Pers Ubiquit Comput (2013) 17:1063–1072
DOI 10.1007/s00779-012-0552-z

ORIGINAL ARTICLE

Introducing the use of depth data for fall detection

Rainer Planinc · Martin Kampel

Received: 23 January 2012 / Accepted: 25 March 2012 / Published online: 13 May 2012

Abstract Current emergency systems for elderly contain at least one sensor (button or accelerometer), which has to be worn or pressed in case of emergency. If elderly fall and loose their consciousness, they are not able to press the button anymore. Therefore, autonomous systems to detect falls without wearing any devices are needed. This paper presents three different non-invasive technologies: the use of audio, 2D sensors (cameras) and introduces a new technology for fall detection: the Kinect as 3D depth sensor. Our fall detection algorithms using the Kinect are evaluated on 72 video sequences, containing 40 falls and 32 activities of daily living. The evaluation results are compared with State-of-the-Art approaches using 2D sensors or microphones.

Keywords Fall detection · Depth sensor · Kinect · Autonomous system

1 Introduction

Emergency call buttons are provided by caretaker organizations and have the main drawback that no information is available about an occurred incident prior the button press. Moreover, people suffering from dementia are not able to react on emergency situations properly [20]. In case of an emergency and if elderly are able to press the button, they have to convey incident details to the operator. If the elderly is not able to talk to the operator for any reason (e.g., due to the lost of consciousness), there is no information about the type of incident at all. Furthermore, especially when dealing with dementia, it is important to reduce the cognitive load on the user [23]. Hence, sensors acting autonomously are needed.

In the field of smart homes autonomously acting sensors are used to fulfill core functions [11]: system control, emergency aid, water and energy monitoring, automatic lighting, door surveillance, cooker use safety, etc.... Due to various reasons summarized by Aldrich [2], smart homes have not been established yet. One of the reasons is costs [2]: it is both easier and less expensive to integrate smart home technology into new than into already existing buildings. This results in a demand for a robust system, which can be integrated into existing buildings easily. Furthermore, assistance means covert assistance regarding physical or intellectual impairment for as long as possible, being hidden from visitors, especially ones not belonging to the family or the innermost circle of friends. A small, unobtrusive system would fulfill that demand [24].

Approaches in the field of Ambient Assisted Living assist elderly to enable them to stay at home independently. These systems are dealing with the management of chronically disease (e.g., diabetes [13]) or activities of daily living (e.g., managing medication [16]). In contrast to these approaches, our approach enhances the safety of elderly. As falls are considered to be a major risk for elderly, Willems et al. showed an overview on automatic fall detection [36]. Not only the falls itselves, but also the consequences of a fall are a risk, especially for elderly. Noury et al. [25] have shown that getting help quickly after

This work was supported by the European Union within the AAL-JP project "fearless" under grant AAL 2010-3-020.

R. Planinc (✉) · M. Kampel
Computer Vision Lab, Vienna University of Technology, Favoritenstrasse 9-11/183-2, 1040 Vienna, Austria
e-mail: rainer.planinc@tuwien.ac.at

M. Kampel
e-mail: martin.kampel@tuwien.ac.at

a fall reduces the risk of death by over 80 % and the risk of hospitalization by 26 %.

Considering these facts, the use of computer vision and audio is feasible as they are able to overcome the limitations of other sensor types [24], but raising privacy issues. Different attempts to detect falls using audio exist, for example, Litvak et al. make use of sound and floor vibrations to detect falls [22]. Computer Vision algorithms detect falls by, for example, using a 3D reconstruction of combined 2D images using multiple cameras [4] or from the change of human shape [31]. When using vision-based approaches, privacy aspects need to be considered, but according to Mihailidis et al. very little research on privacy issues has been conducted so far [24]. Furthermore, other limitations (e.g., camera field of view, occlusions) need to be considered.

The rest of this document is structured as followed: Sect. 2 provides an overview of the State-of-the-Art. The methodology is shown in Sect. 3, and an evaluation can be found in Sect. 4. Finally, a conclusion is presented in Sect. 5.

2 State-of-the-art

We propose to classify fall detection approaches into the following categories, depending on the technology to be used: (1) wearable sensors (e.g., accelerometers to analyze acceleration for fall detection [21]), (2) robots (e.g., a robotic dog following elderly and detecting falls [6]), (3) audio-based approaches using microphones, (4) 2D sensors providing pictures (cameras), (5) 3D sensors providing depth information (e.g., Kinect). As we are only dealing with stationary sensors, only the latter three approaches will be discussed.

2.1 Audio-based approaches

Audio provides information for activity and event detection, and thus, several sound classes (e.g., door sound, human sound, baby noise, and loud noise) can be differentiated [27]. A fall detection system combining audio information together with accelerometers was proposed by Doukas et al. [8]. The sensors are attached to the person's body (i.e., foot), and movement is classified into walking, falling and running using a support vector machine. The main drawback of this system is the need for wearing sensors.

An approach proposed by Litvak et al. [22] detects falls not only by sound, but also by floor vibrations and uses non-wearable devices, thus overcoming the drawback of Doukas et al. [8]. The accelerometer (to detect vibrations of the floor) and the microphones are placed in one corner of the room, and falls are detected using an energy-based event detection algorithm. Afterward, the event is classified to distinguish human falls and falls of objects. For evaluation purposes, a human mimicking doll has been used, resulting in a sensitivity and specificity of 95 % each. These results are promising, although the fall of a human mimicking doll cannot be compared to a fall of a person. Hence, further evaluation of this approach is needed.

Popescu et al. proposed the use of one-class classifiers [28], that is not differentiating between different classes but focusing on only one class: the class of falls. They argue that it is possible to specify only the class of falls, as it is impossible to define all events that are no-falls. Therefore, all other events that are not classified as a fall are not a fall event. Their results are obtained by using a fuzzy logic-based approach combined with height information obtained by an audio array, as it has been shown that information about the height of the sound reduces the false alarm rate [27]. The idea of integrating the height of the sound seems feasible, although the false alarm rate still needs to be reduced to be used in practice.

2.2 2D sensors

The general methodology of a computer vision–based fall detection systems using 2D sensors (cameras) is described in Willems et al. [36]. The first step is the separation of people from the background, which is achieved by means of motion detection and background subtraction. Once the person is detected within the video, different kinds of fall detection approaches are used.

When using a 2D sensor, only pictures of the person are available. The 2D shape of a person implies the orientation and thus is used to distinguish whether a person is in an upright position or not. The use of the bounding box aspect ratio (width to height ratio) to detect falls is proposed by Anderson et al. [3]. If people are in an upright position, the bounding box aspect ratio is bigger than one. In case of a fall, the ratio rapidly changes to a value smaller than one. Another approach presented by Rougier et al. uses information of an approximated ellipse instead of a bounding box [31]. Falls are detected by analyzing the orientation of the ellipse as well as the ratio of the major axis of the ellipse. Figure 1 depicts the shape of a person during a normal activity and during a fall. Furthermore, the corresponding bounding boxes and ellipses to analyze the bounding box aspect ratio and the orientation of the ellipse are illustrated. The use of a bounding box and an approximate ellipse for fall detection is feasible, but depends on the quality of the background segmentation. Assuming that the background segmentation yields in robust results, the fall detection also yields in robust results. A fall into the direction of the camera cannot be recognized by both

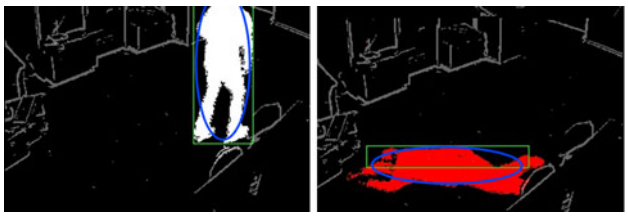

Fig. 1 Analysis of the bounding box aspect ratio and the orientation of the ellipse to detect falls

approaches, as the change of orientation of the person cannot be detected.

Traditional approaches (i.e., not using 3D sensors) making use of 3D information by reconstructing humans from silhouettes gained by different camera views [4]. Hence, the human is represented by the use of voxels allowing to identify different states (upright, on–the–ground and in–between). The quality of this approach also depends on the quality of background segmentation, but having the main drawback of needing a calibrated camera setup. Another approach by Rougier et al. [30] uses 3D information obtained by using one single camera to track the head of the person and to obtain its trajectory. Not only the head position but also the motion speed is taken as an indicator for falls as the motion speed during a fall is typically higher than during usual activities of daily living.

Zambanini et al. propose a method to detect falls and distinguish between a single camera (2D) and a multiple camera (3D) approach [37]. If using a 2D approach, scene analysis is performed on each camera individually. Afterward, the individual results are combined to get an overall decision, as shown in Fig. 2. This approach is called late fusion, as the combination of information of different sources applies at a late stage of the processing pipeline. If information from multiple cameras is combined to reconstruct the person in 3D space, the combination takes place at an early stage. Feature extraction is done on the 3D reconstruction of the person, and a decision whether a fall occurred or not is made afterward. This approach is depicted in Fig. 3 and is called early fusion approach. Compared to other works (e.g., [1]), their system is not vulnerable to low-quality images (e.g., high noise and low resolution) as only basic information (i.e., silhouettes) is extracted from the image anyway.

The approaches of Rougier et al. [30] as well as Zambanini et al. [37] consider motion speed to detect falls, as they assume that the velocity is higher during a fall than during activities of daily living. From our point of view this assumption should not be made, as falls can also occur slowly and thus are not detected using these approaches.

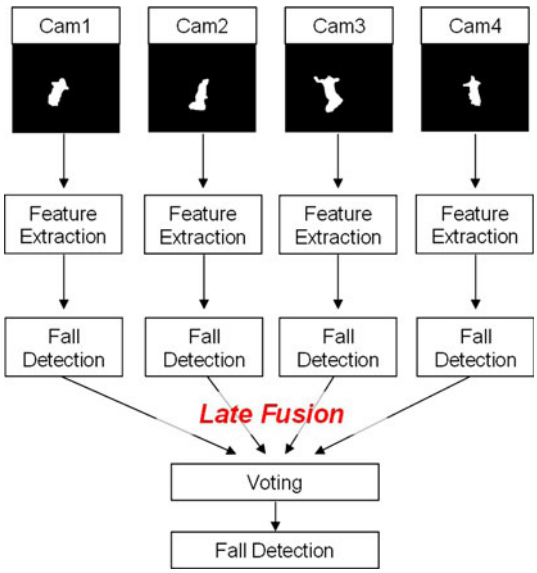

Fig. 2 Late Fusion for multicamera fall detection taken from [37]

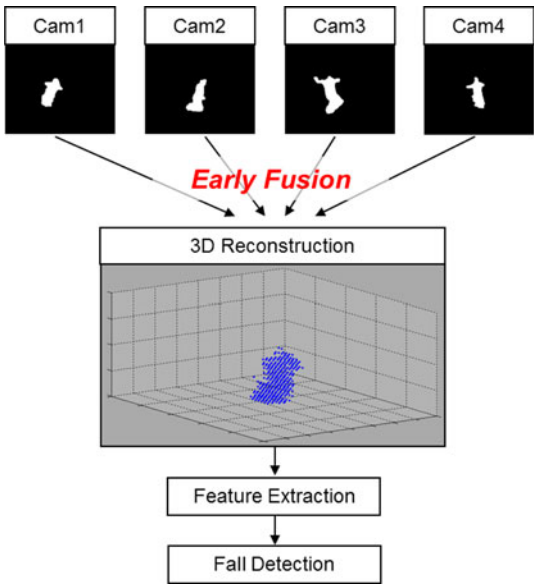

Fig. 3 Early Fusion for multicamera fall detection taken from [37]

2.3 3D sensors

Stereo vision sensors to detect falls are used by Belbachir et al. [5]. These biologically inspired sensors feature a massively parallel pre-processing and reduce the amount of data in comparison with stereo vision cameras dramatically as they are not frame based, but event based. Hence, the motion of people can be determined and the position of the person can be extracted. A fall is detected by tracking the position and velocity of the head, as the position of the head changes rapidly during a fall.

Time-of-Flight cameras [26] are generating depth maps and can be used for fall detection [7]. As a first step, moving regions are detected within the 3D points cloud. The person (foreground) is segmented from the background, and the distance of the person's centroid to the ground floor is analyzed. This results in an efficiency of 80 % and a reliability of 97.3 % when using a centroid-ground floor distance of 0.4 m as threshold [7]. Furthermore, they propose to extract the skeleton from the depth data to analyze the orientation of the person's spine. Another approach using time-of-flight cameras mentions the higher accuracy in contrast to stereo vision [15]. Jansen et al. propose a system for pose recognition, discriminating the poses standing, sitting or lying by thresholding the height of the centroid [15]. They state that their proposed approach works in nursing homes reliably, but not in real homes due to false alarms.

Since the introduction of the Kinect sensor in 2010, a new 3D sensor is available. Stone et al., for example, use the Kinect for obtaining measurements of temporal and spatial gait parameters [10]. Using the Kinect sensor for fall detection was proposed by Rougier et al., but they focus on low-level vision tasks like foreground/background segmentation and detecting the ground plane [29]. Their proposed fall detection algorithm analyzes the distance between the centroid of the body and the ground floor as well as motion speed. As already stated before, motion speed is not a suitable feature for fall detection as the motion speed is not necessarily high during a fall.

3 Methodology

3.1 Audio-based approaches

A multistage approach is performed to recognize events from audio data, consisting of the following steps [38]:

1. *Silence elimination*: first audio is checked for a processable signal in order to prevent further processing of audio in case of only silence. This is achieved by comparing the audio power against a threshold value estimated from a long-term analysis for each microphone.
2. *Feature extraction*: audio signals are represented in either time (time-amplitude representation) or frequency domain (frequency-magnitude representation). Features used in the time domain are the average energy, zero crossing rate or silence ratio, and bandwidth, energy distribution or harmonicity in the frequency domain [32, 33].
3. *Audio pre-classification*: audio data are classified into common types of audio such as speech, sounds or noise. This is done by either using each feature individually in different classification steps or using a set of features combined to a vector to calculate the closeness of the input to the training sets.
4. *Final audio classification*: based on the output of the previous step, each specific audio type is further processed in a different way. For speech recognition, techniques based on Hidden Markov Models [12] are applied as they obtain high recognition performance. For sound recognition, Dynamic Time Warping and Artificial Neural Networks have shown promising results in the past.

3.2 2D sensors

To be able to visually detect risks, the following steps are applied [36]:

1. *Motion detection*: First, motion detection is performed on the video to segment motion (e.g., the person) from the background. For this purpose, a robust background model has to be established, which is able to adapt to changing conditions (e.g., lighting) as well as to reject motion in the background (e.g., a TV). A recently upcoming and promising concept for background modeling is boosting [14], which permits the rejection of recurrent motion in the background during run-time without any presumptions. To increase the system robustness, color information is also exploited for shadow detection [17].
2. *Feature identification*: According to the different risks that have to be detected, a collection of features is extracted. Fall detection requires features describing the human posture [4]. Specific actions are detected by the use of space-time interest points [19]. Other risks such as smoke are detected for example, by using a wavelet transformation or dynamic texture change as feature.
3. *Risk detection*: Different risks (e.g., falls, fire, flooding,...) are pre-defined to interpret the appropriate features and relate them to the risks by using confidence values. Zweng et al. define empirical, semantic driven rules using features with fuzzy boundaries introduced in [9] to analyze the scene and make the decisions [40]. The final decision for a single camera is made by a voting step, which combines the individual confidences. By the use of multiple cameras, the overall robustness and reliability of the system is increased since the voting neglects individual wrong detections. Moreover, the problem of occlusions (e.g., by furniture) is solved implicitly.

3.3 3D sensors

When using the Kinect as 3D sensor, the low-level vision tasks motion detection, foreground/background

segmentation as well as pose estimation are preprocessed in the Software Development Kit[1] (SDK) [34]. As a result, high-level data (i.e., coordinates of specific body junctions) are accessed directly. Since the use of the integrated preprocessing steps offers high-level data, no low-level vision algorithms (e.g., foreground/background segmentation) need to be applied any more.

As the Kinect is a 3D sensor, depth information is available and thus having the main advantage of localizing features in a 3D space. Our proposed algorithms use the orientation of the person's major axis and the height of the spine (relative to the ground floor) as features. In contrast to other works [29], feature analysis is not done on a low-level using the camera picture or the depth picture, but the proposed features are directly applied to the skeleton information.

To determine whether the person is in an upright position or not, the orientation of the major axis based on the body joints position is calculated. To be able to distinguish between similar activities, for example, falling to the ground and lying down in the bed (in both scenarios the orientation of the body is the same!), the height of the spine is used as additional feature. Therefore, we propose and analyze the following two approaches: (1) mapping the 3D body joint coordinates to the 2D depth image and calculating the features using image coordinates; (2) analyzing features directly in the 3D space using world coordinates. Currently two different thresholds are used: a similarity threshold as well as a threshold for the height of the spine.

3.3.1 3D sensor using image coordinates

This approach uses the 3D skeleton information and maps the coordinates to a 2D image space. The orientation of the major axis is calculated using the coordinates of the head, shoulder, spine, hip and knee joints. Using the least squares algorithm to fit a straight line to the data points results in the orientation of the major axis. Afterward, the angle between this line and the horizontal line is calculated. For calculating the height of the spine, an estimation of the ground plane is needed. The ground plane is estimated using the v-disparity map [18, 39]. The basic idea of this approach is that the depth linearly increases on the ground floor. Hence, the depth information of all pixels is analyzed and those pixels having a linear increase in depth are part of the ground plane. This approach assumes that the ground plane is visible in the depth map. After creating the ground plane estimation (which only needs to be done once per scene), the distance of the spine to the ground plane is calculated.

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

3.3.2 3D sensor using world coordinates

Our fall detection algorithm calculates the major orientation of the persons body in 3D space by using the skeleton information. For calculation of the orientation, the head, shoulder (center), spine, hip and the mean position of the knees were taken into consideration. Furthermore, the 3D ground floor is estimated and the spine distance to the ground floor is calculated. If the major orientation of the person is parallel to the ground floor and the height of the spine is near the ground floor, a fall occurred.

Using 3D depth data and world coordinates overcomes the limitations of 2D camera approaches, for example, the problem of falling in the direction of the camera as the distance to the camera is analyzed. Figure 4 depicts the similarity of a person in an upright position and a fall in direction of the camera. Using a 2D single-camera approach and the orientation of the major axis as single feature, a fall in direction of the camera cannot be recognized, because the orientation of the major axis does not change. Figure 5 illustrates the depth image with a person and the corresponding major axis.

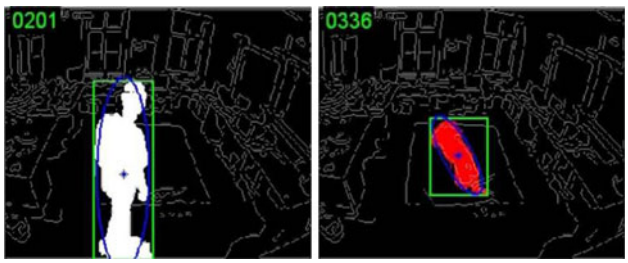

Fig. 4 Fall in direction of the camera

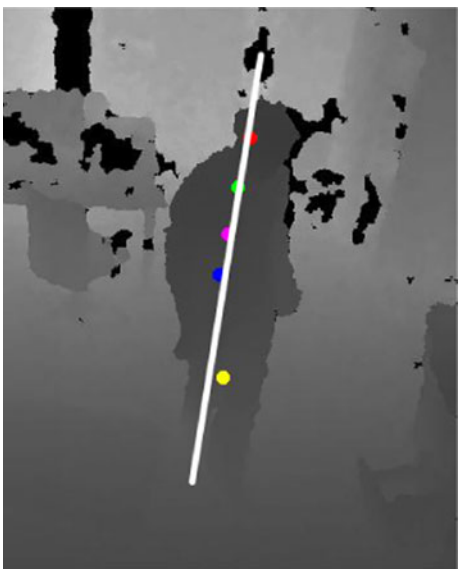

Fig. 5 Major axis calculated using data obtained by the Kinect

4 Evaluation

This section provides an evaluation of our developed fall detection algorithms using image and world coordinates of skeleton points obtained by the use of a 3D sensor (Kinect). Furthermore, the results are discussed and compared with other approaches.

4.1 Experiments using a 3D sensor (Kinect)

To be able to evaluate the fall detection algorithm, the video data are annotated to obtain ground truth information. Therefore, the frame number where the fall began and the frame number where the person is in an fully upright position are annotated. A true positive (TP) of the algorithm is obtained if it detects the fall between the first frame where it began and the last frame, where it ends. As the video sequences do not only include falls but also activities similar to falls, true negatives (TN) are marked (there is no fall, and the fall detection algorithm does not detect a fall). Furthermore, false positives (FP) and false negatives (FN) are analyzed by examining the results of the algorithm. False positives mean that the person does not fall, but the algorithm detects a fall; a false negative is obtained if the person falls, but the algorithm does not detect it. Each video sequence contain one fall at most, but it is possible that the algorithm results in a TP (i.e., fall detected correctly) and a FP (i.e., fall detected without a person falling) within the same video sequence.

The quality of the algorithm is measured using the standard measurements recall, precision, F-score, true negative rate and accuracy. They are defined as follows [35]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

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

$$\text{True negative rate} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{FN}}.$$

All tests have been conducted under laboratory settings, the room setup is shown in Fig. 6. The size of our laboratory is approximately 7 × 6 m, whereas the camera field of view was set to an area of approximately 5.5 × 5.3 m. The Kinect sensor was placed in the middle of the wall at a height of 2.4 m, which is a typical position for surveillance cameras. One frame of the room setup using the Kinect to illustrate the camera field of view is shown in Fig. 7.

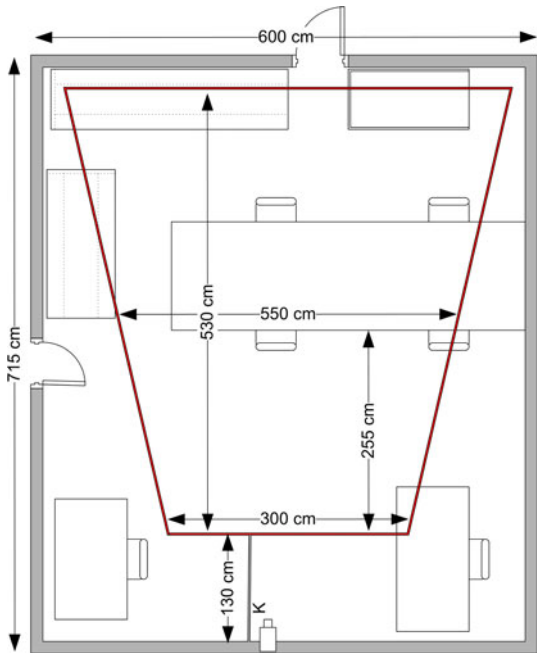

Fig. 6 Room plan showing the room setup for the evaluation

Fig. 7 One frame of the Kinect Sensor, illustrating the camera field of view

The falls were simulated and similar to the definition of falls by Noury et al. [25], but using an extended version of scenarios, depicted in Table 1. The additionally added scenarios are "sitting down on a chair and fall while getting up", "to lie down to a bed and fall out of the bed" and "fall into camera direction". These scenarios are added to enhance the quality of evaluation. Furthermore, two scenarios were taken out from the original definition of Noury at al. since we do not agree with the uniqueness of the outcome. The modification results in 18 different sequences, containing ten falls and eight no-falls. These scenarios were simulated by two subjects, simulating each scenario

Table 1 Definition of scenarios similar to Noury et al. [25]

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

Fig. 8 RGB Video frames of a scenario containing a simulated fall

twice. This results in an overall set of 72 videos, containing 40 falls and 32 no-falls.

Figure 8 shows five video frames taken out of a test sequence, showing a simulated fall according to the scenario "fall backward, ending lying on the ground". To prevent injuries, falls are simulated using a mat. The corresponding depth frames with skeleton points of the head, shoulder, spine, hip and the average of both knees are shown in Fig. 9.

The video frames of an activity of daily living are shown in Fig. 10. This figure shows a person picking something up from the floor. Due to the orientation of the body, this scenario provoke FP, as the body may be parallel to the ground floor. The corresponding depth data of this sequence are depicted in Fig. 11.

Results of evaluating our proposed approaches using the Kinect as 3D sensor on 72 sequences are depicted in Table 2. The absolute values for TP, TN, FP and FN are shown, and an comparison between our approach using image coordinates and world coordinates is given. Using the above specified measures, a comparison of the results of our two approaches according to these measures is shown in Table 3. This table shows that the results of both approaches are similar at first glance. Analyzing the evaluation results in detail shows that at least in five sequences errors occur due to not correctly ending the tracking process when the person leaves the frame. Thus, improving the tracking of the skeleton will improve the obtained results. Assuming that tracking works correctly (i.e., filtering out the last frames of the videos where tracking problems occurred) lead to the results depicted in Table 4. Analog to Table 2, the absolute values for TP, TN, FP and FN are shown. The corresponding measures are shown in Table 5. After the elimination of the tracking errors, a comparison of our approaches shows that the use of the Kinect as 3D sensor with world coordinates clearly outperforms our approach using the Kinect with image coordinates.

4.2 Discussion and comparison of technologies

The evaluation of our proposed algorithms using the Kinect as 3D sensor (together with image and world coordinates) is compared to results of an audio-based algorithm [28] and

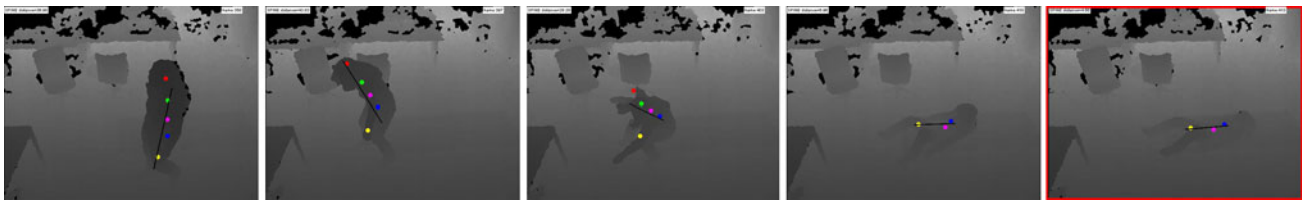

Fig. 9 Depth frames of a scenario containing a simulated fall

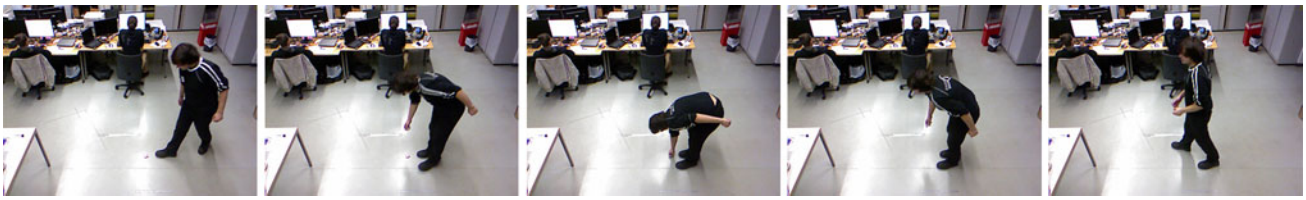

Fig. 10 RGB Video frames of a scenario where the subject picks something up from the ground

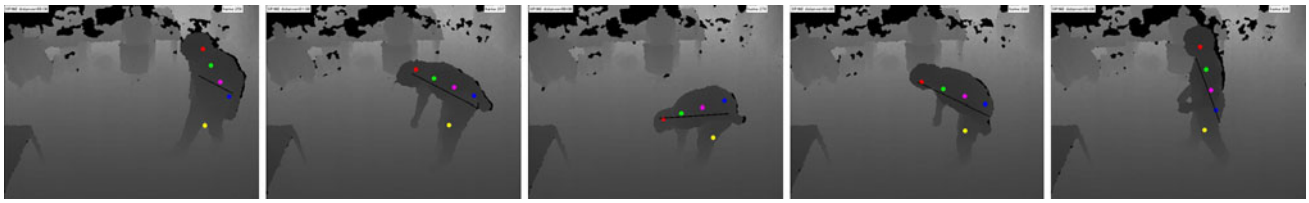

Fig. 11 Depth frames of a scenario where the subject picks something up from the ground

Table 2 Results for evaluating our fall detection approaches

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

Table 3 Measures for evaluating our fall detection approaches

	3D sensor using image coordinates	3D sensor using world coordinates
Recall	0.775	0.925
Precision	0.969	0.881
F-score	0.861	0.902
True negative rate	0.969	0.857
Accuracy	0.861	0.893

Table 4 Results obtained after eliminating tracking errors

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

Table 5 Measures obtained after eliminating tracking errors

	3D sensor using image coordinates	3D sensor using world coordinates
Recall	0.775	0.925
Precision	1	1
F-score	0.873	0.961
True negative rate	1	1
Accuracy	0.875	0.958

the fall detection algorithm using a 2D sensor and a statistical model [40].

The algorithm of Popescu et al. [28] is evaluated on a database containing ten falls and approximately 2 h of "normal" activities. The database for evaluating the audio-based approach [28] results in eight TP, two FP and two FN.

Table 6 Comparison of different technologies for fall detection

	Audio-based approach	2D sensor based approach	3D sensor using image coordinates	3D sensor using world coordinates
Recall	0.8		0.775	0.925
Precision	0.8	0.77	1	1
F-score			0.873	0.961
True negative rate			1	1
Accuracy			0.875	0.958

The database used by Zweng et al. [40] consists of 73 video sequences containing 49 falls and 24 video sequences with activities of daily living (e.g., sneezing, picking something up,...). They tested a single as well as a multiple camera approach under laboratory conditions, resulting in the precision specified in Table 6.

The comparison of results is shown in Table 6. This comparison shows that the audio-based and 2D sensor approaches perform similar and are outperformed by our algorithm using depth information. Our algorithm is implemented in C/C++ and is able to detect falls in real-time, that is, 30 fps on a Intel Core i7-2620M Quad Core CPU @2.7 GHz and 8 GB RAM.

Evaluation has shown that it is not feasible to detect the fall event itself. Therefore, we propose to detect situations where help is needed rather than focusing on the fall event. These situations are detected by the information that a person is in an upright position, is on the floor afterward and does not get up to an upright position within an specified amount of time. Especially for real-world applications, it does not make any difference, why the person is not able to get up from the floor. Hence, it is not of interest whether a fall occurred or whether the person intentionally lay down on the ground—if the person is not able to get up any more, help is needed in any case!

When using autonomous systems within the homes of elderly, privacy and the protection of data becomes an essential aspect to be considered. To ensure the dignity of elderly, the anonymization of data is required. Therefore, the video stream from the Kinect sensor is not analyzed at any time, and only depth data are processed. Hence, people and objects cannot be identified any more. Figure 12 depicts the depth images and thus automatically anonymized snapshots of the 3D Kinect sensor containing major body joints and the major orientation of the person. In contrast, Fig. 13 shows anonymized snapshots of the camera based on Zambanini et al. [37]. They anonymize the camera pictures by applying edge detection algorithms to ensure the dignity of elderly.

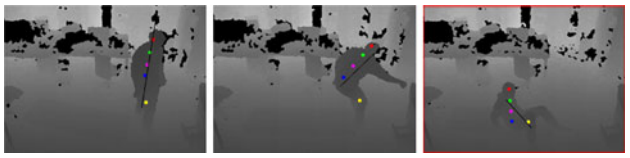

Fig. 12 Anonymized snapshots using the 3D sensor (Kinect)

Fig. 13 Anonymized snapshots using the system proposed by Zambanini et al. [37]

5 Conclusion

This article shows an overview of different non-invasive fall detection approaches and introduces a fall detection approach using Microsoft's Kinect as a 3D sensor. A comparison of methods for fall detection using 2D sensors, microphones and Kinect as a 3D sensor is shown. Our proposed algorithms for the Kinect make use of the orientation of the body and the height information of the spine, using either image or world coordinates. While having two parameters (threshold for similarity to the ground and the height), evaluation has shown that our algorithms outperform other computer vision algorithms as well as algorithms based on audio information. Furthermore, it is shown that our algorithm using world coordinates outperformed our approach using image coordinates. Future work will deal with the tracking problems and a combination of different approaches discussed in the evaluation to enhance the robustness of our approach.

References

1. Aghajan H, Wu C, Kleihorst R (2008) Distributed vision networks for human pose analysis. In: Signal processing techniques for knowledge extraction and information fusion, pp 181–200
2. Aldrich F (2003) Smart homes: past, present and future. In: Harper R (eds) Inside the smart home, Springer, London, pp 17–39
3. Anderson D, Keller J, Skubic M, Chen X, He Z (2006) Recognizing falls from silhouettes. In: 28th Annual international conference of the IEEE engineering in medicine and biology society (EMBS '06), New York, pp 6388–6391
4. Anderson D, Luke R, Keller J, Skubic M, Rantz M, Aud M (2009) Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Comput Vis Image Underst 113:80–89

5. Belbachir AN, Lunden T, Hanák P, Markus F, Böttcher M, Mannersola T (2010) Biologically-inspired stereo vision for elderly safety at home. e & i Elektrotechnik Informationstechnik 127(7):216–222
6. Cai Y (2010) Mobile intelligence. J Univers Comput Sci 16(12):1650–1665
7. Diraco G, Leone A, Siciliano P (2010) An active vision system for fall detection and posture recognition in elderly healthcare. In: Design, automation test in Europe conference exhibition (DATE), Dresden, pp 1536–1541
8. Doukas C, Maglogiannis I (2008) Advanced patient or elder fall detection based on movement and sound data. In: Second international conference on pervasive computing technologies for healthcare, pp 103–107. IEEE, Tampere
9. Doulamis AD, Doulamis ND, Kollias SD (2000) A fuzzy video content representation for video summarization and content-based retrieval. Signal Process 80(6):1049–1067
10. Erik Stone MS (2011) Evaluation of an inexpensive depth camera for in-home gait assessment. J Ambient Intell Smart Environ 3(4):349–361
11. Fisk MJ (2001) The implication of smart home technologies. In: Inclusive housing in an aging society: innovative approaches. The Policy Press, Bristol, pp 101–124
12. Gales M, Young S (2008) The application of hidden Markov models in speech recognition. Found Trends Signal Process 1(3):195–304
13. García-Vázquez J, Rodríguez M, Andrade A, Bravo J (2011) Supporting the strategies to improve elders medication compliance by providing ambient aids. Pers Ubiquit Comput 15(4):389–397
14. Grabner H, Leistner C, Bischof H (2008) Time dependent on-line boosting for robust background modeling. In: Proceedings of the international conference on computer vision, imaging and computer graphics theory and applications, pp 612–618
15. Jansen B, Temmermans F, Deklerck R (2007) 3D human pose recognition for home monitoring of elderly. In: Conference of the IEEE on engineering in medicine and biology society, Lyon, pp 4049–4051
16. Jara A, Zamora M, Skarmeta A (2011) An internet of things-based personal device for diabetes therapy management in ambient assisted living (AAL). Pers Ubiquit Comput 15(4):431–440
17. Kampel M, Wildenauer H, Blauensteiner P, Hanbury A (2007) Improved motion segmentation based on shadow detection. Electronic Lett Comput Vis Image Anal 6(3)
18. Labayrade R, Aubert D, Tarel JP (2002) Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In: IEEE Intell Vehicle Symp 2:646–651
19. Laptev I, Caputo B, Schüldt C, Lindeberg T (2007) Local velocity-adapted motion events for spatio-temporal recognition. Comput Vis Image Underst 108(3):207–229
20. Leikas J, Salo J, Poramo R (1998) Security alarm system supports independent living of demented persons. Gerontechnology: a sustainable investment in the future. Technol Inf 48:402–405
21. Lindemann U, Hock A, Stuber M, Keck W, Becker C (2005) Evaluation of a fall detector based on accelerometers: a pilot study. Med Biol Eng Comput 43:548–551
22. Litvak D, Zigel Y, Gannot I (2008) Fall detection of elderly through floor vibrations and sound. In: 30th Annual international conference of the IEEE engineering in medicine and biology society, (EMBS '08), vol 2008, pp 4632–4635
23. Lubinski R (1991) Dementia and communication. B.C. Decker, Inc., Hamilton
24. Mihailidis A, Carmichael B, Boger J (2002) The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. IEEE Trans Inf Technol Biomed 8(3):238–247
25. Noury N, Rumeau P, Bourke A, OLaighin G, Lundy J (2008) A proposal for the classification and evaluation of fall detectors. Biomed Eng Res (IRBM) 29(6):340–349
26. Oggier T, Lehmann M, Kaufmann R, Schweizer M, Richter M, Metzler P, Lang G, Lustenberger F, Blanc N (2004) An all-solid-state optical range camera for 3D real-time imaging with sub-centimeter depth resolution (SwissRanger). In: Proceedings of SPIE, vol 5249, SPIE, pp 534–545
27. Popescu M, Li Y, Skubic M, Rantz M (2008) An acoustic fall detector system that uses sound height information to reduce the false alarm rate. In: 30th Annual international conference of the IEEE engineering in medicine and biology society (EMBS '08), pp 4628–4631
28. Popescu M, Mahnot A (2009) Acoustic fall detection using one-class classifiers. In: Symposium of the association for the advancement of artificial intelligence, AAAI 2009, pp 3505–3508
29. Rougier C, Auvinet E, Rousseau J, Mignotte M, Meunier J (2011) Fall detection from depth map video sequences. In: Abdulrazak B, Giroux S, Bouchard B, Pigot H, Mokhtari M (eds) Toward useful services for elderly and people with disabilities, *Lecture Notes in Computer Science*, vol 6719. Springer, Berlin / Heidelberg, Montreal, pp 121–128
30. Rougier C, Meunier J, St-Arnaud A, Rousseau J (2006) Monocular 3d head tracking to detect falls of elderly people. In: 28th Annual international conference of the IEEE on engineering in medicine and biology society (EMBS '06), New York, pp 6384 – 6387
31. Rougier C, Meunier J, St-Arnaud A, Rousseau J (2007) Fall detection from human shape and motion history using video surveillance. In: 21st International conference on advanced information networking and applications workshops (AINAW '07), vol 2, Niagara Falls, pp 875–880
32. Saunders J (1996) Real time discrimination of broadcast speech/music. In: Proceedings of the international conference on acoustics, speech, signal processing (ICASSP), pp 993–996
33. Scheirer EMS (1997) Construction and evaluation of a robust multifeature speech/music discriminator. In: Proceedings of the international conference on acoustics, speech, signal processing (ICASSP), pp 1331–1334
34. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: IEEE conference on computer vision and pattern recognition (CVPR), 2011, pp 1297–1304
35. Van Rijsbergen CJ (1979) Information retrieval. Butterworths, London
36. Willems J, Debard G, Bonroy B, Vanrumste BTG (2009) How to detect human fall in video? An overview. In: Positioning and context-aware international conference (POCA)
37. Zambanini S, Machajdik J, Kampel M (2010) Early versus late fusion in a multiple camera network for fall detection. In: 34th Annual workshop of the Austrian Association f. Pattern recognition (ÖAGM 2010), vol 819862, Zwettl, Austria, pp 15–22
38. Zhang T, Kuo CC (1999) Hierarchical classification of audio data for archiving and retrieving. In: Proceedings of the international conference on acoustics, speech, signal processing (ICASSP), vol 6, pp 3001–3004
39. Zhao J, Katupitiya J, Ward J (2007) Global correlation based ground plane estimation using V-disparity image. In: IEEE international conference on robotics and automation, pp 529–534
40. Zweng A, Zambanini S, Kampel M (2010) Introducing a statistical behavior model into camera-based fall detection. In: Proceedings of the 6th international symposium on visual computing (ISCV), vol 6453, Las Vegas, Nevada, pp 163–172