each other in the selection of local interest point detectors, descriptors and encoding methods. The selections have significant influence on overall performance that will be discussed in the following.

5.3.1 Local spatio-temporal interest point detectors and descriptors

In 2D domain, local interest points are referred to any point in the image where image values vary spatially significantly (both dimensions) [Schmid et al., 2000], for example, corner points. Local interest points contain rich information of local image structure, described by feature descriptors which are computed by using image measurement in the supported regions of these points, for example, histogram of oriented gradients. A compact representation of an image can be efficiently produced by using feature descriptors.

Similarly to produce a compact representation of video data and interpret spatio-temporal events, we need to extend spatial interest points into 3D spatio-temporal domains of video data [Laptev, 2005] by treating video as a sequence of images.

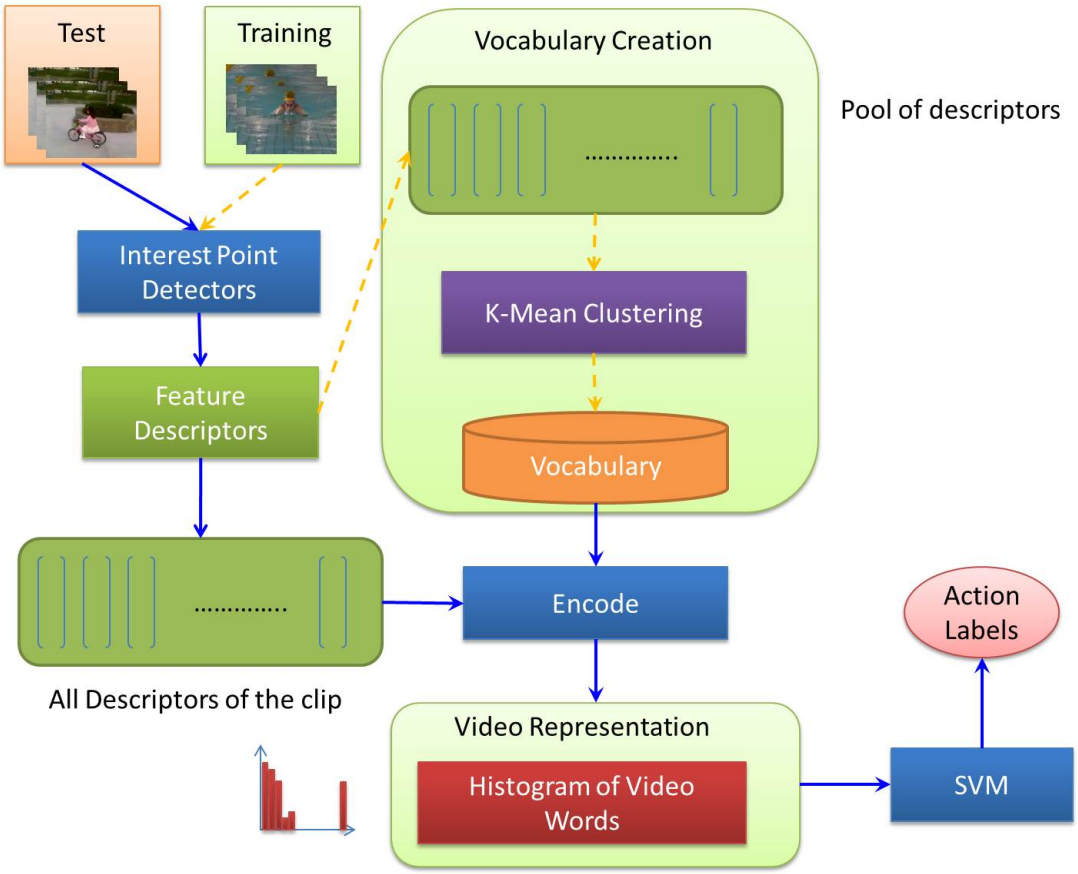


Figure 5-1: The common pipeline of BoVW approach to human action recognition

Spatio-temporal interest points capture shape and motion of video data. Various spatio-temporal interest point detectors have been proposed in the literature by taking a variety of such extensions into account. In this section, we recapitulate some prominent spatio-temporal interest point detectors and descriptors in human action recognition.

1. Harris3D detector and HOG/HOF descriptor [Laptev, 2005; Laptev et al., 2008]

   Ivan Laptev was the first to propose extension of Harris corner detector in 2D image into space-time domain, so-called Harris3D detector. Harris3D detector selects a point with high variation of the intensity in space and non-constant motion in time, determined at multiple spatial and temporal scales. However, this assumption discards spatially salient points containing some certain sorts of motion like periodic motion.

To describe motion and appearance in the 3D neighborhood of detected interest points (space-time supported volume), he also introduced HOG/HOF descriptors [Laptev et al., 2008]. Each volume, centered at detected interest points, is divided into a (n_x, n_y, n_t) grid of cuboids for computing coarse histogram of oriented gradients (HOG) and histogram of Optical Flow (HOF). These histograms are normalized and concatenated into a single HOG/HOF descriptor vector.

2. Cuboid detector and descriptor [Dollár et al., 2005]

   Dollár *et al.* considered a stack of image $I(x, y, t)$ for localizing interest points in spatio-temporal domain rather than $I(x, y)$ in spatial domain. He criticized the Harris3D detector for being unable to detect spatial corners points containing periodic or gradually changed motion. He argued that interest points must be not only salient in space but also have temporal extent. Hence, he concentrated on the temporal domain by applying 1D Gabor filter temporally to detect periodic frequency components. It is showed that Cuboid detectors can also response strongly in as the same range of motion as Harris3D detectors do. As a result, cuboid detector produces more local interest points but is more computationally expensive than Harris3D detector.

   Dollár *et al.* also presented Cuboid descriptor by using local histograms of gradient, inspired by 2D SIFT descriptor [Lowe, 2004]. The cuboid is divided into a number of regions for computing local histograms of each region which are subsequently concatenated into a single vector. PCA is applied to reduce the dimensionality of the final descriptor.

3. Hessian detector and extended SURF descriptor [Willems et al., 2008]

   Willems *et al.* determined spatio-temporal interest points at a certain scale based on the determinant of the 3D Hessian matrix, so-called Hessian detector. It is considered as the spatio-temporal extension of the saliency measure for blob detection in [Beaudet, 1978]. In contrast to Harris3D detector whose scale is selected by iterative manners, the scale-normalized determinant of the Hessian

facilitates both scale invariance and good scale selection simultaneously without iteration. Hence, Hessian detector is the most efficient but produces sparest features among three detectors reviewed so far [Wang et al., 2009].

An extension of 2D SURF descriptor [Bay et al., 2006] into spatio-temporal domain, so-called extended SURF descriptor for video, was also proposed in [Willems et al., 2008]. Like these above methods, each space-time supported volume is also divided into (n_x, n_y, n_t) subvolumes but each of which contains the vector $\upsilon = (\sum d_x, \sum d_y, \sum d_t)$ where d_x, d_y, d_t are weighted sums of uniformly sampled responses of Haar-wavelets over space and time.

4. Motion SIFT (MoSIFT) detector and descriptor [Chen and Hauptmann, 2009]

   The authors were inspired by impressive performance of the Scale Invariant Feature Transform (SIFT) [Lowe, 2004] in 2D domain. Their endeavor was to develop a counterpart for detecting video interest points by treating spatial and temporal dimensions separately. SIFT and optical flow are combined to form a motion-based feature, so-called Motion SIFT or MoSIFT. A SIFT point is selected as a feature point if optical flow near the point is large enough. In other words, MoSIFT point is a SIFT point containing significant motion.

   A MoSIFT descriptor is also proposed to represent (1) spatial appearance of the feature point by an aggregated histogram of gradients and (2) the motion of the feature point by an aggregated histogram of optical flow. The aggregation of histograms makes the descriptor more invariant to any deformation.

5. HoG3D descriptor [Kläser et al., 2008]

   HOG3D is constructed based on histograms of oriented 3D spatio-temporal gradients that are computed memory-efficiently by using integral videos [Kläser et al., 2008]. Eventually, the authors extended integral images, proposed by Viola and Jones [Viola and Jones, 2001] for efficient computation of Haar features of an image, to integral videos for efficient computation of 3D gradient vectors. Regular polyhedrons, i.e., dodecahedron (12-sided) and icosahedron

(20-sided) are adopted for gradient quantization. Similar to the computation of above descriptors, space-time volumes are divided into (n_x, n_y, n_t) sub-volumes in which local orientation histograms are computed before concatenated to one feature vector.

6. Dense sampling

   Video data is divided into regular stacks of images or volumes of images at multiple spatial and temporal scales. A descriptor is applied to each stack to describe its shape and motion. Video representation based on these dense sampling incorporates not only shape and motion of objects but also context information surrounding the objects. It is argued that some actions take place in particular environments, for example, swimming in a swimming pool and sailing in a lake or a river, etc. In such cases, context information play an important role in distinguishing actions. It has been demonstrated empirically in [Wang et al., 2009] that dense sampling outperforms other interest point detectors, i.e., Harris3D, Cuboid, and Hessian in some datasets, like Hollywood-II and UCF datasets. But it performs less effective in KTH dataset than the others. The reason is that background images in the two former datasets change considerably between actions. Context information provided by dense features has effect to raise performance. While all actions in KTH dataset took place in relatively same plain background, performance of dense features is poorer than that of the others.

   In the context of fall detection, surveillance cameras monitor a person of interest living in their home in which all of his/her actions take place. It means that the background does not change much between actions. Dense sampling seems to be inappropriate for recognizing falls.

7. Dense trajectories and motion boundary descriptors [Wang et al., 2013]

   Inspired by the standout performance of dense sampling in action recognition, [Wang et al., 2013] extract dense features and track them by using dense optical flow algorithm to produce a dense trajectories-based video representation. This

representation contains rich foreground motion and context information. Motion Boundary Histogram (MBH) descriptor for human detection proposed by [Dalal et al., 2006] is employed as motion descriptor for dense trajectories. Since MBH relies on differential optical flow, locally constant camera motion is removed but the motion boundaries are retained. Dense trajectories and motion boundary descriptors are considered as the state-of-the-art in action recognition by outperforming the other methods on nine popular action datasets like Hollywood2, UCF50, and HMDB51, etc.

In this chapter, we empirically study the effectiveness of BoVW approaches to fall detection by performing an experiment using a BoVW approach which is widely accepted as a standard in action recognition [Soomro et al., 2012] to provide the baseline results on Multiple Camera Fall Dataset. Harris3D detector and HOF/HOF descriptors are employed in this standard approach.

5.3.2 Encoding and Pooling methods

Various encoding and pooling methods have been proposed in the literature, for example, vector quantization [Csurka et al., 2004], soft-assignment encoding [Gemert et al., 2008], sparse encoding [Yang et al., 2009], locality-constrained linear encoding [Wang et al., 2010], Fisher Kernel encoding [Perronnin et al., 2010], sum pooling [Lazebnik et al., 2006] and max pooling [Yang et al., 2009]. The selection of these methods in the pipeline of BoVW approaches have great influence on final performance, in terms of accuracy and computational cost as described in an empirically comparative study by Wang *et al.* [Wang et al., 2012]. In our experiment, we follow the standard evaluation protocol in [Wang et al., 2009] by using vector quantization encoding and sum pooling methods.

Suppose that we use D-dimensional descriptor to describe shape and motion in the supported volumes of N local spatio-temporal interest points, extracted from a video. That is, we have a set $X = [x_1, x_2, ..., x_i, ..., x_N] \in \mathbb{R}^{D \times N}$ of local descriptors. Also suppose that we are given a vocabulary composing of K video words, i.e., $V =$

$[v_1, v_2, ..., v_j, ..., v_K] \in \mathbb{R}^{D \times K}$. The objective of encoding method is to find a word v_j in the set V as a representation of a descriptor x_i in the set X. Vector quantization which is known as Hard-assignment coding, assigns each descriptor x_i to the nearest video word v_j based on a distance measure, i.e., Euclidean distance. Let denote C as code matrix

$$C = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,j} & \cdots & c_{1,K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{i,1} & c_{i,2} & \cdots & c_{i,j} & \cdots & c_{i,K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{N,1} & c_{N,2} & \ddots & c_{N,j} & \ddots & c_{N,K} \end{pmatrix} \in \mathbb{R}^{N \times K} \tag{5.1}$$

In vector quantization encoding method, element $c_{i,j}$ of code matrix C is determined by

$$c_{i,j} = \begin{cases} 1 & \text{if } j = argmin_j ||x_i - v_j|| \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

where $||.||$ denotes Euclidean distance. We take histograms of occurrence words or take sum pooling method to create a feature vector $p \in \mathbb{R}^{1 \times K}$ as a holistic representation of the video. Each element p_j of the feature vector p is determined by

$$p_j = \sum_{i=1}^{N} c_{i,j} \tag{5.3}$$

Finally, the feature vector is normalized by using ℓ_2 norm.

$$p_j = \frac{p_j}{\sum_{k=1}^{K} \sqrt{p_k^2}} \tag{5.4}$$

5.4 Experiments and performance evaluation

In this section, we carry out our experiments on "multiple camera fall dataset" to evaluate the effectiveness of BoVW approach in discriminating falls from other actions

of daily living. We follow the common evaluation protocol, proposed in [Wang et al., 2009]. STIP and its HOG/HOF descriptors are extracted from all video samples of the dataset, resulting in a huge set of STIP and its descriptors. We use the executable of Harris3D detector, provided by Ivan Laptev [Laptev, 2005], run in Linux environment. The STIP and its HOG/HOF descriptors are saved into text files which are imported into Matlab for further processing. We randomly choose 100,000 STIP and its descriptors from the original set for the vocabulary creation by using K-mean clustering. Since actions of daily living are dominant in the dataset, we must carefully choose enough STIP and its descriptors, extracted from sequences containing fall actions. To this end, we manually classify STIP and its descriptors in the original set into two subsets. One is extracted from sequences containing fall actions and the other from sequences containing actions of daily living. Subsequently, we perform choosing 100,000 STIP and its descriptors randomly from these two subsets for the vocabulary creation. Here we set the number of video words in the vocabulary V to 4000 that has empirically demonstrated to produce good results for a variety of datasets [Wang et al., 2009]. We adopt vector quantization encoding method with Euclidean distance, sum pooling, and ℓ_2 norm to determine the final feature vector.

Since each video sample of the dataset composes of a series of actions, probably including falls, we use sliding-window method to locate fall actions. As figured out in [Noury et al., 2008], a fall usually lasts within one second. Therefore, a window of 30 frames with a sliding step of two frames is adopted in our implementation. That is, the window with the length of 30 frames is shifted along the time axis every two frames

Table 5.1: Experimental Results of BoVW approach to fall detection against "multiple cameras fall dataset." In the table, a fall, correctly detected, is denoted by 1. A fall, not detected, is denoted by -1. Since there is no fall incident in the scene 24, BoVW approach does not produce any false alarm. Consequently, the column of scene 24 is left blank.

		Scenes																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23		24
Views	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	-1	1	1	-1	-1	
	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	
	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	-1	1	1	
	5	1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	-1	-1	-1	1	-1	1	-1	-1	1	
	6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	-1	1	1	1	-1	
	7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	1	1	1	
	8	1	1	1	1	1	1	1	-1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	-1	-1	

The classification is performed by using non-linear two-class SVM [Vedaldi and Fulkerson, 2008] with Chi-squared kernel since with vocabulary/histogram representations, Chi-square kernel performs better than other kernels like linear, quadratic and Radial Basis Function (RBF) kernels [Zhang et al., 2007]. All actions of daily living are combined into one class, that is, the non-fall class. The other one is fall class.

We perform leave-one-out cross validation in the whole dataset. Since the dataset composes of 24 scenarios, captured simultaneously from 8 cameras. We leave one scenario (8 video samples captured from 8 views) intact for testing and use the other video samples in the rest of 23 scenarios for training. Such experiments are repeated 24 times in order to complete the leave-one-out cross validation. By doing that, we are also able to evaluate performance of BoVW approach to fall recognition across multiple views.

To evaluate the recognition accuracy, we manually annotate the ground truth for each video sample. The ground truth in this context is a time window of occurring a fall. That is, we manually measure the first frame when the person starts falling and the last frame when the person starts hitting the ground. If the SVM classifier responses positively within or very near the ground truth window, a fall is claimed. The results of our experiments are reported in Table 5.1.

Specificity and *sensitivity* of BoVW approach are also shown in Table. 5.2. In comparison with state-of-the-art methods [Hung et al., 2013; Hung and Saito, 2013; Auvinet et al., 2011; Rougier et al., 2007b, 2011b], tested on the same dataset, the performance of BoVW approach is comparable (please compare with Table 4.1). However, the computational cost of BoVW approach is more expensive than the others. For localizing STIP and determining HOG/HOF descriptors, the average frame rate is about 1.44 fps, by using the executable provided by the author in [Laptev, 2005]. That is, the real-time performance of BoVW approaches seems to be poor for practical usage of fall detection.

Table 5.2: Specificity and Sensitivity of BoVW approach, tested on "multiple cameras fall dataset."

		Sensitivity	Specificity
Views	1	83.3%	100%
	2	95.8%	100%
	3	100%	100%
	4	91.7%	100%
	5	70.8%	100%
	6	87.5%	100%
	7	95.8%	100%
	8	83.3%	100%

5.5 Conclusions

In this chapter, various common human action recognition datasets are summarized to highlight the fact that fall actions are treated separately from actions of daily living in the context of dataset creation. Despite good performance of BoVW approaches as reported on a wide range of such datasets in the literature, it is subjective to draw a similar good performance of BoVW approaches to discriminating fall actions from other actions of daily living. Hence, we have presented an empirical study to evaluate the effectiveness of BoVW approaches to fall detection. Experiments of BoVW approach with STIP, HOG/HOF descriptors and nonlinear Chi-Square kernel SVM classifier on "multiple cameras fall dataset" are carried out. The recognition accuracy of BoVW approach, in term of specificity and sensitivity is comparable with that of state-of-the-art methods [Hung et al., 2013; Hung and Saito, 2013; Auvinet et al., 2011; Rougier et al., 2007b, 2011b], tested on the same dataset. However, the computational cost of BoVW approach is more expensive than the others.

Chapter 6

Conclusions and future works

6.1 Dissertation conclusions

In this dissertation, we have addressed the problem of detecting fall incidents in an effort to support the elderly living alone safely at home. It is figured out that accidental falls are the most common cause of injuries for the elderly [Yu, 2008] and the sixth leading cause of death [MacCulloch et al., 2007]. Our research outcomes are capable of helping the elderly reach the instant treatment just after suffering from accidental falls, in turn, not worsening their injuries or even saving their lives.

Detecting fall incidents by using vision technology is challenging. On the one hand, we must discriminate falls carefully from usual activities, particularly the confounding or like-fall ones. The accuracy of fall detection methods must be high since it relates to the human safety. The false alarms also must be kept as low as possible otherwise it may bother emergency response centers which are always ready to offer immediate helps to fallen people. On the other hand, we need to handle other common challenges of vision technology, such as low image quality, viewpoint variations, illumination variations, cluttered background, occlusion by furniture and real-time implementation.

To tackle these challenges, we have introduced in this dissertation the 3D spatial features, i.e., the combination of heights and occupied areas, extracted from 3D cuboids or 3D bounding boxes of the person of interest and Human-Ground Contact

Areas (HGCA). We have demonstrated the effectiveness of these 3D spatial features in discriminating falls from other usual activities, in terms of both high detected rate, low false alarms and real-time processing. We argue that people in lying states occupy larger areas than those in sitting and standing states. The heights of standing people are greater than that of sitting and lying people. Therefore, the combination of heights and occupied areas are highly discriminative in classifying human states to perform fall inference. We also argue that people always make contact with the ground during usual activities mainly by the feet. People often lie completely on the ground after suffering from accidental falls. Therefore, contact areas between human and ground contain rich information to distinguish human states, leading to our proposal of Human-Ground Contact Areas. Falls are discriminated from usual activities by analyzing human state transition.

In implementation, we configure two cameras whose fields of view are relatively orthogonal to simplify the 3D cuboid reconstruction. 2D bounding boxes extracted from two views are served as two orthographic projections of the 3D cuboid. As a result, the 3D cuboid reconstruction are very straightforward. We also suggest using Local Empirical Templates to normalize the reconstructed 3D cuboids in order to make them view-invariant across the viewing windows, facilitating the human state classification. To estimate HGCA, we propose projecting foreground across views by using homography of the ground between views. There exist overlap regions between foregrounds where people and the ground are in contact, i.e., feet during usual activities and almost whole body after falls. We carry out experiments of our approaches on a common dataset of fall detection, "multiple camera fall dataset," demonstrating favorably comparable performance with state-of-the-art methods [Auvinet et al., 2011; Rougier et al., 2007b, 2011b], tested on the same dataset, but with lower computational cost. Our solutions are low-cost and can be run in real-time by a common PC with standard resolution video surveillance cameras, i.e., 320x240.

In conclusions, our major contributions are to introduce new 3D spatial features and low-cost schemes of estimating these features which are good for fall detection, as the summarization below.

1. The combination of heights and occupied areas.

    - Proposal of using two cameras whose fields of view are relatively orthogonal to approximate the person of interest of 3D cuboids or 3D bounding boxes.
    - Suggestion of using Local Empirical Templates to normalize the reconstructed 3D cuboids, making them view-invariant across the viewing windows, facilitating the human state classification.

2. Human-Ground Contact Areas

    - Proposal of using foreground projection across views based on the homography of the ground between views and measuring overlap regions between foregrounds as HGCA that is view-invariant across the viewing windows.
    - Investigation into the relationship between HGCA and human states, i.e., standing, sitting, and lying in consideration of various poses and viewpoints via human state simulation.

In addition, the third contribution of this dissertation comes from our empirical study of the effectiveness of BoVW approaches to fall detection. Since BoVW approaches have demonstrated good performance on a wide range of common datasets of human actions, its performance in discriminating fall actions and other actions of daily living is unknown. We have summarized various common datasets of human actions to figure out the fact that fall actions are often treated separately from other actions of daily living, in the context of dataset creation. Despite good performances of BoVW approaches on a variety of common datasets as reported in the literature, it is subjective to draw its similar good performance in fall detection in consideration of the above fact. To the best of our knowledge, there is no research work evaluating the performance of BoVW approaches on a dataset, exclusively dedicated to fall detection. That is, the dataset must contain both fall actions and a variety of actions of daily living. Hence in this dissertation, we carry out an experiments of a standard BoVW approach with nonlinear Chi-square kernel SVM classifier on "multiple

camera fall dataset" by using a common PC. We conclude that BoVW approaches produce favorably comparable performance with state-of-the-art methods (including our proposed solutions), but at the expense of expensive computational cost.

We do hope that our research outcomes will contribute considerably a step toward the commercialization of vision-based fall detection technology, in particular and video surveillance systems for healthcare applications, in general. Such systems will not only bring huge benefits to the elderly but also bring smile back to the faces of their families, caregivers, and the governments. The quality of life, the quality of healthcare, the autonomy, the freedom, and the safety, etc. will be ameliorated in an comfortable way since vision technology has been increasingly accepting by a majority of the elderly community.

6.2 Future directions to fall detection

In this section, we delineate several research directions to fall detection.

6.2.1 Creation of benchmark dataset

The main application of fall detection methods is to help the elderly live alone at home in safety. Benchmarking fall detection methods by real falls of the real elderly is much better than by simulated falls performed by young actors. The creation of a benchmark dataset containing real falls of the real elderly in real home environments will evaluate fall detection methods precisely and promote its development significantly, even though as we mentioned in Sect. 2.2.1 that it is a daunting task. To the best of our knowledge, such real datasets are not publicly available. People, especially the elderly, may be hesitated to appear in the publicly available datasets.

6.2.2 Fall detection on a mobile robot

A common assumption of fall detection methods is the usage of stationary cameras so that foreground segmentation can be easily performed to detect moving people.

Stationary cameras facilitate a common trend of analyzing silhouettes for fall detection, as illustrated in Sect. 2.3. However, we can relax this assumption by using BoVW approaches as described in chapter 5, since BoVW approaches have demonstrated good performance on datasets containing camera motions, like UCF101 and HMDB51. Although BoVW approaches have expensive computational cost, the advances of GPU may help facilitate the real-time performance. Once the assumption of stationary cameras is relaxed, it leads to an interesting applications of detecting falls from a mobile robot. We can integrate the functionality of safety guards for assistive social robots which have been recently developing to provide social interaction to the elderly living alone, in order to relieve the depression and isolation [Broekens et al., 2009]. Moreover, mobile robots are able to travel various rooms to follow the elderly and monitor them, rather than placing networks of cameras in multiple rooms. The combination of a Kinect and a mobile robot seems to be a promising solution. On the one hand, the mobility of the robot help us overcome the depth range limitation. Both color and depth image sequences can be used to locate and describe interest points. On the other hand, Kinect is able to work in low-lighting conditions, especially at night without ambient light.

6.2.3 Multiple-target fall detection

The fall detection methods proposed in this dissertation is very suitable to the elderly living alone at home. That is, the methods only cope with the presence of single person. In section 1.2, we argue that it is meaningless to use fall detection method in an environment with two-people presence. When one person falls down, the other one will easily detect the incident. However, the situation becomes dangerous if the two people are unresponsive or slowly responsive. Even the person witnesses the incident but is unable to response or help the victim immediately. This case requires the extension of fall detection methods to cope with multiple targets. We need to integrate multiple-target tracking algorithms into fall detection methods and take human occlusion into consideration. In chapter 3, we should build the ground plane model in 3D world and make the projections of the positions of 3D cuboids into the

ground plane model for tracking and computing the features. It is straightforward to extend the method in chapter 4 since multiple-target tracking by using planar homography of the ground between views was well studied by [Khan and Shah, 2006]. They are some possible ways of extending and improving the quality of our works as well as contributing to the development of fall detection methods.

Bibliography

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 *International Conference on Information and Communication Technologies: From Theory to Applications*, pages 1003 – 1007.

Anderson, D., Keller, J. M., Skubic, M., Chen, X., and He, Z. H. (2006). Recognizing falls from silhouettes. In *IEEE International Conference on Engineering in Medicine and Biology Society*, pages 6388 – 6391.

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*, 113:80 – 89.

Auvinet, E., Mullon, F., St-Arnaud, A., Rousseau, J., and Meunier, J. (2011). Fall detection with multiple cameras: An occlusion-resistant method based on 3d silhouette vertical distribution. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):290 – 300.

Auvinet, E., Reveret, L., St-Arnaud, A., Rousseau, J., and Meunier, J. (2008). Fall detection using multiple cameras. In *IEEE International Conference on Engineering in Medicine and Biology Society*, pages 2554 – 2557.

Auvinet, E., Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2010). Multiple cameras fall dataset. Technical report, Universitè de Montreal.

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf: Speeded up robust features. In *European Conference on Computer Vision*, volume 1, pages 404 – 417.

Beaudet, P. (1978). Rotationally invariant image operators. In *International Joint Conference on Pattern Recognition*, pages 579 – 583.

Benezeth, Y., Jodoin, P. M., Emile, B., Laurent, H., and Rosenberger, C. (2010). Comparative study of background subtraction algorithms. *Journal of Electronic Imaging*, 19(3):1–12.

Bharucha, A. J., Anand, V., Forlizzi, J., Dew, M. A., Reynolds, C. F., Stevens, S., and Wactlar, H. (2009). Intelligent assistive technology applications to dementia care: current capabilities, limitations and future challenges. *American Journal of Geriatric Psychiatry*, 17(2):88 – 104.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *IEEE International Conference on Computer Vision*, pages 1395 – 1402.

Broekens, J., Heerink, M., and Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2):94 – 103.

Butler, A., Izadi, S., Hilliges, O., Molyneaux, D., Hodges, S., and Kim, D. (2012). Shake'n'sense: reducing interference for overlapping structured light depth cameras. In *ACM Conference on Human Factors in Computing Systems, CHI*, pages 1933 – 1936.

Chan, M., Estève, D., Escriba, C., and Campo, E. (2008). A review of smart homes - present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91:55 – 81.

Chaquet, J. M., Carmona, E. J., and Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117:633 – 659.

Charif, H. N. and McKenna, S. J. (2004). Activity summarization and fall detection in a supportive home environment. In *International Conference on Pattern Recognition*, pages 323 – 326.

Chen, M. Y. and Hauptmann, A. (2009). Mosift: Recognizing human actions in surveillance videos. Technical report, Carnegie Mellon University.

Chen, X., He, Z., Keller, J. M., Anderson, D., and Skubic, M. (2006). Adaptive silhouette extraction and human tracking in complex and dynamic environments. In *IEEE International Conference on Fuzzy Systems*, pages 16 – 21.

Chen, Y. T., Lin, Y. C., and Fang, W. H. (2010). A hybrid human fall detection scheme. In *International Conference on Image Processing*, pages 3485 – 3488.

Cowan, D. and Turner-Smith, A. (1999). The role of assistive technology in alternative models of care for older people. *Royal Commission on Long Term Care*, 2(4):325 – 346.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, pages 1–22.

Cucchiara, R., Grana, C., Neri, G., Piccardi, M., and Prati, A. (2002). The sakbot system for moving object detection and tracking. In *Video-based surveillance systems*, pages 145 – 157. Springer.

Cucchiara, R., Grana, C., Piccardi, M., Prati, A., and Sirotti, S. (2001). Improving shadow suppression in moving object detection with hsv color information. In *IEEE International Conference on Intelligent Transportation Systems*, pages 334 – 339.

Cucchiara, R., Grana, C., Prati, A., and Vezzani, R. (2005). Probabilistic posture classification for human behavior analysis. *IEEE Transactions on Systems, Man, and Cybernestics, Part A: Systems and Humans*, 35(1):42 – 54.

Cucchiara, R., Prati, A., and Vezzani, R. (2007). A multi-camera vision system for fall detection and alarm generation. *Expert Systems*, 24(5):334 – 345.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histogram of flow and appearance. In *European Conference on Computer Vision*, volume 2, pages 428 – 441.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*.

Doughty, K., Cameron, K., and Garner, P. (1996). Three generations of telecare of the elderly. *Journal of Telemedicine and Telecare*, 2(2):71 – 80.

Doukas, C., Maglogiannis, I., Tragas, P., Liapis, D., and Yovanof, G. (2007). Patient fall detection using support vector machines. In *IFIP The International Federation for Information Processing*, volume 247, pages 147 – 156.

Dubey, R., Ni, B., and Moulin, P. (2012). A depth camera based fall recognition system for the elderly. In *International Conference on Image Analysis and Recognition*, volume 2, pages 106 – 113.

Elgammal, A., Harwood, D., and Davis, L. (2000). Non-parametric model for background subtraction. In *European Conference on Computer Vision*, volume 2, pages 751 – 767.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259 – 275.

Faugeras, O. and Lustman, F. (1988). Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 2:485 – 508.

Gao, Z., Chen, M. Y., Hauptmann, A. G., and Cai, A. (2010). Comparing evaluation protocols on the kth dataset. In *International Workshop on Human Behavior Understanding*, pages 88 – 100.

Gemert, J. C. V., Geusebroek, J. M., Veenman, C. J., and Smeulders, A. W. M. (2008). Kernel codebooks for scene categorization. In *European Conference on Computer Vision*, volume 3, pages 696 – 709.

Ghasemzadeh, H., Jafari, R., and Prabhakaran, B. (2009). A body sensor network with electromyogram and inertial sensors: multi-modal interpretation of muscular activities. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):198–206.

Haritaoglu, I., Harwood, D., and Davis, L. (2000). W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809 – 830.

Hartley, R. I. and Zisserman, A. (2004). *Multiple view geometry in computer vision*. Cambridge University Press, 2nd edition.

Hazelhoff, L., Han, J., and de With, P. H. N. (2008). Video-based fall detection in the home using principal component analysis. In *Advanced Concepts for Intelligent Vision Systems*, pages 298 – 309.

Hossain, M. A. and Ahmed, D. T. (2012). Virtual caregiver: an ambient-aware elderly monitoring system. *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1024 – 1031.

Hossain, M. S., Goebel, S., and Saddik, A. E. (2012). Guest editorial: Multimedia services and technologies for e-health (must-eh). *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1005 – 1006.

Htike, Z. Z., Egerton, S., and Chow, K. Y. (2011). A monocular view-invariant fall detection system for the elderly in assisted home environments. In *International Conference on Intelligent Environments*, pages 40 – 46.

Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernestics, Part C: Applications and Reviews*, 34(3):334 – 352.

Huang, B., Tian, G. H., and Wu, H. (2008). A method for fast fall detection based on intelligent space. In *IEEE International Conference on Automation and Logistics*, pages 2260 – 2265.

Hung, D. H., Chung, S. L., and Hsu, G. S. (2010). Local empirical templates and density ratios for people counting. In *Asian Conference on Computer Vision*, volume 4, pages 90 – 101.

Hung, D. H., Hsu, G. S., Chung, S. L., and Saito, H. (2012). Real-time people counting in crowded areas by using local empirical templates and density ratios. *IEICE Transactions on Information and Systems*, E95-D(7):1791 – 1803.

Hung, D. H. and Saito, H. (2012). Fall detection with two cameras based on occupied areas. In *18th Japan-Korea Joint Workshop on Frontier in Computer Vision*, pages 33 – 39.

Hung, D. H. and Saito, H. (2013). The estimation of heights and occupied areas of humans from two orthogonal views for fall detection. *IEEJ Transactions on Electronics, Information and Systems*, 133(1):117 – 127.

Hung, D. H., Saito, H., and Hsu, G. S. (2013). Detecting fall incidents of the elderly based on human-ground contact areas. In *Asian Conference on Pattern Recognition*.

Ikizler, N. and Duygulu, P. (2007). Human action recognition using distribution of oriented rectangular patches. In *Workshop on Human Motion Understanding, Modeling, Capture and Animation*, pages 271 – 284.

INRIA (2006). Inria xmas motion acquisition sequences. http://4drepository.inrialpes.fr/public/viewgroup/6.

Javed, O. and Shah, M. (2008). *Automated Multicamera Surveillace: Algorithm and Practice*. Springer, 1st edition.

KaewTraKuPong, P. and Bowden, R. (2001). An improved adaptive background mixture model for real-time tracking with shadow detection. In *European Workshop on Advanced Video-Based Surveillance Systems*.

Khan, S. and Shah, M. (2006). A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *European Conference on Computer Vision*, volume 4, pages 133 – 146.

Khan, Z. A. and Sohn, W. (2011). Abnormal human activity recognition system based on r-transform and kernel discriminant technique for elderly home care. *IEEE Transactions on Consumer Electronics*, 57(4):1843 – 1850.

Kharicha, K., Iliffe, S., Harari, D., Swift, C., Gillmann, G., and Stuck, A. E. (2007). Health risk appraisal in older people 1: are older people living alone an "at risk" group? In *British Journal of General Practice*, volume 57, pages 271 – 276.

Kim, K., Chalidabhongse, T. H., Harwood, D., and Davis, L. (2005). Real-time foreground-background segmentation using codebook model. *Journal of Real-Time Imaging*, 11(3):172 – 185.

Kläser, A., Marzalek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d gradients. In *British Machine Vision Conference*, pages 275 – 284.

Koch, S. (2006). Home telehealth - current state and future trends. *International Journal of Medical Informatics*, 75:565 – 576.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). Hmdb - a large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, pages 2556 – 2563.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107 – 123.

Laptev, I., Marzalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1 – 8.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2169 – 2178.

Lee, T. and Mihailidis, A. (2005). An intelligent emergency response system: preliminary development and testing of automated fall detection. *Journal of Telemedicine and Telecare*, 11(4):194 – 198.

Liao, Y. T., Huang, C. L., and Hsu, S. C. (2012). Slip and fall event detection using bayesian belief network. *Pattern Recognition*, 45:24 – 32.

Lin, C. S., Hsu, H. C., Lay, Y. L., Chiu, C. C., and Chao, C. S. (2007). Wearable device for real-time monitoring of human falls. *Measurement*, 40:831 – 840.

Liu, C. L., Lee, C. H., and Lin, P. M. (2010). A fall detection system using k-nearest neightbor classifier. *Expert Systems with Applications*, 37:7174 – 7181.

Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos "in the wild". In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1996 – 2003.

Liu, J. and Shah, M. (2008). Learning human actions via information maximization. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1 – 8.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91 – 100.

Luke, R. H., Anderson, D., Keller, J. M., and Skubic, M. (2008). Moving object segmentation from video using fused color and texture features in indoor environments. *Journal of Real-Time Image Processing, Submitted for publication*.

MacCulloch, P. A., Gardner, T., and Bonner, A. (2007). Comprehensive fall prevention programs across settings: a review of the literature. *Geriatric Nursing*, 28(5):306 – 311.

Mastorakis, G. and Makris, D. (2012). Fall detection system using kinect's infrared sensor. *Journal of Real-Time image Processing*, pages 1–12.

Mathie, M. J., Coster, A. C. F., Lovell, N. H., and Celler, B. G. (2004). Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement. *Physiological Measurement*, 25(2):1 – 20.

McKenna, S. J., Jabri, S., Duric, Z., Rosenfeld, A., and Wechsler, H. (2000). Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42 – 56.

Messing, R., Pal, C., and Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In *IEEE International Conference on Computer Vision*, pages 104 –111.

Mubashir, M., Shao, L., and Seed, L. (2013). A survey on fall detection: principles and approaches. *Neurocomputing*, 100:144 – 152.

Niebles, J. C. and Li, F. F. (2007). A hierarchical model of shape and appearance for human action classification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1 – 8.

Niebles, J. C., Wang, H. C., and Li, F. F. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79:299 – 318.

Noury, N., Fleury, A., Rumeau, P., Bourke, A. K., ÓLaighin, G., Rialle, V., and Lundy, J. E. (2007). Fall detection: principles and methods. In *IEEE International Conference on Engineering in Medicine and Biology Society*, pages 1663 – 1666.

Noury, N., Rumeau, P., Bourke, A. K., ÓLaighin, G., and Lundy, J. E. (2008). A proposal for the classification and evaluation of fall detectors. *Ingénierie et Recherche Biomédicale (IRBM)*, 29:340 – 349.

Nowozin, S., Bakir, G., and Tsuda, K. (2007). Discriminative subsequence mining for action classification. In *IEEE International Conference on Computer Vision*, pages 1 – 8.

Nyan, M. N., Tay, F. E. H., and Seah, K. H. W. (2006). Garment-based detection of falls and activities of daily living using 3-axis mems accelerometer. *Journal of Physics: Conference Series (International MEMS Conference)*, 34:1059–1067.

Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large scale image classification. In *European Conference on Computer Vision*, volume 4, pages 143 – 156.

Piccardi, M. (2004). Background subtraction technique: a review. In *IEEE International Conference on Systems, Man, and Cybernestics*, pages 3099 – 3104.

Planinc, R. and Kampel, M. (2012a). Introducing the use of depth data for fall detection. *Personal and Uniquitous Computing*, 17(6):1063 – 1072.

Planinc, R. and Kampel, M. (2012b). Robust fall detection by combining 3d data and fuzzy logic. In *ACCV Workshop on Color Depth Fusion in Computer Vision*, volume 2, pages 109 – 120.

Pogorelc, B., Bosnić, Z., and Gams, M. (2012). Automatic recognition of gait-related health problems in the elderly using machine learning. *Multimedia Tools and Applications*, 58(2):333 – 354.

Poppe, R. (2010). A survey on vision-based on human action recognition. *Image and Vision Computing*, 28:976 – 990.

Reddy, K. K. and Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971 – 981.

Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1 – 8.

Rohrbach, M., Amin, S., Andriluka, M., and Schiele, B. (2012). A database for fine grained activity detection of cooking activities. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1194 – 1201.

Rougier, C., Auvinet, E., Rousseau, J., Mignotte, M., and Meunier, J. (2011a). Fall detection from depth map video sequences. In *International Conference on Smart Homes and Health Telematics*, pages 121 – 128.

Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2006). Monocular 3d head tracking to detect falls of elderly people. In *IEEE International Conference on Engineering in Medicine and Biology Society*, pages 6384 – 6387.

Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2007a). Fall detection from human shape and motion history using video surveillance. In *International Workshop on Advanced Information Networking and Applications*, pages 875 – 880.

Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2007b). Procrustes shape analysis for fall detection. In *International Workshop on Visual Surveillance*.

Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2011b). Robust video surveillance for fall detection based on human shape deformation. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(5):611 – 622.

Ruggiero, C., Sacile, R., and Giacomini, M. (1999). Home telecare. *Journal of Telemedicine and Telecare*, 5(1):11 – 17.

Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151 – 172.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local svm approach. In *International Conference on Pattern Recognition*, pages 32 – 36.

Scott, T. E. (2000). Bed exit detection apparatus. *US Patent 6067019*.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *ACM International Conference on Multimedia*, pages 357 – 360.

Shoaib, M., Dragon, R., and Ostermann, J. (2010). View-invariant fall detection for elderly in real home environment. In *Pacific-Rim Symposium on Image and Video Technology*, pages 52 – 57.

Shoaib, M., Dragon, R., and Ostermann, J. (2011a). Context-aware visual analysis of elderly activity in a cluttered home environment. *EURASIP Journal on Advances in Signal Processing*, 2011:1 – 14.

Shoaib, M., Dragon, R., and Ostermann, J. (2011b). Fall dataset. http://www.tnt.uni-hannover.de/ shoaib/fall.html.

Sixsmith, A. and Johnson, N. (2004). A smart sensor to detect falls of the elderly. *IEEE Pervasive Computing*, 3(2):42 – 47.

Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human action classes from videos in the wild. Technical report, CRCV-TR-12-01.

Spasova, V. and Iliev, I. (2014). A survey on automatic fall detection in the context of ambient assisted living systems. *International Journal of Advanced Computer Research*, 4-1(14):94 – 109.

Stauffer, C. and Grimson, W. L. R. (2000). Learning patterns of activities using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747 – 757.

Tamura, T., Yoshimura, T., Sekine, M., Uchida, M., and Tanaka, O. (2009). A wearable airbag to prevent fall injuries. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):910–914.

Tenorth, M., Bandouch, J., and Beetz, M. (2009). The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE International Conference on Computer Vision Workshops*, pages 1089 – 1096.

The-World-Factbook (2013a). Germany population. https://www.cia.gov/library/publications/the-world-factbook/geos/gm.html.

The-World-Factbook (2013b). Italy population. https://www.cia.gov/library/publications/the-world-factbook/geos/it.html.

The-World-Factbook (2013c). Japan population. https://www.cia.gov/library/publications/the-world-factbook/geos/ja.html.

Thome, N., Serge, M., and Ambellouis, S. (2008). A real-time multiview fall detection system: A lhmm-based approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1522 – 1532.

United Nation (2002). World population ageing: 1950 - 2050. In *World Assembly on Ageing*.

United Nation (2007). World population aging 2007. Technical report, Department of Economic and Social Affairs, Population Devision, United Nation.

Vedaldi, A. and Fulkerson, B. (2008). Vlfeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511 – 518.

Wang, C. C., Chiang, C. Y., Lin, P. Y., Chou, Y. C., Kou, I., Huang, C. N., and Chan, C. T. (2008). Development of a fall detecting system for the elderly residents. In 1362, ., editor, *IEEE International Conference on Bioinformatics and Biomedical Engineering*.

Wang, H., Kläser, A., Schmid, C., and Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103:60–79.

Wang, H., Ullah, M. M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, pages 1 – 11.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T. S., and Gong, Y. (2010). Locality-constrained linear coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3360 – 3367.

Wang, X., Wang, L. M., and Qiao, Y. (2012). A comparative study of encoding, pooling and normalization methods for action recognition. In *Asian Conference on Computer Vision*, volume 3, pages 572 – 585.

Ward, G., Holliday, N., Fielden, S., and Williams, S. (2012). Fall detectors: a review of the literature. *Journal of Assistive Technology*, 6(3):202 – 215.

Willems, G., Tuytelaars, T., and Gool, L. V. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision*, volume 2, pages 650 – 663.

Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfinder: real-time tracking of human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780 – 785.

Yang, J., Yu, K., Gong, Y., and Huang, T. S. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1794 – 1801.

Yu, M., Rhuma, A., Naqvi, S. M., and Chambers, J. (2011). Fall detection for the elderly in a smart room by using an enhanced one class support vector machine. In *International Conference on Digital Signal Processing*, pages 1 – 6.

Yu, M., Rhuma, A., Naqvi, S. M., Wang, L., and Chambers, J. (2012). A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment. *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1274 – 1286.

Yu, X. G. (2008). Approaches and principles of fall detection for elderly and patient. In *IEEE International Conference on e-Health, Networking, Applications and Services*, pages 42 – 47.

Zambanini, S., Machajdik, J., and Kampel, M. (2010). Detecting falls at homes using a network of low-resolution cameras. In *IEEE International Conference on Information Technology and Applications in Biomedicine*.

Zhang, C., Tian, Y., and Capezuti, E. (2012a). Privacy preserving automatic fall detection for elderly using rgbd cameras. In *International Conference on Computers Helping People with Special Needs*, pages 625 – 633.

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213 – 238.

Zhang, Z., Liu, W., Metsis, V., and Athitsos, V. (2012b). A viewpoint-independent statistical method for fall detection. In *International Conference on Pattern Recognition*, pages 3626 – 3630.

Zhuang, X., Huang, J., Potamianos, G., and Hasegawa-Johnson, M. (2009). Acoustic fall detection using gaussian mixture models and gmm supervectors. pages 69–72.

Zivkovic, Z. (2004). mproved adaptive gausian mixture model for background subtraction. In *International Conference on Pattern Recognition*, pages 28 – 31.

Zweng, A., Zambanini, S., and Kampel, M. (2010). Introducing a statistical behavior model into camera-based fall detection. In *International Symposium on Visual Computing*, pages 163 – 172.

List of Publications

Journal Articles

1. Dao Huu Hung and Hideo Saito, "The estimation of heights and occupied areas of humans from two orthogonal views for fall detection," *IEEJ Transactions on Electronics, Information, and Systems*, Vol. 133, No. 1, pp. 117 - 127, January 2013.

2. Dao Huu Hung, Gee-Sern Hsu, Sheng-Luen Chung, and Hideo Saito, "Real-time counting people in crowded areas by using local empirical templates and density ratios for people counting," *IEICE Transactions on information and Systems*, Vol. E95-D, No. 7, pp. 1791 - 1803, July 2012.

Conference Proceedings Articles

1. Dao Huu Hung, Hideo Saito, and Gee-Sern Hsu, "Detecting fall incidents of the elderly based on human-ground contact areas," *Asian Conference on Pattern Recognition*, pp. 516 - 521, Okinawa, Japan, November 2013.

2. Dao Huu Hung and Hideo Saito, "Fall detection with two cameras based on occupied area," *Joint Japan-Korea Workshop on Frontier in Computer Vision*, pp. 33 - 39, Kawasaki, Japan, February 2012.

3. Dao Huu Hung, Sheng-Luen Chung, and Gee-Sern Hsu, "Local empirical templates and density ratios for people counting," *Asian Conference on Computer Vision*, Vol. 4, pp. 99 - 101, Queensland, New Zealand, November 2010.