AN INTELLIGENT CAMERA FOR THE HEALTHCARE

Master Thesis | A.J.J. Aertssen

Supervisors | M. Rudinac, Prof.dr.ir. P Jonker
"Research is to see what everybody else has seen, and to think what nobody else has thought."

Albert Szent-Györgyi (1893-1986)
Maybe it was just luck, or maybe there is no such thing as luck. Working on this master project gave me the insight on what I wanted to do in my future work. I remember the day I came in contact with my professor Pieter Jonker and explained him the assignment I wanted to do which was based on the detection of humans leaving the bed. He talked about a similar project using camera’s to detect human actions. I was not experienced in the field of computer vision but the feeling was good enough to go for it. I want first to thank Prof.dr.ir Pieter Jonker for giving me this assignment and for having enough confidence in me to make this a successful research. Secondly I would like to give my special thanks to Maja Rudinac, PhD student at the TU Delft and an expert on computer vision. She encouraged me and guided me through the whole process of my master project. I learned that working on computer vision topics in a group that is highly multi-disciplinary, is not so straight forward. With the new techniques in computer games and networking, computer vision is one of the fastest going fields and a lot of literature can be found on different topics. Maja guided me trough this literature and stimulated me to continuously improve my work. This work is not the result of one person but it is more a team effort. I’ve enjoyed the freedom of exploring and defining my research direction and I thank her for being patient and critical with my work. The discussions and collaboration resulted in two accepted conference papers on this topic.

REAL TIME FALL DETECTION AND POSE RECOGNITION IN HOME ENVIRONMENTS.

FALL AND ACTION DETECTION IN ELDERLY HOMES
AAATE : Association for the Advancement of Assistive Technology in Europe, 11th European AAATE Conference, Maastricht

I would also like to thank my colleagues from DAZA who supported me trough this project and giving me their knowledge about the healthcare. Also for giving me free time to finish my master thesis and helping me with designing a fisheye camera and making the scientific data. Also I would like to thank the volunteers from the Rubenshof for letting me monitor them during a day. It is because of them that I was able to finish this research successfully.

My family have been a great support both financially and mentally trough my years of college. They taught me to never give up and keep believing in myself which is the spirit you need for research. And finally I would like to thank my girlfriend Wendy, for her patient, positive spirit and love. She stood by me the whole process of my master study and she deserves more than just a ‘thank you’.

Jerry Aertssen, Augustus 2011
SUMMARY

Objective: Injuries caused by falls of elderly people are a common worldwide problem and ageing of population will even further increase related burdens and costs. Recent technology using active monitoring systems have proven their success in order to analyze human actions. What is lacking in these researches is implementation in real elderly home environments. Most of the healthcare researches are focusing on the detection of falls and not on the detection of normal daily actions. We present a single camera with a fisheye lens which is capable of monitoring an entire room. The use of only one camera reduces the costs and simplifies the computational burden which results in a real time system. While different research is done on the detection of such actions, none of these is done using real data by elderly people in their own living environment. Using this data will increase the difficulty level of the action recognition, because every living environment will have different settings and noise factors.

Main: We developed an action detection system which monitors the actions of elderly people in their homes during normal daily activities with the idea to raise the alarm in the case of danger. Our system is equipped with a single wide angle camera mounted on the ceiling of an elderly home. This gives a topview image of the environment resulting in a clear map of household objects without any occlusions. The main idea is to monitor the motion information of elderly and to model actions as a change of motion or poses in time that leads to a specific action. After background subtraction using Gaussian Mixture Models, the motion information is extracted using the Motion History Images method and analyzed to detect important actions. We propose to model actions as the shape deformations of the motion history image in time. Every action is defined at several moments in time, called “Action peaks” using different features, the holistic area, contour and location measurements as well as the Fourier shape descriptors. We combine all the measurements into the Bag of Word model and create unique action representations called ‘Action Signatures’. These action signatures are then transformed and combined using feature fusion in order to learn the optimal combination of features for each action. Learning the optimal feature fusion is performed using Support Vector Machines. The final trained system is used to classify each new action.

Results: the result section is divided into 2 sections. First the scientific data is used which is recorded in a testing room, simulating elderly home, with colleagues and students. We recorded and detected multiple actions: Bending, Walking, Falling, Collapsing, all with very high accuracy rates, above 93%. Finally real data is recorded in real elderly homes observing 4 elderly people. Different actions are monitored: Walking, Sitting, Open Door, and Eating. Results in a real environment depict high detection rates and prove that the system is able to detect multiple human actions using only one single camera.
CONTENTS

Acknowledgements .................................................................................................................. 3
Summary ..................................................................................................................................... 4
Contents ...................................................................................................................................... 5
1. Introduction .......................................................................................................................... 7
   1.1 Fall Incidents ................................................................................................................... 7
   1.2 Risk Factors .................................................................................................................... 8
   1.3 Monitoring systems ......................................................................................................... 9
   1.4 Research Focus .............................................................................................................. 9
   1.5 Overview ...................................................................................................................... 10
2. Related Work ....................................................................................................................... 11
   2.1 Human detection ............................................................................................................ 11
   2.2 Action Recognition ....................................................................................................... 14
      2.2.1 Environmental settings .......................................................................................... 14
      2.2.2 Intra- and interclass variability ............................................................................ 14
   2.3 Available datasets ......................................................................................................... 14
      2.3.1 Weizmann & KTH ............................................................................................... 15
      2.3.2 inria xmas ........................................................................................................... 15
   2.4 Detecting Motion ............................................................................................................ 15
   2.5 Wearable Sensors ......................................................................................................... 16
   2.6 Healthcare Applications ............................................................................................... 16
   2.7 Conclusion .................................................................................................................... 17
3. System ................................................................................................................................... 18
   3.1 Camera SetUp ............................................................................................................... 18
      3.1.1 Topview ............................................................................................................... 18
      3.1.2 Barrel Distortion ................................................................................................. 19
   3.2 Human and Action Detection ....................................................................................... 19
4. System Overview ................................................................................................................. 20
   4.1 Motion Detection .......................................................................................................... 21
      4.1.1 Frame Differencing .............................................................................................. 21
      4.1.2 Gaussian Background Model ............................................................................... 22
   4.2 Motion History Image ................................................................................................... 23
   4.3 Feature Measurements .................................................................................................. 24
      4.3.1 Blob Area ............................................................................................................. 24
      4.3.2 Position & Change of Position in time ................................................................. 24
      4.3.3 shape measurements ........................................................................................... 25
   4.4 Action Trigger peaks ................................................................................................. 25
4.5 Bag of words ........................................................................27
  4.5.1 Action Signatures .................................................................27
  4.5.2 Feature Fusion ..................................................................29
4.6 Methods overview ..................................................................30
  4.6.1 Method 1 ........................................................................30
  4.6.2 Method 2 ........................................................................32
5. Experimental setup and Results ..............................................34
  5.1 First Experiment: experimental room ....................................34
    5.1.1 Fall and Walk Detection ..................................................35
    5.1.2 Bend and Collapse Detection .........................................37
    5.1.3 results first experiments ................................................37
  5.2 Second experiment: Realistic Data ........................................38
    5.2.1 top view Camera .............................................................38
    5.2.2 Walking Action ...............................................................38
    5.2.3 Bending & Sitting Action ...............................................39
    5.2.4 Testing & Results ..........................................................39
  5.3 experiment 3: monitoring real elderly homes ......................40
    5.3.1 Training & Testing ........................................................41
    5.3.2 inter patient Training ....................................................42
    5.3.3 INTRA patient Training .................................................44
6. Conclusion ..............................................................................46
  6.1 Recommendations .............................................................47
References ..................................................................................48
Appendix ....................................................................................50
1. INTRODUCTION

The improvement of healthcare in the last decades has lead to lower mortality rates and longer lives, with many beyond the 100 years. This aging population however requires many adjustments both economically and socially, in particular in the work sector, pensions and healthcare institutes. The number of elderly people aged 65+ increases rapidly, and the amount of children (under the age of 15) decreases. Today the ratio of young people with respect to the elderly has decreased from six to three children for one older person. The expected growth of elderly people will result in a ratio where there are more elderly than younger people. Because of this ageing problem, serious questions arises about the financial viability of healthcare systems for the elderly. [5].

The increase in the population of elderly will result in an increase in healthcare expenses [13], but with the decreasing number of younger people, and therefore the decrease in nursing staff, many of the elderly will not receive the adequate healthcare they need. This aging problem results in new trends where elderly care is shifted from healthcare institutes to healthcare at home. Besides the financial benefits, elderly people also maintain their independence, allowing them to live in their own home environments. Because of this shift to elderly care at home there is a need for new technological systems which can provide the necessary information for an adequate healthcare. Such monitoring systems would improve the quality of life for the elderly and delay the moment of transition to costly facilities [14].

1.1 FALL INCIDENTS

A common worldwide problem, which will increase because of the aging population, are injuries caused by falls of elderly people. Around 30% of people aged 65 years or older living in the community and more than 50% of those living in residential care facilities or nursing homes fall every year, and about half of those who fall do so repeatedly [11]. Although not all falls lead to injury, about 20% need medical attention, 5% result in a fracture, and other serious injuries, such as severe head injuries, joint distortions and dislocations, and soft-tissue bruises, contusions, and lacerations, arise in 5–10% of falls. These percentages can be more than doubled for women aged 75 years or older (see Figure 1) [2]. Injury is the fifth leading cause of death in elderly adults, and most of these fatal injuries are related to falls. Falls account for over 80% of injury-related admissions to hospital of people older than 65 years. A fall and related injury, or even a fear of their consequences, such as social withdrawal, loss of independence and confidence, and admission to a long-term care facility, can cause severe depression and anxiety. In the Netherlands 92,000 elderly need medical care after a fall, 36,000 of them must stay in the hospital and nearly 1800 patients die each year. This leads to a economic cost of €5600; for each fall incident [17].
Almost half of the fall incidents of elderly people take place inside of their house. There are lots of different factors that increase the chance of fall incidents. Because of the decreasing muscle force and movement speed, it is harder to keep the body balanced. Besides that, the reaction time decreases which result that elderly cannot judge dangerous situations in time. Most of fall incidents with elderly are caused by tripping or bumping into objects. The risk factors can be divided into intrinsic and extrinsic factors.

**INTRINSIC RISK FACTORS:**
- physiologic
- neurologic
- pathologic

**EXTRINSIC RISK FACTORS:**
- living environment of the house
- living style
- public room
- adverse reactions to medicine

In order to prevent elderly to fall, some determinants can be influenced. One can influence the physiological state of the human, learn to anticipate on human behavior, perform adjustments of the surrounding environment, adjustment of living style and/or adjustments in medicine [17]. In this thesis, we limit our focus to the anticipation of human behavior.
1.3 MONITORING SYSTEMS

Many home and care giving facilities have implemented different kinds of alarming systems. The most commonly used alarms are the wearable pressure buttons and mounted pull cords which will alarm the caregivers when activated. Still these devices require manual activation by the elderly which is in some cases not possible. More advanced systems are designed to recognize alarm situations automatically. The main advantage of such systems is that they do not need users to activate the alarm, when an alarm situation occurs. These systems will protect the elderly and will be able to offer help when needed. When using active monitoring system one can also apply preventive care for the elderly. The elderly is monitored and warned when a dangerous or abnormal action is performed. When applying long term monitoring, one can use the normal activity as an indicator of the health of the elderly [12]. Recent studies prove that actions can be monitored using intelligent camera systems. These systems are evaluated in the laboratory settings in which an office is transformed to a living room. In this thesis our approach is evaluated using realistic data recorded in elderly homes consisting of a large variety of actions performed by the elderly during their normal daily activities.

1.4 RESEARCH FOCUS

Recent technologies using active monitoring systems have proven their success in analyzing human actions. What is lacking in these researches is the implementation in real elderly home environments. Most of the healthcare researches are focusing on the detection of falls and not on the detection of normal daily actions. In this thesis we give answers to the following questions:

*Can we design an intelligent monitoring system which can work in real time and will be financially beneficial to be applied in the elderly care system?* We present a single camera with a fisheye lens which is capable of monitoring an entire room. The use of only one camera reduces the costs and simplifies the computational burden which results in a real time system.

*How can we detect multiple actions performed by the elderly in a realistic environment?* We developed an action detection system which monitors the motion information of elderly and model the actions as a change of motion or poses in time that leads to a specific action. These actions are recorded in real elderly home environments with elderly people in their normal daily activities.

The focus of this research will mainly be on the detection of human actions. While most state of the art research is done on the detection of falls, this theses will take it to the next level and focus on the detection of different actions. Main problems that we had to overcome are occlusions in the camera field of view and the fact that different actions have partially the same motion. Normal actions during daily activities are: Walking, Bending, Standing Up & Sitting Down, while the dangerous action is falling...
down. While different research is performed on the detection of such actions, none of these is done using real data by elderly people in their own living environment. Using this data will increase the difficulty level of the action recognition, because every home environment will have different settings and noise factors. The next chapter will further describe state of the art on human detection and action recognition.

1.5 OVERVIEW

Chapter 2 will give an overview of the previous work done on human action recognition and the detections of humans in images. Recent literature is shown which show results in healthcare vision applications

Chapter 3 describes the system configuration and the use of the camera.

Chapter 4 will explain all methods proposed during this research for the detection of human actions.

Chapter 5 shows all the results split up into 3 different parts of the research.
2. RELATED WORK

This chapter describes the research done on the detection of humans and the recognition of actions. The first part focuses on the detection of humans using images, while the second part gives an inside to the healthcare applications.

2.1 HUMAN DETECTION

The detection of human subjects from recorded data is an important first step for the analysis of a human pose, human actions or the presence of multiple humans. Recent state of the art research has proven to be very successful in the detection of humans using vision based applications.

Papageorgiou et al [18] describes a method for object detection in unconstrained, cluttered scenes. The first step in their approach is representing the image from pixel space into the space of wavelet coefficients. Wavelets employ a dynamic set of basis functions that represents the input function in the most efficient way. Thus wavelets are able to provide a great deal of compression and are therefore very popular in the fields of image and signal processing. To differentiate between classes they propose a Support Vector Machine which constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class. mention that one of the main attraction of a SVM is that it is capable of learning in high-dimensional spaces with very few training examples. Their proposed method proved to give an accurate result for people detection with a success rate of 90% with 1 false positive for every 10,000 patterns processed. An important factor which inhibits the system for being used in real time applications is the slow processing speed. In order to increase the processing speed they propose different methods. They conclude that this detector should work more appropriate in combination with another system, and making use of dynamical information when processing video sequence.

Depoortere et al [19] proposed some methods to increase the processing speed of human recognition using the same principles as Papageorgiou and Poggio [18]. The first step in order to reduce the amount of computational work was to reduce the dimensions via feature selection. Papageorgiou et al [18] applied feature selection based on the mean of a wavelet coefficient, relative to the overall mean of all coefficients of the same class. Depoortere proposed 3 other methods of feature selection. The first method uses Bhattacharyya measures, the second feature selection procedure that was tested was AdaBoost based. The third feature selection procedure makes use of the decision surface of a SVM that was trained with the full set of features. The decision surface is then analyzed to identify those features that influence the decision the most. The second step in order to reduce the amount of computational work was dimension reduction by projecting the feature vector in a low dimensional subspace. The third method that was proposed was to reduce the number of support vectors. They accomplished that by using Support Vector Regression Machines (SVRMs). This method was even better than
the proposed re-implemented approach of Papageorgiou et al [18] and was about 3 orders of magnitude faster.

Gavrila et al [10] proposed a method for shape-based object detection using Distance Transforms (DT). The graylevel image which results from the distance transform are compared with predefined templates. The templates which are similar are grouped into a prototype template which result in faster detection rates. This grouping is done at various levels using a bottom-up approach and a K-means algorithm. This results in a hierarchy of prototypes. The features used in this method are oriented edges, which were divided in 8 values. For detecting they compare the templates with new data using the chamfer distance. The matching remains dependent on a reasonable contour segmentation and it might not work when pedestrians are very close to the camera when shape variations become even larger.

Viola et al [25] propose a method which combines static en dynamic information into a pedestrian recognizer. They measure the difference between region averages at various scales, orientations and aspect ratios. Absolute, motion shear, magnitude and appearance filters are used. Scale invariance is achieved during training process simply by scaling the training images to a base resolution. The scale invariance of the detection is achieved by operating on image pyramids. AdaBoost is used to select a subset of features and construct a classifier. The resulting classifier balances intensity and motion information in order to maximize detection rates. To reduce computation time cascade architecture is used, where simple detectors are placed earlier in the cascade. Information about the direction of motion can be extracted from the difference between shifted versions of the second image in time with the first image. They conclude that using the simple filters in combination of static and dynamic information results in a extremely low computation time for detecting pedestrians and gives a good detection rate.

Navneet Dalal and Bill Triggs [6] proposed a method to detect humans by evaluating well normalized local histograms of image gradient orientations in a dense grid, using simple architecture with a single detection window. An image is divided in small regions called cells. On each cell a histogram of gradient directions or edge orientations over the pixels of the cell is calculated. This results in a distribution of each cell which represents local object appearance and shapes. They propose to normalize each cell over somewhat larger spatial regions. This method is referred as Histogram of Oriented Gradients (HOG). For classification they use a soft linear support vector machine. In the results they show that fine scale gradients, fine orientation binning, relative coarse spatial binning and high quality local contrast normalization in overlapping descriptor blocks are all important for good performance. They recommend that it is useful to develop a coarse-to-fine style detector based on HOG descriptors. Also using motion information should improve the results.
<table>
<thead>
<tr>
<th>Paper</th>
<th>Features</th>
<th>Approach</th>
<th>Classification</th>
<th>Advantage / Disadvantage</th>
<th>Applicable for this research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papageorgiou et al [18]</td>
<td>Haar Wavelets (intensity differences), Feature selection based on the mean of a wavelet coefficient, relative to the overall mean of all coefficients of the same class</td>
<td>Holistic</td>
<td>SVM</td>
<td>An important factor which inhibits the system for being used in real time applications is the slow processing speed.</td>
<td>Not applicable because of slow processing speed.</td>
</tr>
<tr>
<td>Depoortere et al [19]</td>
<td>Haar Wavelets (Based on Papageorgiou et al [19]) Feature selection: • Bhattacharyya measures • AdaBoost • decision surface of a SVM dimension reduction by projecting the feature vector in a low dimensional subspace reduce the number of support vectors</td>
<td>Holistic</td>
<td>Cascade of SVM's</td>
<td>Their proposed method was better and faster than Papageorgiou et al [8]. But still sometimes pedestrians were not found or shaped liked trees were classified as humans.</td>
<td>Not applicable because of the performance is still too low for healthcare purpose.</td>
</tr>
<tr>
<td>Gavrila et al [10]</td>
<td>Distance Transforms: oriented edges</td>
<td>Holistic</td>
<td>Template matching using chamfer distance</td>
<td>Method allows real time detection, but really depends on a reasonable contour segmentation. Pedestrians close to the camera are not detected because shape variations become larger.</td>
<td>Not applicable because of low human detection rate, because it needs perfect contour segmentation.</td>
</tr>
<tr>
<td>Viola et al [25]</td>
<td>Motion patterns from pairs of sequences of images by: • absolute difference • comparing sums within the same motion images • magnitude of motion in one of the motion image Appearance (rectangle) Filter Integral images</td>
<td>Holistic</td>
<td>Cascade of Adaboost</td>
<td>Using motion information improves the detection rate of humans and reduces false positives.</td>
<td>In some extend motion detection might be useful to reduce falls positives. But not the way to detect humans which are not moving.</td>
</tr>
<tr>
<td>Navneet Dalal and Bill Triggs [6]</td>
<td>Histograms of Oriented Gradients: On each cell a histogram of gradient directions or edge orientations over the pixels of the cell is calculated</td>
<td>Holistic</td>
<td>SVM</td>
<td>Using HOG they outperformed other approaches using it for human detection. Approach could be optimized using motion detection , part based detection and course to fine approach</td>
<td>Applicable for this research because of accurate human detection, but some improvement could be on computational speed.</td>
</tr>
</tbody>
</table>

**TABLE 1: HUMAN RECOGNITION OVERVIEW**
2.2 ACTION RECOGNITION

Computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, or multidimensional data from a medical scanner. As a technological discipline, computer vision seeks to apply its theories and models to the construction of computer vision systems. There already exists a number of working systems that perform parts of this task in specialized domains. However, the generic "Vision Problem" is far from being solved. No existing system can come close to emulating the capabilities of a human. Vision is therefore one of the problems of computer science most worthy of investigation because we know that it can be solved, yet we do not know how to solve it well. The main challenge in human action recognition is to point out a label which represents a certain action. The common approach is to extract features from the video images and analyze them. By using predefined actions learned from training data the action can be classified. The choice of training data and features to analyze is the biggest challenge and depends totally on the actions you want to detect. Next we discuss the difficulties which need to be solved during action recognition.

2.2.1 ENVIRONMENTAL SETTINGS

Different scenes can lead to different variations for the action recognition process. Outdoor scenes give variations in lightning conditions, moving objects because of the wind. While indoor scenes are less vulnerable for lightning and wind conditions, there can be a large burden in cluttered scenes. Indoor objects like furniture can block parts of the moving subject which can result in poor action recognition results. Different viewpoints of the camera can result in different classifications for the same action. The system is getting more complex when there are multiple background motions or when the camera position changes in time.

2.2.2 INTRA- AND INTERCLASS VARIABILITY

Actions performed by humans have large variations even when they perform the same action. Simple walking actions can differ in speed, posture of the upper body, movement of the limbs etc. There can be an overlap of certain actions which will make the classification between actions harder. A good human action classification system must be able to separate different actions and generalize actions from the same class.

2.3 AVAILABLE DATASETS

In the field of vision based research there are some publicly available datasets containing human actions. These datasets can be used for the comparison of the obtained results.
2.3.1 WEIZMANN & KTH

These recordings contain different actions, 6 for the KTH and 10 for the Wiezmann, which are recorded with a static background. The actions are performed by different actors with variations in the performance and the duration. The actions are recorded using only one camera in different scenarios, mostly outdoors.

2.3.2 INRIA XMAS

A total of 14 actions are performed by 11 actors while recorded at 5 different viewpoints. The cameras are fixed, while the actions are in arbitrary direction. The background in this dataset is also static.

2.4 DETECTING MOTION
When using static background data, one can obtain the region of interest, ROI, by applying background subtraction. The resulting information contains a silhouette of the moving object. Although silhouettes contain lots of noise and are influenced by orientation and translation, they still contain lots of information which can be used for the human action detection. Bobick and Davis proposed a method to subtract background by looking at the differences between two sequent frames looked from a single viewpoint, also referred as frame differencing. Also the difference between frames of a total action sequence is extracted which results in a Motion Energy Image (MEI). Finally they construct a Motion History Image (MHI) where pixel intensity is used in the silhouette motion [3].

2.5 WEARABLE SENSORS

Most commonly used applications in healthcare make use of wearable sensors. Often these sensors are worn by the elderly on their body, like heart rate sensors. Or they are mounted inside clothes to track the position of the user and recognize certain activities [16]. Most wearable sensors use accelerometers to measure the magnitude and direction of acceleration. One of the approaches is based on the wearable systems which are able to detect falls. Zhang et al [26] uses a non-negative matrix factorization method for feature extraction. The major advantage of this method is the accuracy of detecting a fall. Using wearable systems mainly provides human position and movement information. Such systems give limited information which makes it difficult to recognize specific actions. Wearable systems are easy to implement but there is always a tradeoff with human comfort. The biggest problem with wearable sensors is the fact the people might forget to wear the sensor.

2.6 HEALTHCARE APPLICATIONS

Recent research done on computer vision in healthcare applications has proven to be successful in detecting dangerous situations. While different actions are analyzed, the main focus of these researches is on the detection of a fall incident. J. Toa presents an intelligent video surveillance system to detect human fall incidents for enhanced safety in indoor environments. The system consists of two main parts: a vision component which can reliably detect and track moving people in the view of a camera, and an event-inference module which parses observation sequences of people features for possible falling behavioral signs. In particular, they extract the aspect ratio of a person as observation feature, based on which fall incidents are detected as abrupt changes in the feature space. Their experiments show that the proposed approach can robustly detect human falls in real time [23]. A similar research done by Anderson et al also applies background subtraction which results in a single silhouette. To reduce the noise caused by shadow they measure color information. Finally they apply a bounding box around the resulting silhouette and analyze the length/width ratio of it. They train hidden Markov Models for the action classifications [1]. Both methods are very successful in
detecting fall incidents, but are limited in detecting other actions. An even bigger problem is the occlusion of furniture with both researches. The positioning of the camera is crucial and the occlusion of objects with the moving subject will make it impossible to detect certain actions. Diraco et al use 3D depth information using a Time of Flight camera which is mounted on the wall. An off-line calibration is applied to analyze the room and camera parameters. The distance of the 3D human centroid from the floor plane is evaluated by using the previously defined calibration parameters and the corresponding trend is used as feature in a thresholding-based clustering for fall detection. The fall detection rate is very high but the system will be limited by its financial costs. Due to the fact they make use of a Time of Flight camera which are relatively expensive compared to normal camera’s [7].

To overcome the occlusion problems Nait-Charif proposed a method based on the previous mentioned methods from J.Tao and D. Anderson. They make use of a single camera with a fisheye lens which is mounted on the ceiling. The created topview images give a clear overview of the room and reduces the occlusion problem. The bounding box principle is used in order to detect fall incidents [15].

2.7 CONCLUSION

An overview of related work on action and human detection is presented in this chapter. Recent analyses have proven the success of human action detection in the healthcare systems. There are various methods applied on the available datasets and self acquired date, all with their own advantages and disadvantages. While most methods are not specifically designed for healthcare applications, still these algorithms can provide enough information to be applied in healthcare. The biggest gap between these researches and the healthcare is the integration of these methods in real life. Most researches using vision still need fast computers in order to analyze the data and apply the algorithms in real time. Another important issue is the expected price of the final system. The growth in elderly population will result in less money for the healthcare. There must be a tradeoff between the “perfect” product and the end price which nursing facilities are willing to pay. Most researches do not focus on these problems which can result in a high cost complex system that needs multiple camera’s, which will be unaffordable for the institutes. This research focuses on the detection of human actions, while keeping in mind that the end product must work in real time using only standard home computers and that the price of the product must be kept as low as possible. During this research, we also acquired our own dataset since the existing datasets were not applicable for our methods. The next chapter will describe the proposed system; the camera setup and choice of camera will be explained, and the methods applied to perform the action recognition explained.
3. SYSTEM

Looking at previous studies one can ask why there is currently no existing action recognition system for the healthcare. The challenges in detecting human actions with all the varieties in motion and human sizes have proven to be really hard to overcome. In this chapter we discuss the overview of the challenges we had to overcome in the design of our action recognition system. The choice of camera and position of the camera is discussed and finally human action detection methods will be described.

3.1 CAMERA SETUP

Inside elderly homes there are lots of objects in the room that can block the view of the camera to the elderly. This is one of the biggest problem to overcome when applying camera vision in a living environment. Occlusions can cause a dramatic decrease in the fall detection rate [9].

FIGURE 5: OCCLUSION PROBLEM [9]

The choice of camera position is crucial for the accuracy of human action recognition. To reduce this problem, one can implement multiple camera’s at different positions. This will give a clear overview of the room and when occlusions occur system can switch between different cameras to get another viewpoint. The main drawback of such a system is the high market price of buying more cameras as well as the computational price when combined information from all cameras is analyzed.

3.1.1 TOPVIEW

In this research we propose the use of the Geovision Gv-Fe111 camera mounted on the ceiling. With a fisheye camera a 180 degree view of the room looked from above is obtained. In this way occlusions are brought back to a minimum while a clear overview of an entire living room will be given. Also use of just a single camera highly reduces the total cost of the system.
3.1.2 BARREL DISTORTION

The problem with using a fisheye lens is that they suffer from some amount of distortion, better known as "Barrel distortion". In "barrel distortion", image magnification decreases with distance from the optical axis. This results in an image which has been mapped around a sphere (barrel). Radial distortion, whilst primarily dominated by low order radial components, can be corrected using Brown’s distortion model [4].

\[
x_u = X_d + (X_d - X_c)(K_1 r^2 + K_2 r^4 + \ldots) + (P_1 (r^2 + 2(X_d - X_c)^2) + 2P_2 (X_d - X_c) (Y_d - Y_c)) \left(1 + P_3 r^2 + \ldots\right)
\]
\[
y_u = Y_d + (Y_d - Y_c)(K_1 r^2 + K_2 r^4 + \ldots) + (P_2 (r^2 + 2(Y_d - Y_c)^2) + 2P_1 (X_d - X_c) (Y_d - Y_c)) \left(1 + P_3 r^2 + \ldots\right)
\]

Where:

\((X_u, Y_u) = \text{undistorted image point}, \quad (X_d, Y_d) = \text{distorted image point},\)
\((X_c, Y_c) = \text{centre of distortion}, \quad K_n = \text{nth radial distortion coefficient},\)
\(P_n = \text{nth tangential distortion coefficient}, \quad r = \sqrt{(X_d - X_c)^2 + (Y_d - Y_c)^2}, \text{and } \ldots = \text{an infinite series}\)

Compensating the distortion will result in better image data but will also increase the computational power. For this research no compensation is applied because the biggest distortion takes place at the outer side of the video. Because we are monitoring within elderly homes, these distortion is mostly at the walls where no human action will take place. To increase the computational speed we choose not to apply Brown’s distortion model.

3.2 HUMAN AND ACTION DETECTION

As seen in the previous work section, lots of research is done on the detection of humans in images. Most of them need additional training and require high computational power. While most methods acquire high detection rates, most of them are used in an external settings. We have the advantage that we use internal settings with a static camera. While all other objects in the room have no motion, one can detect humans using motion information. While all other objects in the room stay static, the human will have motion information which can be detected by applying background subtraction logarithms. The resulting information is the motion silhouette of each frame which can be analyzed for the detection of actions. When the system detects a motion it will classify it as a specific action. If this action is classified as dangerous it can automatically send an alarm to the nursing system. When the system fails to detect a dangerous action and the human is still in the room but there is no motion for a specific amount of time, an alarm will be generated by the inactivity measurement. In this way a complete alarm system will be provided. This holistic approach of human and action detection gives the advantage of its robustness and real time implementation. The system only uses one camera which results in a cheap solution for the improvement of healthcare. In the next chapter the specific algorithms for the background subtraction and action recognition methods will be described.
As mentioned before, action recognition using camera’s is a challenging task. Different viewpoints, people, lightning conditions make it hard to design an universal system with accurate action detection. In this research we have the advantage that our camera is in a static place with external conditions that does not change much in time. This chapter will describe the methods used and designed for the detection of humans and the recognition of actions.
4.1 MOTION DETECTION

We define a human as a moving object in the living room in time, while an action is defined as a change in motion during time. To localize a moving object we proposed to apply background subtraction. There are many different methods to segment the background [1]. The main goal of background subtraction is removing information from the image data which did not change during time. A common used method in the field of background subtraction is the Frame differencing method.

4.1.1 FRAME DIFFERENCING

Frame differencing is a non-recursive technique which uses a sliding-window approach for background estimation. It stores a buffer of the previous video frames, and estimates the background image based on the temporal variation of each pixel within the buffer. Non-recursive techniques are highly adaptive as they do not depend on the history beyond those frames stored in the buffer [20]. Frame differencing uses the image information of frame at time ‘t’ and subtract that image from the image in frame ‘t-1’. The resulting image is a silhouette containing only information of the changed pixels. The major advantage of Frame Differencing is its simplicity and fast computation so it can be applied in real time applications. It does not require any prior processing and it is independent of the environmental conditions, such as the specific room type or illumination conditions. However, motion segmentation using this method is very coarse and dependant on the shadows in the scene.

Since it uses only a single previous frame, frame differencing may not be able to identify the interior pixels of a large, uniformly-colored moving object. This is commonly known as the aperture problem. To eliminate these effects and acquire a more stable result we applied Doubled Frame Differencing [3]. After capturing three successive frames in a video, two separate difference images (‘t-1’ and ‘t’) and (‘t ’ and ‘t+1’) are generated. These difference images are now binarized and summed up using the “and” operation. The resulting image is now the binary Double Difference Image (DDI). The resulting image is called “Silhouette of moving object”.

![Frame Difference Example](image)
4.1.2 GAUSSIAN BACKGROUND MODEL

A different approach of applying background subtraction is using a "Mixture of Gaussians (MoG)" which is a recursive technique. The main difference with Frame Differencing is that this method makes a background and foreground model which is continuously update in time. As a result, input frames from distant past could have an effect on the current background model. The pixel-level Mixture of Gaussians (MOG) background model has become very popular because of its efficiency in modeling multi-modal distribution of backgrounds (such as waving trees, ocean waves, light reflection, etc.), its ability to adapt to a change of the background (such as gradual light change, etc.) and the potential to implement the method in real time. Friedman and Russell [8] modeled the intensity values of a pixel by using a mixture of three Normal distributions and applied the proposed method to traffic surveillance applications. Stauffer and Grimson [21] presented a method which assumes that the time series of observations, at a given image pixel, is independent of the observations at other image pixels. They also assume that these observations can me modeled by a mixture of K Gaussians. Let x be a pixel value observed at time t [21].

\[ p(X_t) = \sum_{i=1}^{K} \frac{W_{i,t}}{n \sum_i \sum_t} \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2} (x_t - \mu_{i,t})^T \Sigma_{i,t}^{-1} (x_t - \mu_{i,t})} \]

Where:
- \( W_{i,t} \) = weight,
- \( \mu_{i,t} \) = Mean value,
- \( \Sigma_{i,t} \) = the covariance matrix for the ith Gaussian distribution at time t

The distributions are then sorted and only the first are used as background. MoG has proven that it is very robust as a background subtraction method, while it is more computational expensive than Frame Differencing.

The resulting information now only contains motion information and all static background information is removed from every frame. Knowing the fact that our system is in a static indoor environment we use this motion information for the detection of humans inside the living room. When a motion occurs which is larger than threshold X then this motion is classified as human motion. After the detection of humans we can than focus on the recognition of actions.
Now we need to capture the sequence of motion change (DDI images) belonging to the one action in a single image. For that we apply the method of Motion History Images (MHI) [3]. Basic idea is to model the motion by accumulating intensity changes of pixels. Now we can define the intensity as a function of the temporal history of motion at that point. The MHI at time $t$ is calculated according to Equation 1, where $D(x,y,t)$ represents DDI image at time $t$ and pixel position $(x,y)$. The variable $\tau$ represents the duration of movement, in consecutive frames, and $\text{MHI}_\tau(x,y,t)$ temporal history of motion at point $(x,y,t)$ occurring during the $\tau$ frames.

\[
\text{MHI}_\tau(x,y,t) = \begin{cases} 
\tau \left( \max (0, \text{MHI}_\tau(x,y,t - 1)) - 1 \right) & \text{If } \ldots D(x,y,t) = 1 \\
\text{Otherwise} & 
\end{cases}
\]

Resulting MHI is now a scalar valued image where more recently moved pixels appear brighter as can be seen on Figure 9: MHI Examples (Left: Bending, Right: Walking). Such MHI is useful for our application since we only need to know the shape and location of the motion change, not the direction.
The resulting silhouette image and motion history image can be used to apply feature measurements. These features or combination of features can be used to describe a certain action. Most feature measurements done on silhouette and on MHI images are features which contain useful information about the size of the silhouette, Area, Orientation, Shape etc. The features we measured for this research are described below.

### 4.3 FEATURE MEASUREMENTS

#### 4.3.1 BLOB AREA

The first measurement we perform is the Area of the Silhouette and MHI. This Area is defined as the number of positive pixels inside of the image.

During the performance of an action the Area of both the silhouette and the MHI will increase in time. The amount of Area increase depicts amount of movement that is used during certain action and can be useful for classification of different actions (e.g. compare Walking with Eating).

#### 4.3.2 POSITION & CHANGE OF POSITION IN TIME

During an action one is constantly moving its body. This movement can be measured by finding the center of mass of the silhouette and measure the change of this position in time. The center of mass is calculated with:

\[
X_{center} = \frac{1}{A} \sum X.\text{Coordinates of Blob}
\]

\[
Y_{center} = \frac{1}{A} \sum Y.\text{Coordinates of Blob}
\]

Where A is the Area of the blob. We measure the center of mass of the silhouette during specific intervals in time, as can be seen in Figure 10.

![Figure 10: Center of Mass](image)

The position of the center of mass is used to describe where an actual action is performed and used later to train the system. If the position of the center of mass is taken at multiple time periods, we would learn exact place where a certain action is performed (e.g. such as dining or sleeping).
4.3.3 SHAPE MEASUREMENTS

Finally we measure the change of the shape of the MHI. At first a contour of a MHI is generated, and afterwards described using Fourier Descriptors. The major advantage of using Fourier Descriptors is because of their invariance on translations, rotations and scale. The contour of a silhouette is described in the frequency domain in such a way that the lower frequencies describe the general contour of the silhouette while the higher frequencies describe the fine detail of the contour. In our application fine details of the contour are not useful for global contour discrimination. Therefore only a subset of the Fourier Descriptors is sufficient to describe the global contour of the silhouette. This reduces the dimension of the descriptors and increases the speed, which is a big advantage for applying it real time. For a given contour \( s(t) \) which is normalized to \( N \) points, the discrete Fourier transform is given by Equation:

\[
F_d = \frac{1}{N} \sum_{t=0}^{N-1} s(t) e^{-j \frac{2 \pi m t}{N}}, n = 0,1, ..., N - 1
\]

This results in a vector of complex numbers where the magnitude of the descriptors is divided by the magnitude of the second descriptor in order to apply scale normalization. This results in:

\[
F_d = \begin{bmatrix}
\frac{d_2}{d_1}, \frac{d_3}{d_1}, ..., \frac{d_{N-1}}{d_1}
\end{bmatrix}
\]

Scale invariance is now obtained by dividing the magnitude values of the Fourier descriptors by the first component. After that the first descriptor is discarded since it only gives information about the position of the contour and it is not describing the contour itself. [27]. Examples of the shape contours can be seen in Figure 11.

![Bending Shape Contour](image)

**FIGURE 11: BENDING SHAPE CONTOUR**

4.4 ACTION TRIGGER PEAKS

The features we measure on the blobs are giving important information of a specific action. The moment to analyze and measure these features in time is very important, because all of these measurements differ in time. When a person is performing one of the actions we want to detect, we can
observe a large increase of both the Area of silhouette and the Area of motion history image. Once the action is finished this area will decrease, which results in area peaks of the silhouette and motion history image. These peaks define when an action has happened and when to analyze it and are referred to as Action Trigger Peaks and are shown in Figure 12.

![Figure 12: Action Trigger Peaks](image)

The Action Trigger Peak is found by subtracting the Area at time $t$ and time $t-1$ checking the resulting value. The positive or negative resulting value corresponds to an ascending or descending slope. To overcome the problem that small peaks are found we apply this method on 10 sequential frames for the ascending and descending slope. When for 10 frames the slope is increasing and followed by 10 frames of descending a slope is detected.

$$ \text{Slope}(t) = \left( \text{MHI}_{\text{Area(t)}} - \text{MHI}_{\text{Area(t-1)}} \right) $$

- $\text{Slope}(t - 5) ... \text{Slope}(t) > 0 \implies \text{Increase}$
- &
- $\text{Slope}(t) ... \text{Slope}(t + 5) < 0 \implies \text{Decrease}$

$$ t = \text{Trigger Peak} $$

After the detection of a trigger peak we measure all features which are described above at the time of a trigger peak.
Because of the large variety of poses of the human body during one action we need to train the system in such a way that it detects actions with a high accuracy. We calculated multiple features but do not know in advance which features are good and relevant for a specific action. To solve this problem we applied the “Bag of Words” approach which is widely used in the field of action recognition. Using this approach, we calculate a dictionary, set of features that are describing a specific action. Given a video “V” and the feature vector “f” and number of frames N, we define bag of the words in following paragraph:

\[ V = \{ f_i \} (1 \leq i \leq N) \]

We calculate the features described above at the moment of the Action Trigger Peak, during a number of frames. This will result in a sequence of feature vectors:

\[ D = \{ d_i \} (1 \leq i \leq N) \]

Were \( d_i \) is a specific feature which forms the input of the Bag of Words.

When we train the system with multiple training action movies and store the feature vectors, we get a database of predefined “Words”. The next step in this system will be the classification of new data to the predefined words. We do this by analyzing the new feature vectors and calculate its distance to the predefined words. The three words with the minimal distance to the new feature vector are stored [22].

\[
Distance = |D_{\text{predefined}} - D_{\text{new}}| = \sqrt{\sum_{i=1}^{N} (D_{\text{predefined}} - D_{\text{new}})^2}
\]

What we end up with is a Histogram which reflects the distribution of all words which are describing the action. We call these histograms the “Action Signatures”.

### 4.5.1 ACTION SIGNATURES

A few examples of action signatures for the walking actions are shown in the Figure 13 and Figure 14. From here we immediately see that the results from the Bag of Words contain multiple “Words” or in our case Actions. This means that some features are more important for a certain action than other features. To make a decision which action is performed by analyzing all the different features in the Bag of Words method, we apply feature fusion method.
FIGURE 13: ACTION SIGNATURE FROM THE FEATURE "CENTER OF MASS" OF A WALKING MOVIE

FIGURE 14: ACTION SIGNATURE FROM THE FEATURE "AREA MHI"OF A WALKING MOVIE
4.5.2 FEATURE FUSION

Feature fusion is performed using the results of all the action signatures from the training data. Each action signature of each specific action is combined into one final feature fusion signature. The final histogram will then represent the combination of features describing specific action. In this case each action is represented with 5 features as it is shown in Figure 15.

![Feature Fusion Examples](image)

**FIGURE 15: FEATURE FUSION SIGNATURE EXAMPLES**

However, now the system needs to learn what is the optimal combination of features for every action. Therefore we train separate classifier for each class of actions. Resulting signatures of feature fusion are used as the input to the Support Vector Classifier. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class. It often happens that the sets to discriminate are not linearly separable in that space, therefore it was proposed that the original finite-dimensional space be mapped into a higher-dimensional space. To keep the computational load reasonable, the mapping used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function $K(x_i, x_j)$ selected to suit the problem [24]. For this research we choose the kernel function as a Gaussian Radial Basis Function:

$$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

as proposed in [6] for optimal speed and a high performance. Once the separate SVMs are trained for each action, joint decision is made using combined classifier. Trained learning network is learning optimal combinations of features for every action, and is used later in testing. Adding novel type of
features can be done without any further adjustment, but adding new actions, requires adding and training additional SVM.

4.6 METHODS OVERVIEW

In this section we have presented different parts of our action recognition system. The main goal was to design a stable and reliable system which can work in a real time with a high performance. Recent literature have proven that the individual methods described above work in a real time. The first part of our system removes background information in the movie. This is done with Gaussian background modeling due to the fact that this is best background subtraction technique which can still work in real time. Secondly all the features are calculated which are chosen carefully by extensive testing and research with our data. To combine these measurements we used a Bag of Words approach in combination of action signatures and feature fusion to come up with the input for our SVM classifier. The choice of the classifier is based on the literature study and the fact that it must work real time. The next chapter will describe our experimental setup and results of this system. We tested the system using scientific data produced by students and real data which is captured in real elderly homes.

4.6.1 METHOD 1

Figure 16 presents the total block scheme of the first method we applied on our scientific data. We first applied the background subtraction using Double Frame Differencing to obtain the Silhouette of moving object. Using this image, we now calculate the Area of the silhouette image and the Area of motion history image. We observed that the average maximum Area measurements of “Falling” and “Walking” exceed the “Bending” and “Collapsing” values by 20%. This is used to define the Area MHI threshold. Further analyzing the Action Trigger Peaks of the Area of Silhouettes, we observed that the average maximum values of “Falling” actions exceed the average maximum “Walking” value with 15%. This is used to define the Area Silhouette threshold. Combining these two thresholds we can distinguish between “Walking” and “Falling”. For all the actions in the training data, we extracted the action models in an offline step and formed the database. In an online step, when we want to detect a certain action, we first search for an Action Trigger Peak to detect the action culmination. Once it occurs and the Area value is below the Area MHI threshold (which eliminates “Walking” or “Falling” actions), model of that action is extracted using the Fourier Descriptors. Now to recognize which action it is, we compare the extracted model with all the models from the database. For comparison the Euclidian distance is used. Using the combination of the area threshold and the Fourier descriptors will classify the data with its belonging action.

The main drawback of this method was the manually defined threshold for action recognition. Therefore we improved the method and designed a learning network that will automatically calculate the optimal thresholds and feature combinations. This method is described in further text.
FIGURE 16: BLOCK SCHEME METHOD 1
4.6.2 METHOD 2

A complete overview of the final methods can be found in Figure 17. First, the background information of the input images from the camera are removed by applying the Gaussian mixture model. This will then result in a silhouette of the moving object:
- This silhouette is improved by removing the noise and shadows by applying morphological filtering operations.
- The final silhouette is then implemented and stored into the Motion History image, where the latest motion appears brighter than the previous motion.

After preprocessing the input frames from the camera, we go on analyzing the motion. In total, we calculate 5 features of the silhouette and motion history image:
- Area silhouette
- Area MHI
- Change of the position of the center of mass
- Position of silhouette
- Fourier descriptor

Using the information of the Area MHI, we can check whether there is an action occurring based on the action trigger peaks. When this happens, all calculated features are stored and analyzed:
- Create Action signature using the Euclidean distance to the training data

The output of all 5 action signatures are combined by feature fusion and will be the input to the support Vector Machines, which are trained from the training data for every single action. In total, we have 4 actions, so we also have 4 SVM’s:
- Feature Fusion
- Input to SVM’s
  - Classify with Dining SVM
  - Classify with Walking SVM
  - Classify with Sitting SVM
  - Classify with Door SVM

The outputs of all SVM’s are combined into the final decision classifier which classifies the motion to its belonging action.

The explanation of the experimental setup and results of both method 1 and 2 will be explained in the next chapter. We performed multiple experiments in different household conditions using scientific and realistic data. Here, we show the detection rate and the improvement by using the self learning algorithms.
5. EXPERIMENTAL SETUP AND RESULTS

To test our system, we used both synthetic and real data. Synthetic data was generated in experimental room in which students/actors performed dangerous actions (falling and bending) that would be impossible to obtain in a real environment. For the rest of experiments we placed camera in real elderly homes and monitored elderly people during their regular daily activities. This allowed us to test system in very challenging conditions, where illumination was changing, obstacles were present, layout of different homes was changing and different elderly people moved in very specific manner. Results and detailed description of all experiments is presented in the further text.

5.1 FIRST EXPERIMENT: EXPERIMENTAL ROOM

The first part of this research is performed in a experimental room where we used a normal webcam to monitor students in the room. We recorded in total 52 movies containing all the dangerous actions that elderly person can perform:

- Walking
- Bending
- Collapsing
- Falling

All data is recorded in a standard room during daytime under different illumination conditions. For the multiple action data another 16 movies were recorded using the same conditions. To make our testing more robust, we used in total 4 different subjects who all differ in height and wear different clothes. Example images of the data can be found in picture 15.

During this part of the research we did not apply the complete system as described in section 4. For this experiment, we extracted the motion of elderly using the Double Frame Differencing method described previously to remove the background and only remain with the silhouette of the moving object. To further improve these results we applied morphological filtering.

Secondly we applied the Motion History Imaging technique and calculated following 3 features.

- Area of the single Silhouette
- Area of the Motion History
- Fourier Descriptors

When action trigger peak occurs, each action was described using 3 features and matched with the database of different actions.
5.1.1 FALL AND WALK DETECTION

We generated our data in a home environment where all actions are performed under different illumination conditions and by multiple people wearing different clothes. Analyzing the Action Trigger Peaks of the MHI Area of the different actions, we observed that the average maximum “Falling” and “Walking” values exceed the “Bending” and “Collapsing” values by 20% (Figure 19). This is used to define the Area MHI threshold. Further analyzing the Action Trigger Peaks of the Area of Silhouettes, we observed that the average maximum values of “Falling” actions exceed the average maximum “Walking” value with 15% (Figure 20). This is used to define the Area Silhouette threshold. Combining these two thresholds we can distinguish between “Walking” and “Falling”. What is important to
notice is that both of these thresholds are learned from manually labeled training data by comparing differences in area values. These thresholds are then applied on the testing data, to detect the “Fall” and “Walking” actions.

**FIGURE 19:** CHANGE OF THE “AREA OF MHI” FOR DIFFERENT ACTIONS IN DIFFERENT CONDITIONS WHERE: “RED – FALLING”; “PINK – WALKING”; “GREEN – COLLAPSING” AND “BLUE – BENDING”

**FIGURE 20:** CHANGE OF THE “AREA OF SILHOUETTE” FOR DIFFERENT ACTIONS IN DIFFERENT CONDITIONS WHERE: “RED – FALLING”; “PINK – WALKING”; “GREEN – COLLAPSING” AND “BLUE – BENDING”
5.1.2 BEND AND COLLAPSE DETECTION

For all the actions in the training data, we extracted the Fourier action models in an offline step and formed the database as is represented in Figure 21. In an online step, when certain action is to be detected, system searches for an Action Trigger Peak to detect the action culmination. Once it occurs and the Area value is below the Area MHI threshold (which eliminates “Walking” or “Falling” actions), the model of that action is extracted using the Fourier Descriptors. Now to recognize which action it is, system compares the extracted model with all the models from the database. For comparison the Euclidian distance is used. Since Action Trigger Peak lasts several frames, for the final classification of an action all the frames from the specific Action Trigger Peak are compared with the predefined models. At every frame the votes for the three best matching models, with the smallest distance are saved. We now apply the voting scheme and the model (action) with most votes is considered as detected.

![Offline Fourier Models](image1)

**FIGURE 21: OFFLINE FOURIER MODELS**

5.1.3 RESULTS FIRST EXPERIMENTS

We trained our data on 5 randomly selected movies with only single actions (or 10% of all data).

<table>
<thead>
<tr>
<th>Action</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>100,00%</td>
<td>100,00%</td>
</tr>
<tr>
<td>Falling</td>
<td>100,00%</td>
<td>100,00%</td>
</tr>
<tr>
<td>Collapsing</td>
<td>91,67%</td>
<td>95,83%</td>
</tr>
<tr>
<td>Bending</td>
<td>91,67%</td>
<td>95,83%</td>
</tr>
</tbody>
</table>

Table 2 shows the results of the data containing all the actions. The actions “Walking” and “Falling” are with all the data correctly classified using only the Area information of the MHI and the Area of individual Silhouette. The actions “Collapsing” and “Bending” have a slightly lower precision and accuracy. The reason for misclassification is the fact that when a human collapses to the ground it sometimes first bends over which will be classified as bending and not as collapsing. Even though the bending action shows some clear models constructed from the Fourier Descriptors, there is a small chance that “Bending” and “Collapsing” are misclassified.
5.2 SECOND EXPERIMENT: REALISTIC DATA

To prove that our system can work in a real elderly home environment, we tested our system inside of an elderly home during single day.

5.2.1 TOP VIEW CAMERA

For this research we used a self made camera with infrared lights with a fisheye lens to get a big overview of the room. Due to the quality of this camera we did not get a very sharp image, see Figure 22.

![TOPVIEW IMAGE](image)

FIGURE 22: TOPVIEW IMAGE

During testing of this system we found out that the double frame differencing method wasn’t capable of removing the background without leaving a lot of noise. This was due to the blurry image of the camera. Therefore we choose to use the Mixture of Gaussians to remove the background which resulted in a clear silhouette without noise.

5.2.2 WALKING ACTION

For walking detection, the change in the area of the MHI was observed. For this research we divided the data in a training and a test dataset. Using training data, we learned the average values of the change in the area that occurred during the action walking. We observed that the average maximum “Walking” values exceed the “Bending” and “Standing Up” values by 25%, Figure 19. This is used to define the Area MHI threshold. When this threshold has been reached the action can be defined as a walking action.
5.2.3 BENDING & SITTING ACTION

For the bending and sitting actions we could not separate using only the Area information. Therefore we used the Fourier Descriptors to describe the contour of the action at the area trigger peaks. The real data show also similar contour shapes within the actions performed by the elderly (Figure 24).

5.2.4 TESTING & RESULTS

For testing we monitored an elderly person in its household during normal daily activities and during various periods in the day. We monitored the elderly for 10 hours which resulted in an action database of 22 walking actions, 7 Bending actions and 5 Standing Up actions. We trained the action models on all the Bending and Standing action movies. During testing we applied leave one out cross validation method due to the low amount of available data.
Table 3 Detection rate using scientific data

<table>
<thead>
<tr>
<th>Action</th>
<th>Accuracy %</th>
<th>Precision %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>100,00</td>
<td>100,00</td>
</tr>
<tr>
<td>Bending</td>
<td>95,83</td>
<td>91,67</td>
</tr>
</tbody>
</table>

Table 4. Detection rate using real data

<table>
<thead>
<tr>
<th>Action</th>
<th>Accuracy %</th>
<th>Precision %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>93,62</td>
<td>88,00</td>
</tr>
<tr>
<td>Bending</td>
<td>87,50</td>
<td>87,50</td>
</tr>
<tr>
<td>Standing Up</td>
<td>80,00</td>
<td>80,00</td>
</tr>
</tbody>
</table>

Table 3 shows the results using the scientific data in a testing room. The results showed a very high detection rate and were therefore the reason to test this on real data. The results of the real data in an elderly home are shown in table 4. What we can observe is that the detection rate seems to slightly decrease. This can be explained by the conditions of the environment; the elderly room had more noise which made the detection more difficult. The area information gives a high walking detection rate of almost 94% but using the contour models seems to be less accurate due to the low amount of training data. Another problem that we observed is that the Standing Up action is sometimes recognized as a bending action due to the bended pose of the elderly during standing up.

5.3 EXPERIMENT 3: MONITORING REAL ELDERLY HOMES

Using the knowledge of the previous experiments we have chosen a different camera which has an integrated fisheye lens to get a better quality image as is displayed in Figure 25. We enlarged our database by monitoring 4 different elderly people living independently in different houses. These elderly people were monitored during single day which resulted in a database containing 4 classes actions per each patient.

- Walking
- Eating Diner
- Sitting
- Open Door

Looking at these actions we see that the elderly people did not bend, so this action could not be trained. Also the falling and collapsing actions were not performed because of the danger of injury of the elderly people. While analyzing the movies we detected that 2 new actions were performed by the elderly. The action Easting is performed by all 4 elderly people during the monitoring. 3 of the 4 elderly also opened and closed door which were inside of the living room. To analyze these actions we used the Bag of Words method in combination with the action signature and feature fusion technique described in the previous section.
5.3.1 TRAINING & TESTING

Table 5. Final Dataset

<table>
<thead>
<tr>
<th>Human</th>
<th>Eating</th>
<th>Walking</th>
<th>Sitting</th>
<th>Door</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>35</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>23</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>20</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>9</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

Because of the large variety of the number of actions of each person and in order to perform equal training, we randomly selected 5 movies for each action for every patient. To train the system we used the leave one out method.

- First background subtraction is done with the Mixture of Gaussians
- Features are calculated and stored in the Bag of Words

FIGURE 25: EXAMPLE IMAGES IN REAL ELDERLY HOMES
• Actions signatures are trained using the leave one out method using the Euclidian Distance
• Feature Fusion is applied to come up with the final input
• Support Vector Machine is trained for every action using the feature fusions as the input.

Because we train for every action its corresponding support vector machine classifier, we can test every classifier with the movie which is left out using the leave one out method. The resulting output classifies the movie as belonging that that action or not.

5.3.2 INTER PATIENT TRAINING

The first training and testing we did within each patient itself using the leave one out method. The final result will classify the input and is separated as follows:

True Positives : Action classified correctly
True Negative : Action rejected correctly
False Positive: Action incorrectly classified
False Negative: Action incorrectly rejected

Knowing these values we can calculate the Accuracy and Precision of the system:

\[ \text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \]

which will tell use the proportion of true results against all positives

\[ \text{Precision} = \frac{TP}{TP + FP}, \]

which will tell use the proportion of true results against all positives

<table>
<thead>
<tr>
<th>Table 6. Final Dataset Client1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Client1</strong></td>
</tr>
<tr>
<td>tp</td>
</tr>
<tr>
<td>tn</td>
</tr>
<tr>
<td>fp</td>
</tr>
<tr>
<td>fn</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Precision</td>
</tr>
</tbody>
</table>
We can see in the tables 6-10 that the final results shown a very good performance on the detection of multiple actions with different actions. The action “Dining” is classified correctly with 3 of the 4 patients. The reason why with patient 1 the accuracy is slightly lower is because in one of the dining movies the patient is repositioning himself and causes more motion which is therefore classified as sitting instead of dining. Still this dining action has a very high accuracy and can these extra motions can be vital information for the healthcare to detect. Also action eating caries very valuable information for healthcare, since many incidents occur during this action (eg. choking).
The action “Door” is only performed by humans 2, 3 and 4. The results show a high performance with client 4 while, with client 3 all Door movies are classified correctly while 2 other movies are false classified as the action Door. The worst classification with this action happened with person 2, this is mainly because of the room environment of this person. Inside of the room the lightning conditions where much darker then in the other rooms which caused some more noise.

Looking at the actions Walking and Sitting we can see that with all the 4 patients the accuracy is between the 85-95 % which is very high and proven to be better than the method without using the Bag of Words principle, see table 5.

Finally, table with combined measurements shows overall results above 85%, which encourage us to apply such a system in real homes. Also increase of the available training data will further improve the performance of the system.

5.3.3 INTRA PATIENT TRAINING

To see whether this system is dependent on the training specific to one person we trained the system on one client and tested it on the other clients. In this way we get an indication whether actions are dependent on the patient or the training data of one patient can be used for other clients as well. In tables 11-17 the results are presented. When we look at these results we can see that some actions still perform quite well. The action Dining has an average accuracy of 92%. This is mainly because of the similarity of the actions of all the humans. While the action Doors has a lower accuracy is still reaches up to 90% which indicates that this actions is also performed similar by all the humans. Both Walking and Sitting are giving lower accuracy rates because of the large variability of motion of the humans. Some elderly walk with helping aids and differ in speed in size. Also the position of the actions is different for every individual person which will drop the detection rate when you train on one and test on others. Still one can see that the methods used for the action detection are very stable and work for different actions.
### Table 11: Train on Client 2 Test Client 3

<table>
<thead>
<tr>
<th></th>
<th>Dining</th>
<th>Door</th>
<th>Sit</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>85,00</td>
<td>85,00</td>
<td>50,00</td>
<td>80,00</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>75,00</td>
<td>75,00</td>
<td>0,00</td>
<td>60,00</td>
</tr>
</tbody>
</table>

### Table 12: Train on Client 2 Test Client 4

<table>
<thead>
<tr>
<th></th>
<th>Dining</th>
<th>Door</th>
<th>Sit</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>100,00</td>
<td>90,00</td>
<td>80,00</td>
<td>100,00</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>100,00</td>
<td>80,00</td>
<td>60,00</td>
<td>100,00</td>
</tr>
</tbody>
</table>

### Table 13: Train on Client 3 Test Client 2

<table>
<thead>
<tr>
<th></th>
<th>Dining</th>
<th>Door</th>
<th>Sit</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>100,00</td>
<td>70,00</td>
<td>50,00</td>
<td>90,00</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>100,00</td>
<td>40,00</td>
<td>0,00</td>
<td>80,00</td>
</tr>
</tbody>
</table>

### Table 14: Train on Client 3 Test Client 4

<table>
<thead>
<tr>
<th></th>
<th>Dining</th>
<th>Door</th>
<th>Sit</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>100,00</td>
<td>100,00</td>
<td>50,00</td>
<td>65,00</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>100,00</td>
<td>100,00</td>
<td>0,00</td>
<td>25,00</td>
</tr>
</tbody>
</table>

### Table 15: Train on Client 4 Test Client 2

<table>
<thead>
<tr>
<th></th>
<th>Dining</th>
<th>Door</th>
<th>Sit</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>100,00</td>
<td>100,00</td>
<td>100,00</td>
<td>85,00</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>100,00</td>
<td>100,00</td>
<td>100,00</td>
<td>75,00</td>
</tr>
</tbody>
</table>

### Table 16: Train on Client 4 Test Client 3

<table>
<thead>
<tr>
<th></th>
<th>Dining</th>
<th>Door</th>
<th>Sit</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>70,00</td>
<td>90,00</td>
<td>100,00</td>
<td>90,00</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>44,44</td>
<td>80,00</td>
<td>100,00</td>
<td>100,00</td>
</tr>
</tbody>
</table>

### Table 17: Final Average Result

<table>
<thead>
<tr>
<th></th>
<th>Dining</th>
<th>Door</th>
<th>Sit</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>92,50</td>
<td>89,17</td>
<td>71,67</td>
<td>85,00</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>86,57</td>
<td>79,17</td>
<td>43,33</td>
<td>73,33</td>
</tr>
</tbody>
</table>
Due to the ageing population and the decrease of nursing staff, there is an upcoming problem in the healthcare. This causes that elderly people are living independently for a longer period of time. To increase the safety of the elderly people who live independently at home, we aimed to develop a new intelligent camera system. The goal of this development was to make a robust and real time application that can detect multiple actions performed by the elderly in their own living environment.

After doing an intensive literature survey, we started our research by monitoring students in an experimental environment. Using only a simple webcam mounted on the ceiling we monitored different actions:

- Walking
- Bending
- Falling
- Collapsing

Using only 3 different features, Area Silhouette, Area MHI and Fourier Descriptor, we obtained the detection rates of 100% for both Falling and Walking, while the Bending and Collapsing actions were detected with 96%. With this knowledge and high accuracy rates we were encouraged to start the second experiment with a real data. We used a different camera with a fisheye lens mounted on top of the room to get an overview of the entire room. In this way, we reduced the costs of using multiple cameras while obtaining the similar performance. The results of these measurements had some decrease in accuracy due to the real environment challenges but still proved to be successful in the detection of multiple actions. However, both methods used manually predefined thresholds to detect certain actions and did not have any learning network behind. Therefore, we concluded that in order to use system online it must automatically learn which features are useful to use.

To improve the system we finally monitored 4 different elderly people during their daily activities and under different lightning and room conditions. To obtain robust detection system, we described every action using five different features and created learning network that is calculating optimal combination of features for every action. To combine features we first applied Bag of Words method to create dictionary of features and action signatures. We finally applied feature fusion and used the result as the input for the learning network. As network set of parallel support vector machines was used, one for each class of actions that we want to detect. Final result is obtained using combined classifier from different SVMs. In this way, in training the system optimal combination of features for every action is learned. In data analysis, we included two new actions, Dining and Opening Doors. To test the performance of our system we performed two different testing methods. We first wanted to see performance of system if it is customized to specific person so we training and tested the system using data from same patient and obtained accuracy rates of 80-100%. Secondly, we wanted to see if the learned
network on one patient can be transferred to other patients as well. Therefore we trained our system on one elderly person and tested on others. Obtained results show that the action Dining and Doors have the best detection rates, which concludes that these actions are quite similar for different people. The actions Walking and Sitting have some slightly smaller detection rates which shows the variability of these actions for each different person. That can be explained by fact that each person has unique way of moving.

Finally from all obtained results, we can draw following conclusions:

- Camera system that we proposed can be used in healthcare applications to detect dangerous actions and provide safety and care to the elderly
- Multiple actions can be detected in real time with a single camera
- System must be trained on real data from elderly homes, and therefore we obtained database of actions from real elderly homes
- To have an optimal solution, one needs to train the system for each specific person.

6.1 RECOMMENDATIONS

The focus of this thesis was to find out if we can design system that can recognize multiple human actions using a single camera. Looking back at this research we learned that human actions performed by actors/students are different than in a real life. It is very important to gather real data from real elderly in their own living environment. Our final data is therefore recorded with 5 elderly people living independently. In an offline process we learned the system to recognize multiple different actions with very high accuracy rates. From different experiments, we also concluded that the system must be trained for every individual person and adding new class of actions requires retraining the system.

If we take a look at the ultimate goal of this project, we would like to design a system that is self learning and working online, which can not only detect actions but also recognize who is performing that action. In that way multiple users can be monitored in one living room. It must also be possible to recognize living patterns of a person in time. If the system can detect an abnormal pattern, the prevention instead of only the detection of a dangerous action can be performed. Therefore more data is needed obtained from recording person for a longer period of time. The self learning camera must then be able to adjust itself and recognize new actions performed by a specific individual.

Finally, this system will be used in combination with other healthcare devices and implemented within the healthcare alarming systems. In this way can create a complete healthcare system which can be used in healthcare institutes and elderly people living at home. It will provide safety by alarming people when a dangerous situation occurs. Injuries and deceases can be prevented by using the knowledge on living patterns of a person, learned for every individual and by adjusting the care to that persons needs.
REFERENCES


27. D.S. Zhang, J.Lu, Comparative study on shape retrieval using Fourier descriptors with different shape signatures, In Proceedings of the International Conference on Multimedia and Distance Education, Fargo, ND, USA, pp. 1–9, June 2001
REAL TIME FALL DETECTION AND POSE RECOGNITION IN HOME ENVIRONMENTS

Jerry Aertssen, Maja Rudinac and Pieter P. Jonker  
Delft Biorobotics Lab, Delft University of Technology, Delft, The Netherlands

Keywords: Action recognition, Motion detection, Shape descriptors, Home monitoring, Application in elderly care.

Abstract: Falls are one of the major obstacles for independent living of elderly people that can be severely reduced by introducing home monitoring systems that will raise the alarm in the case of emergency. In this paper we present an inexpensive and fast system for fall detection and dangerous actions monitoring in home environments. Our system is equipped only with a single camera placed on the ceiling and it performs room monitoring based on the motion information. After background subtraction, motion information is extracted using the method of Motion History Images and analysed to detect important actions. We propose to model actions as the shape deformations of motion history image in time. Every action is defined with the specific shape parameters taken at several moments in time. Model shapes are extracted in offline analysis and used for comparison in room monitoring. For testing, we designed a special room in which we monitored in various environmental conditions a total of four different actions that are dangerous for elderly people: “walking”, “falling”, “bending” and “collapsing”. Obtained results show that our system can detect dangerous actions in real time with high recognition rates and achieves better performance comparing to the state of the art methods that use similar techniques. Results encourage us to implement and test this system in real hospital environments.

1 INTRODUCTION

The amount of elderly people will strongly increase during the next decennia. The strong increase of elderly people has some social effects especially on healthcare and elderly care. There is a large trend in the displacement of elderly care from healthcare institutes to healthcare at home. Prevent care on elderly is done in order to keep elderly home and independent as long as possible. Falls in elderly population are large hazard for their health and produce high costs for social system as well. Almost half of the fall incidents occur in elderly houses and can be prevented by an adequate monitoring system. There are lots of different factors that increase the chance of fall incidents (Kannus et al, 2005). Because of the decreasing muscle force and movement speed due to the aging, it is harder to keep the body balanced. Besides that, the reaction time decreases which result in reduced ability of elderly person to judge dangerous situations in time.

Current systems for fall detection and prevention that are implemented in healthcare are not able to detect multiple dangerous situations and falls without the help of extra electronic devices mounted on the person. And more problems arise when there are multiple persons inside one room (Close et al, 1999).

The goal of our research is to design an intelligent camera system able to detect multiple actions and falls during day and nighttimes using only one camera. The system must work real time with the intended goal to implement this technique into an embedded system. The monitoring system should be designed in such a way that it warns an elderly person on the dangerous pose or action that he or she is performing and to alert the medical services in the case that actual fall occurs. In this paper, we propose such a system for action recognition that uses a webcam mounted on a ceiling pointing directly down, in order to create a top view image. In this way cluttered scenes are brought back to a minimum and there is a clear distinction between different poses. For testing, we designed a test room that will simulate house environment for acquiring our data. The data used for this research is captured during daytime with different illumination conditions. In total we observed 4 actions “Walking”, “Bending”, “Falling” and “Collapsing” (very slow falling) that were performed by multiple people differing in size and wearing different...
clothing in order to create a realistic dataset. In further text we will describe the methods that we used to design our system.

In section two we discuss the related work which already has been done in this field of work. Section three describes the methods form motion detection that we used to describe actions. Chapter four explains how the actual system is working and how the actions are detected while Chapter five shows the results and gives the conclusion.

2 RELATED WORK

Several different techniques and systems were proposed recently that detect dangerous poses or falls of humans inside a room. Most of those systems make use of accelerometers which detect abnormal accelerations and trigger an alarm. One of the approaches is based on the wearable systems which are able to detect falls. (Zhang et al, 2006) uses a non-negative matrix factorization method for feature extraction. The major advantage of this method is the accuracy of detecting a fall, but the major disadvantage is the fact that people have to wear these devices which results in discomfort. This might cause that after some time the devices will be left a side by the user and a fall will not be detected.

Recently, some researchers proposed to detect falls using camera systems. They use single camera to analyze moving object by background subtraction. To detect a fall, the measurements of the length width ratio of the bounding box are calculated. Their results show that this approach works well and that it is able to detect different poses when the camera is placed sideways (Tao et al, 2005) and (Anderson et al, 2006). However, such approaches experience a lot of problems due to occlusions from objects inside of the room.

Other researchers propose to use 3D cameras in order to get specific coordinates of the human inside of the room with respect to the floor (Diraco et al, 2010). This approach proved to have nice results but because of the use of 3D cameras, it is very expensive for the healthcare home environment where you will need multiple cameras to cover all the rooms. (Nait-Charif and McKenna, 2004) proposed a system which uses only a single omnidirectional camera with a wide angle lens placed on the ceiling. This approach reduces the cluttering scenes and can be used to detect multiple objects in the room to define safe regions. Falls were detected using the ratio between the bounding ellipses. However the main drawback is that they are using ratio information which is not sufficient for multiple pose detection and multiple action detection that we would like to perform.

3 MOTION DETECTION

We can define actions as the change of the motion in time. In our method we propose to describe all the events (or motion changes) belonging to a certain action using only a single image, which can be further modelled using a specific shape descriptor. In that way every action is uniquely described with its representative shape models. Such a method requires several steps, and they are explained in more detail in the next chapters.

3.1 Background Extraction

In order to analyze the action in a certain frame first step is to detect the motion change by removing the background information. There are many different methods to segment the background but the easiest way is to use Frame Differencing. The major advantage of this method is its simplicity and fast computation so it can be applied in real time applications. Another advantage is that it does not require any prior processing and it is independent of the environmental conditions, such as the specific room type or illumination conditions. However, motion segmentation using this method is very coarse and dependant on the shadows in the scene. To eliminate these effects and acquire more accurate results we applied Doubled Frame Differencing. After capturing three successive frames in a video, two separate difference images (‘t-1’ and ‘t’) and (‘t


3.1 Motion History Images

Now we need to capture the sequence of motion change (DDI images) belonging to the one action in a single image. For that we apply the method of Motion History Images (MHI) (Bobick and Davis, 2001). Basic idea is to model the motion by accumulating intensity changes of pixels. Now we can define the intensity as a function of the temporal history of motion at that point. The MHI at time \( t \) is calculated according to Equation 1, where \( D(x,y,t) \) represents DDI image at time \( t \) and pixel position \( (x,y) \). The variable \( \tau \) represents the duration of movement, in consecutive frames, and \( MHI_{\tau}(x,y,t) \) temporal history of motion at point \( (x,y,t) \) occurring during the \( \tau \) frames.

\[
MHI_{\tau}(x,y,t) = \begin{cases} 
  \max(0, D(x,y,t-1) - D(x,y,t-\tau)) & \text{if } \ldots D(x,y,t-1) = 1 \\
  0 & \text{otherwise}
\end{cases}
\]

Resulting MHI is now a scalar valued image where more recently moved pixels appear brighter as can be seen on Figure 3. Such MHI is useful for our application since we only need to know the shape and location of the motion change, not the direction.

3.3 Area Measurements

During movement the silhouettes generated from the DDI images are changing. Therefore also the MHI changes during time. We propose to describe the action by analyzing this change of the MHI. We focused on two different measurements, the area change and the shape change.

In every moment \( t \), we now define and measure two parameters: the Area of the Individual Silhouette (from DDI image) and the Area of the MHI. Since both images are binarized, the area is calculated as the sum of all positive pixels in that image. Now we can measure the differences in the area through time which proves to be very useful for “fall” and “walking” detection. Detailed explanation of the detection method follows in the section Action recognition. Results of the area changes for specific actions are presented in Figure 2a and Figure 2b.

3.4 Shape Measurements

As we already described, we measure the change of the shape of the MHI. At first a contour of a MHI is generated, and afterwards described using Fourier Descriptors. The major advantage of using Fourier Descriptors is because of their invariance on translations, rotations and scale. The contour of a silhouette is described in the frequency domain in such a way that the lower frequencies describe the general contour of the silhouette while the higher frequencies describe the fine detail of the contour. In our application fine details of the contour are not useful for global contour discrimination. Therefore only a subset of the Fourier Descriptors is sufficient to describe the global contour of the silhouette. This reduces the dimension of the descriptors and increases the speed, which is a big advantage for applying it real time.

For a given contour \( s(t) \) which is normalized to \( N \) points, the discrete Fourier transform is given by Equation (2)

\[
F_n = \frac{1}{N} \sum_{t=0}^{N-1} s(t) e^{-j2\pi nt/N}, n = 0,1,\ldots,N-1
\]

This results in a vector of complex numbers where the magnitude of the descriptors is divided by the magnitude of the second descriptor in order to apply scale normalization. This results in:

\[
F_d = \left[ \frac{F_{d_0}}{F_{d_1}}, \frac{F_{d_2}}{F_{d_1}}, \ldots, \frac{F_{d_{N-1}}}{F_{d_1}} \right]
\]

Scale invariance is now obtained by dividing the magnitude values of FDs by the \( F_{d_0} \) component. After that the first descriptor \( F_{d_0} \) is discarded since it only gives information about the position of the contour and it is not describing the contour itself. (Zhang and Lu, 2001).

Shape descriptors are used to model the specific actions by describing the change of the MHI of that action. Detailed explanation of modelling and detection of different actions follows in next section.
4 ACTION RECOGNITION

We now combined the methods described in section 3 in order to make a fast, reliable and efficient action detection system. The scheme of this detection system can be found in Figure 1.

4.1 Action Triggering

As can be seen in Figure 2, we first apply the background subtraction using Double Frame Differencing to obtain the Silhouette of moving object. Using this image, we now calculate the Area of the silhouette image and the Area of motion history image, as explained in previous section. When a person is performing one of the actions we want to detect, we can observe a large increase of both the Area of silhouette and the Area of motion history image. Once the action is finished this area will decrease, which results in area peaks of the silhouette and motion history image. These peaks define when an action has happened and when to analyze it and are referred to as Action Trigger Peaks and are shown in Figure 3.

The Action Trigger Peak is found by subtracting the Area at time $t$ and time $t-1$ checking the resulting value. The positive or negative resulting value corresponds to an ascending or descending slope, as defined in Equations 4 and 5.

$$\text{Slope}(t) = \left( MHI_{\text{Area}}(t) - MHI_{\text{Area}}(t-1) \right)$$  \hspace{1cm} (4)

$$\text{Slope}(t - 1) > 0 \land \text{Slope}(t) < 0 \rightarrow \text{Peak} = t$$  \hspace{1cm} (5)

4.2 Fall and Walk Detection

We generated our data in a home environment where all actions are performed under different illumination conditions and by multiple people wearing different clothes. Analyzing the Action Trigger Peaks of the MHI Area of the different actions, we observed that the average maximum “Falling” and “Walking” values exceed the “Bending” and “Collapsing” values by 20% (Figure 2b). This is used to define the Area MHI threshold. Further analyzing the Action Trigger Peaks of the Area of Silhouettes, we observed that the average maximum values of “Falling” actions exceed the average maximum “Walking” value with 15% (Figure 2a). This is used to define the Area Silhouette threshold. Combining these two thresholds we can distinguish between “Walking” and “Falling”. What is important to notice is that both of these thresholds are learned from manually labelled training data by comparing differences in area values. These thresholds are then applied on the testing data, to detect the “Fall” and “Walking” actions.

4.3 Action/Pose Models

Another way to describe a change in the motion history image is by shape descriptors. We now model an action using an exact shape of the MHI that occurs in the peak of that action. For every
specific action the training data is analyzed and at the Action Trigger Peak, a contour segmentation is performed on the MHI. On this contour we calculate the Fourier descriptors and save them as a model for the specific action. Examples of such models can be found in Figure 4, where it is clearly visible that same actions performed by different people in different conditions preserve same shape information.

Figure 4a: Models for the action “Bending”.

Figure 4b: Models for the action “Collapsing”.

### 4.4 Bend and Collapse Detection

For all the actions in the training data, we extracted the action models in an offline step and formed the database. In an online step, when we want to detect certain action, we first search for an Action Trigger Peak to detect the action culmination. Once it occurs and the Area value is below the Area MHI threshold (which eliminates “Walking” or “Falling” actions), model of that action is extracted using the Fourier Descriptors. Now to recognize which action it is, we compare the extracted model \( FD_{contour} \) with all the models from the database. For comparison the Euclidian distance is used. As defined in the Equation (6).

\[
d = \left| FD_{contour} - FD_{model} \right| = \sqrt{\sum_{i=1}^{N} \left( FD_{contour} - FD_{model} \right)^2}
\]  

Since Action Trigger Peak lasts several frames, for the final classification of an action all the frames from the specific Action Trigger Peak are compared with the predefined models. At every frame the votes for the two best matching models, with the smallest distance are saved. We now apply the voting scheme and the model (action) with most votes is considered as detected.

### 5 TESTING AND RESULTS

For the purpose of this research we designed a special testing room, with the web camera mounted on the ceiling to generate a top view image. We used this setting since it is independent of the motion direction angle while providing visibility of the entire room. We recorded in total 52 movies containing all the dangerous actions that elderly person can perform: Walking, Bending, Collapsing and Falling. All data is recorded in a standard room during daytime under different illumination conditions. For the multiple action data another 16 movies were recorded using the same conditions. To make our testing more robust, we used in total 4 different subjects who all differ in height and wear different clothes. We trained our data on 5 randomly selected movies with only single actions (or 10% of all data) Results are presented for the testing data of both single and multiple action movies.

Table 1: Single Action Results.

<table>
<thead>
<tr>
<th>Action</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Falling</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Collapsing</td>
<td>91.67%</td>
<td>95.83%</td>
</tr>
<tr>
<td>Bending</td>
<td>91.67%</td>
<td>95.83%</td>
</tr>
</tbody>
</table>

Table 1 shows the results of the data containing only one single action. The actions “Walking” and “Falling” are with all the data correctly classified using only the Area information of the MHI and the Area of individual Silhouette. The actions “Collapsing” and “Bending” have a slightly lower precision and accuracy. Confusion matrix shows that these two actions are both misclassified with each other. The reason for misclassification is the fact that when a human collapses to the ground it sometimes first bends over which will be classified as bending and not as collapsing. Even though the bending action shows some clear models constructed form the Fourier Descriptors, there is a small chance that “Bending” and “Collapsing” are misclassified.

In normal daily activity multiple actions can happen after each other which might make classification of an action a more challenging task. Table 2 shows the results of these multiple task. The results show that the action “Walking followed by Falling” is classified correctly in all the data. This is caused by the fact that the calculations of the MHI Area and the Silhouette Area are very stable and clearly distinguish from other actions. Combining the other two action who make use of the models
from the Fourier Descriptors, the Precision seems to drop to 83.33%. This is mainly caused by the fact that the MHI contour during “Bending” still contains information of the Walking silhouettes. This extra information changes the contour of the MHI contour. This change in contour deteriorates the results compared with the predefined contours of the “Bending” and Collapsing actions.

Table 2: Multiple Action Results.

<table>
<thead>
<tr>
<th>Action</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking -&gt; Fall</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Walking -&gt; Bending</td>
<td>83.33%</td>
<td>91.67%</td>
</tr>
<tr>
<td>Bending -&gt; Walking</td>
<td>91.67%</td>
<td>95.83%</td>
</tr>
<tr>
<td>Walking -&gt; Bending -&gt; Collapse</td>
<td>83.33%</td>
<td>91.67%</td>
</tr>
</tbody>
</table>

Table 3: Single Action results for (Tao, 2006) method.

<table>
<thead>
<tr>
<th>Action</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>80%</td>
<td>88.89%</td>
</tr>
<tr>
<td>Falling</td>
<td>75%</td>
<td>85.71%</td>
</tr>
</tbody>
</table>

If we compare our results with the state of the art we outperform other techniques based on the bounding box ratio. Table 3 shows the results using the bounding box ratio principle as used in (Tao et al, 2005) and (Anderson et al, 2006) on our dataset. The actions “Walking” and “Falling” are being detected but with a drop in precision and accuracy, while the actions “Bending” and “Collapsing” couldn’t be detected on our dataset at all. Using only the bounding box ratio proved not to be successful on our dataset, but since it shows that in some cases falling and walking can be detected, the combination of using the bounding box ratio together with our method can be promising. If we look at the other methods such as the wearable devices discussed by (Zhang et al, 2006), we achieve slightly better performance on a very similar dataset but these devices have a drawback that patients need to wear them all the time which is very uncomfortable. Regarding speed, our system achieves real time performance.

6 CONCLUSIONS

The research presented in this paper is related to human fall detection and the detection of different actions that can be dangerous for elderly people. Our main goal was to design a system which can work in real time applications and reduce the implementation costs by using only one web camera. Our system is able to detect and distinguish different actions by using a size and shape information of the motion history image that characterizes certain action. It outperforms other methods based on background subtraction and pose recognition using silhouette information. It also gives very high fall detection results, works in real time and is very inexpensive to implement. For further development we plan to implement it in an embedded system and test it in different nursing home environments.

REFERENCES


Close, J. Ellis, M., Hooper, R., Glucksman, E., Jackson, S. Swift C. (1999). Prevention of falls in the elderly trial (PROFET): a randomised controlled trial. Department of Health Care of the Elderly, Guy’s King’s, and St Thomas’ School of Medicine, King’s College Hospital, London, UK.


FALL AND ACTION DETECTION IN ELDERLY HOMES

Jerry Aertssen, Maja Rudinac, Pieter Jonker
Delft Biorobotics Lab, Delft University of Technology, The Netherlands

Objective
Injuries caused by falls of elderly people are a common worldwide problem and ageing of population will even further increase related burdens and costs. In this paper we present a monitoring system that detects dangerous actions of elderly and raises an alarm when necessary. Such a system would improve the quality of life for elderly and reduce costs in healthcare in general. The main requirements in its design are that the system can perform in real time with a high detection rate using realistic data, and to make it affordable so that it can be used as a healthcare monitoring system.

Main content
Our system is equipped with a single wide angle camera mounted on the ceiling of an elderly home. This gives a topview image of the environment resulting in a clear map of household objects without any occlusions. The main idea is to monitor the motion information of elderly and to model actions as change of motion or poses in time that lead to a specific action. After background subtraction, the motion information is extracted using the Motion History Images method and analyzed to detect important actions. We propose to model actions as the shape deformations of the motion history image in time. Every action is defined with the specific shape parameters taken at several moments in time. Shape models are extracted in offline analysis and used for comparison in room monitoring. For testing, we used a real elderly home environment in which we monitored in various environmental conditions a total of three different actions that are dangerous for elderly people: “Walking”, “Bending” and “Getting Up”.

Results
The obtained results from observing real elderly household show that our system can detect dangerous actions in a real time with high recognition rates. It can separate Walking, Bending and Standing Up actions with detection rates up to 94%.

Conclusion
Our system is able to detect multiple actions by analyzing the size and shape information of the motion history image. Because of the high recognition rates we can conclude that the presented method has a large potential to be further implemented in a real healthcare monitoring system which will increase the safety of elderly and improve their quality of life. We are now gathering data from various elderly homes to test our system, further improve detection rates and get a better understanding of actions that are happening. The presented system performed both training and testing on data obtained form a same household environment. We are now working on a design of a more advanced algorithm that will be able to transfer models learned in one environment to other households, which will allow a wider applicability of our system.

Keywords. action recognition, motion detection, shape descriptors, home monitoring, application in elderly care, elderly environment
1. Introduction

In the next 30 years there is a large shift in the human population. While today the majority of the population is still young, it will soon change to a majority of population which is above the age of 65. Elderly people are more vulnerable to physical and psychological impairment but due to this large shift, less young people will be available to nurse the elderly. These facts require new assistive technologies that will improve the quality of life and reduce the working burden of the caregivers. Research has shown that elderly people prefer to stay independent and live in independent household as long as possible [1,2] and therefore assistive technologies need to provide safer environments.

Injuries caused by falls of elderly people are a common worldwide problem and the ageing populations will even further increase related burden and costs. Around 30% of the people aged 65 years or older living in the community and more than 50% of those living in residential care facilities or nursing homes fall every year, and about half of those who fall do so repeatedly [3]. Falls account for over 80% of injury-related admissions to hospital of people older than 65 years. Fractures, severe head injuries, joint distortions and dislocations, soft-tissue bruises, contusions, and lacerations, arise in 5–10% of falls. These percentages can be more than doubled for women aged 75 or older (Figure 1) [3,4]. Injury is also the fifth leading cause of death in elderly adults and most of these fatal injuries are related to falls. Only in the Netherlands 92,000 elderly need medical care after a fall, 36,000 of them must stay in the hospital and nearly 1800 patients die each year. This leads to an economic cost of €5600; for each fall incident [3].

Current systems for fall detection and prevention implemented in healthcare are not able to detect multiple dangerous situations and falls without the help of extra electronic devices that person needs to wear constantly. Also, these methods can not cope with situations when there are multiple persons inside one room [12]. The goal of this research is to design an intelligent camera system which is able to detect multiple actions and falls during day and nighttimes using only one camera. The monitoring system should be designed in such a way that it warns an elderly person for the dangerous pose or action that he or she is performing and to alert the medical services in case an actual fall occurs. In this paper, we propose such a system for action
recognition that uses a wide angle camera mounted on a ceiling pointing directly down, in order to create a top view image (Figure 2). In this way cluttered scenes are brought back to a minimum and there is a clear distinction between different poses. For testing, we monitored a voluntary elderly during day in different parts of the room. During this monitoring we captured 3 different actions: Walking, Bending and Standing up.

![Figure 2: Topview Images](image)

In this paper we will describe the methods that we used to design our system. In section two we discuss the related work which already has been done in this field. Section three describes the methods of motion detection that we used to describe actions. Chapter four explains how the actual system is working and how the actions are detected while Chapter five shows the results and gives the conclusion.

2. Related work

A commonly applied method to detect human falls is with the help of electronic wearable sensors. These sensors use the technique of accelerometers to detect a fall. During a fall, the displacement of the human body is faster and differs from normal daily activity which can be measured with accelerometers [5]. The major advantage of these devices is that they work independently of the surrounding of the elderly and are therefore relatively cheap. The disadvantage is that they are not accurate enough which results in many false alarms and the fact that they must be constantly worn resulting in reported discomfort for the elderly. Another problem is that elderly often forget to wear these devices or do not want to wear it due to unfashionable and cumbersome design and therefore the fall can not be detected.

Recently several methods were developed that detect falls by using camera systems placed inside the home environment of the elderly. Such systems use a single camera mounted in the corner of a room at the table height. The human actions are then analyzed by observing the ratio between the height and length of the moving person. When the human is walking, the length of the person silhouette will be larger than the width because of the shape of the human body. During a fall the width of the silhouette will increase and the height will decrease. The advantage of such an approach is that is very simple and fast and therefore can be applied in the real time systems. The major drawback is the fact that objects inside of the room might block the camera field of
sight and therefore the moving person can not be fully detected [7,8]. In our research we propose to overcome this problem by using the topview of a room with a wide angle camera and by designing more advanced modeling techniques we extended the number of detected actions. In [6] we presented a method for fall detection and pose recognition and tested it using scientific data recorded in a specially designed testing room. In this paper we adapt our method to a real environment and report results of monitoring real elderly household during longer time periods.

3. Motion Detection

The first step in analyzing human behavior using a camera system is to define what an action is. We define actions as the change of a motion in time and propose to describe all events which belong to a certain action using only a single image. These single images can now be analyzed by observing their exact size and shape.

3.1. Removing the background

We are detecting actions using the motion information. The first step in detecting motion is removing the background or all the pixels in image that are not changing in time. There are many different methods to remove the background, but one of the most stable methods is the “Gaussian Mixture Model”. This model describes the variation in the background by modeling the pixel intensity using a mixture of Gaussian distributions. In the case that a mixture component occurs frequently and does not change much, it will be set as the background. The background is first learned from several subsequent frames and then it is subtracted from each frame to obtain motion information. To achieve robustness of the method, after some time the background should be relearned because the conditions can change in time [9]. In our system were we have a very stable indoor environment this method works very efficiently and in real time.

3.2. Motion History Images

In order to recognize an action in a certain amount of time we use the method of Motion History Images (MHI) [10]. When the background is removed from the image, the only remaining part is a silhouette of the moving object. This silhouette is stored and added to a sequence of silhouettes from preceding frames to get the total MHI. Such approach is very useful for our application since we only need to know the shape and location of the motion change.

3.3. Area Measurements

During movement the silhouettes are changing constantly. Therefore also the MHI changes during time. We measure the changes in MHI by observing two different parameters, the change of the area of MHI as well as the change of the shape. If we observe the silhouette of MHI, we can define its area as the number of positive pixels in the image. Now we can measure the differences in this area through time which
proves to be very useful for the “walking” detection. Results of the area changes can be found in Figure 3.

3.4. Shape Measurements

Another parameter that can be measured is the shape of the contour of MHI. We describe the shape of each action using the Fourier Descriptors that define shapes in the frequency domain. The major advantage of this method is that the shape description is invariant to translation, rotation and scale. To reduce the computation time we use only the lower frequencies, which describe the basic shape of the contour with no loss in performance. For a more detailed explanation please refer to [6].

4. Action Recognition

To detect the dangerous actions we used the combination of the techniques described above.

4.1. Action Triggering

The first step in modeling different actions is detecting when to start the measurement. One of the parameters that can be observed is an increase in the area size of the MHI when the action starts. When the action is finished, this area is decreasing, which result in an area peak. These peaks define exact time periods during which the action has happened and can be used as a trigger to the system to start analyzing the action. We refer to them as the Action Trigger Peaks. Examples can be found in Figure 4.
4.2. Walking Detection

For walking detection the change in the area of the MHI was observed. For this research we recorded the data in a real elderly home environment and divided it in training and a test dataset. Using training data, we learned the average values of the change in the area that occurs during the action walking. We observed that the average maximum “Walking” values exceed the “Bending” and “Standing Up” values by 25% (Figure 3). This is used to define the Area MHI threshold. When this threshold has been reached the action can be defined as a walking action.

4.3. Action Models

For the other two actions, the area information alone is not enough to separate them. Therefore, we describe each action using the exact shape that MHI acquires during that action, or in other words using sequence of poses that an elderly takes while performing that action. Shape models of each action are calculated offline using Fourier descriptors on the training data and stored in a database. In an online step, at the action trigger peak, a contour is extracted from the MHI, shape descriptors are calculated and the resulting model is compared with the models from the database. The model that best fits the new motion is defining the actual action that has happened. The results of the models used for detection can be found in figure 5.

Figure 5a: Models for the action "Standing Up"

Figure 5b: Models for the action "Bending"

5. Detection of the safe area

Another part of this research is the detection of safe areas in the room. Dangerous actions such as laying or bending are allowed and safe within these areas. For example if an elderly person is sleeping in a bed, it can often be confused with falling on the floor if such areas are not defined. Therefore we also worked on a design of an algorithm that will detect various objects in the household and define safe actions around these objects. For bed detection we used method proposed by [11], Histogram of Oriented Gradients (HoG). An image is divided in small regions called cells and on
each cell a histogram of gradient directions or edge orientations over the pixels of the cell is calculated. This results in a distribution of each cell which represents local object appearance and shapes. The system is then trained with images of various different beds and HoG features are extracted. In an online step, every cell in an image is compared with a database and cells containing beds are localized. In Figure 6 we show results that we obtained on an online available database. We are now collecting data from a various household environments using the topview wide angle camera which we plan to use for the creation of the database with various furniture models.

Figure 6: Bed detection results

6. Testing and Results

For testing we monitored an elderly person in her household during normal daily activities and during various periods in the day. We monitored the elderly during 10 hours which resulted in an action database of 22 walking actions, 7 Bending actions and 5 Standing Up actions. We trained the action models on all the 7 Bending and 5 Standing action movies. During testing we applied a leave one out cross validation method due to the low amount of available data.

Table 1. Detection rate using scientific data

<table>
<thead>
<tr>
<th>Action</th>
<th>Accuracy %</th>
<th>Precision %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Bending</td>
<td>95.83</td>
<td>91.67</td>
</tr>
</tbody>
</table>

Table 2. Detection rate using real data

<table>
<thead>
<tr>
<th>Action</th>
<th>Accuracy %</th>
<th>Precision %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>93.62</td>
<td>88.00</td>
</tr>
<tr>
<td>Bending</td>
<td>87.50</td>
<td>87.50</td>
</tr>
<tr>
<td>Standing Up</td>
<td>80.00</td>
<td>80.00</td>
</tr>
</tbody>
</table>

Table 1 shows the results using the scientific data in a testing room. The results showed a very high detection rate and were therefore the reason to test this on real data. The results of the real data in an elderly home are shown in table 2. What we can observe is that the detection rate seems to slightly decrease. This can be explained by the conditions of the environment; the elderly room had more noise which makes the detection more difficult. The area information gives a high walking detection rate of almost 94% but using the contour models seems to be less accurate due to the low amount of training data. Another problem that we observed is that the Standing Up action is sometimes recognized as a Bending action due to the bended pose of the elderly during standing up. We are now observing several elderly homes to acquire more data and obtain better action models so we expect that this problem will be solved in future.
7. Conclusion

The research presented in this paper is related to human fall detection and the detection of different actions that can be dangerous for elderly people. Our main goal was to design a system which can work in real time applications with reduced implementation costs due to the use of only one camera. Our system is able to detect multiple actions by analyzing the size and shape information of the motion history image. Because of the high recognition rates on a data obtained by observing an elderly person in her home we can conclude that this method has potential to be used in a real monitoring system that will improve the quality of life for the elderly. For further development we are still gathering data from elderly homes to improve the detection rates. We are also working on a design of a more advanced algorithm that will be able to transfer models learned in one environment to the other households which will allow wide applicability of our system.

References