Computer Vision for Active and Assisted Living

Rainer Planinc, Alexandros Andre Chaaraoui, Martin Kampel and Francisco Florez-Revuelta

1 Introduction

The field of computer vision has been growing steadily and attracting the interest of both researchers and the industry. Especially in the last decade, enormous advances have been made with regard to automated and reliable recognition of image or video content, such as face, object and motion recognition [1–3] and gesture and activity recognition [4, 5]. Additionally, recent advances in 3D scene acquisition, such as the Microsoft Kinect depth sensor, represent a huge leap forward, enabling 3D modelling and body pose estimation in real time with low cost and mostly simple setup solutions. Such advances are very relevant to the field of active and assisted living (AAL). Traditionally binary sensors have been employed to provide the infrastructure of a smart home upon which services can then be provided to assist people and provide comfort, safety and eHealth services, among others. However, binary sensors reach their limits when complex scenarios have to be taken into account that require a broader knowledge of what is happening and what a person is doing. One or more visual sensors can provide very detailed information about the state of the environment and its inhabitants when combined with the aforementioned pattern

Rainer Planinc
Vienna University of Technology, Institute for Computer Aided Automation, Vienna, Austria e-mail: rainer.planinc@tuwien.ac.at

Alexandros Andre Chaaraoui
Google, Inc. e-mail: alexandrosc@google.com

Martin Kampel
Vienna University of Technology, Institute for Computer Aided Automation, Vienna, Austria e-mail: martin.kampel@tuwien.ac.at

Francisco Florez-Revuelta
Kingston University, Faculty of Science, Engineering and Computing, Kingston upon Thames, United Kingdom e-mail: F.Florez@kingston.ac.uk

recognition and machine learning techniques. For this reason, computer vision is becoming more and more popular for assisted living solutions.

In this chapter, we are going to review how cameras are employed in AAL and what the current state of the art is related to applications and recognition techniques. We will distinguish between traditional RGB cameras and depth sensors (or multimodal approaches, where both techniques are combined), whose recent great impact in the field deserves an individual overview. With the goal of introducing professionals of other AAL fields to computer vision in general and specifically to its application to assisted living, we will review the image processing pipeline, from different camera types that can be installed in care centres and people's homes to recent advances in image and video feature extraction and classification. For depth sensors, specific applications, feature estimation techniques and the most successful data representations are reviewed to detail how these differentiate from the traditional RGB approaches.

The remainder of this chapter is structured as follows: Section 2 reviews the most common applications of computer vision in AAL and related projects and works, and provides then an overview of the different stages of a traditional image processing pipeline including insights about its application to AAL. Section 3 introduces depth sensors and their main advantages that have made them so popular, it then continues with the data representations that are used among the state of the art for skeletal human motion analysis. Finally, Section 4 confronts the observed progress and advantages with the concerns of continuous monitoring and private spaces and related limitations, and concludes this chapter.

2 Using RGB cameras

Even though the idea of using cameras to monitor older or impaired people easily raises privacy concerns, computer vision has been considered widely for AAL [6, 7]. This is due to the multiple types of AAL scenarios in which the use of cameras would still be acceptable, such as in public facilities, *i.e.* nursing centres and hospitals, and during specific activities or events, such as tele-rehabilitation or safety assessment. Since image and video can provide very rich data about a person's activity, research has also been carried out to enhance monitoring systems with security [8] and privacy protection techniques [9, 10]. In this sense, cameras can provide rich sensor data for human monitoring, not only complementing systems with networks of binary sensors, but potentially replacing them in a near future.

2.1 Applications

In this section, we will briefly go through the main applications that video cameras have enabled in AAL scenarios. These applications go from event detection to person-environment interaction, support to people with cognitive impairment, affective computing and assistive robots. However, the following applications stand out among the state of the art.

Human behaviour analysis From basic motion tracking [11], through human action and activity recognition [12–14] to long-term behaviour analysis [7], these fields have been studied extensively for AAL applications. Greatest interest receives the potential recognition of activities of daily living (ADLs), which can lead to monitor habits and routines related to a person's health, as well as to abnormal behaviour detection, which is of special interest for early detection of mental impairment. In this sense, performing an ADL, such as a kitchen task, can serve as a functional measure of cognitive health [15]. Recently, there is also an increasing interest in the use of wearable cameras for recognising ADLs [16–18].

Fall detection Over a decade of work can be found on using RGB cameras for fall detection. Early work from Nait-Charif and McKenna relies on tracking and ellipse modelling techniques to detect falls as unusual activities [19]. A very similar work can be found in [20], and multi-camera networks have been employed in [21] and [22] among multiple others [23].

Tele-rehabilitation Therapies based on rehabilitation exercises or gaming can benefit from visual monitoring allowing to apply semi-automated evaluations of the performed tasks. For instance, exergames have been developed for stroke rehabilitation [24] or to rehabilitate dynamic postural control for people with Parkinson's disease [25].

Gait analysis The uniqueness of human gait has traditionally led to its application to human identification [26]. Nonetheless, human gait is also a valuable indicator of the mobility and health of a person. Interestingly, due to the complex mental coordination process involved, physical frailty can also be associated to an increased risk of cognitive impairment [27]. Automatic visual gait analysis has also been employed for fall prevention by assessing the risk [28, 29].

Physiological monitoring Image processing techniques have recently been developed to measure some physiological variables without direct contact with the user, *e.g.* heart rate [30] and respiratory motion [31]. Monitoring over time the semeiotic face signs have also been uses to assess cardio-metabolic risks [32].

2.2 Image processing stages

In order to take advantage of image and video based data from one or multiple RGB cameras, the image streams have to be analysed. For this purpose, different pattern recognition and machine learning techniques are commonly applied depending on the targeted application and the level of temporal and semantic complexity of the event that has to be detected. For this purpose, the video stream is processed through a pipeline of processing stages that allow to apply computer vision techniques from person identification to activity recognition. In this chapter, these have been divided based on the objective of each processing stage, namely image acquisition, image pre-processing, and feature extraction and classification.

In the following, each of these processing stages will be described focusing on how these are applied to AAL scenarios and the related works that can be found among the state of the art.

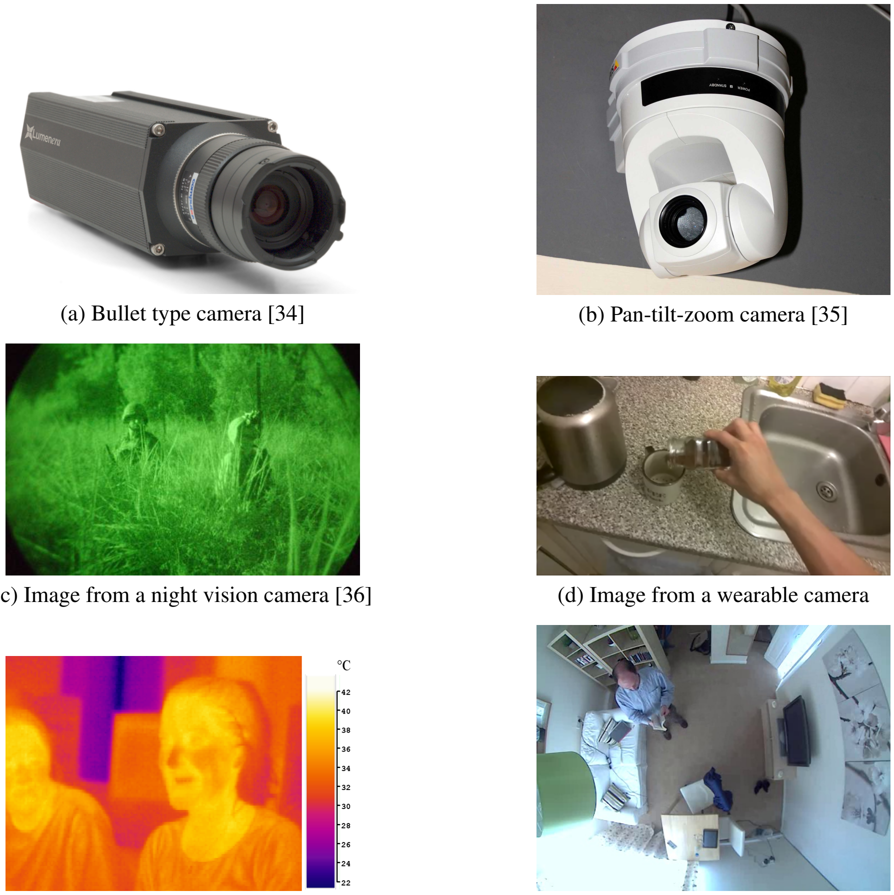
2.2.1 Image acquisition

Nowadays, a variety of video cameras can be found for monitoring and surveillance purposes. Cameras can be divided by their intended place of installation, such as outdoors or indoors, their mechanical capacities, such as bullet type or pan-tilt-zoom cameras, or their in-built features, such as motion detection or night vision, among others. Specifically, in AAL scenarios, mostly traditional indoor bullet type cameras have been used, along with omnidirectional cameras. The latter have the advantage of having an increased field of view by means of a fish-eye lens. This allows to cover, for instance, a complete room as shown in Figure 1 from a centric view point of the ceiling, if its height and the camera's field of view are sufficient. Omnidirectional cameras have been proposed for example in [33] for a home-care robotic systems. However, capturing naturally-occurring activities is challenging due to the inherently-limited field of view of fixed cameras, the occlusions created by a cluttered environment, and the difficulty of keeping all relevant body parts visible, mainly hands, as the torso and head may create occlusions. This is the reason why wearable cameras, such as GoPro® or Google Glass™, are beginning to be employed in assisted living applications.

Besides RGB cameras, several other image capturing technologies have been employed for assisted living scenarios. Depth cameras, based either on time of flight (TOF) or structured light have been very popular recently. Computer vision methods can take advantage of depth data enabling, for example, 3D scene understanding and markerless human body pose estimation. This has led to a significant amount of research effort and results in the state of the art. For this reason, depth sensors are considered separately in Section 3. Thermal cameras, which acquire the infrared radiation of the scene, also facilitate person segmentation and pose estimation.

Although for video surveillance the traditional CCTV is still employed in most cases, in AAL scenarios these have been replaced for internet protocol (IP) cameras, where the image transition occurs over local area networks, which are typically used in smart homes also for other purposes, such as binary sensor networks and internet-based services. A central point of processing, either inside the building or remotely, receives the camera streams for their storage and analysis. Additionally, cameras can provide features such as on-camera recording, and some basic image analysis, as the aforementioned motion detection, which can trigger the recording if desired.

Using networks of multiple cameras leads to additional constraints, since multi-view calibration and multi-camera geometry have to be taken into account. The work from Aghajan & Cavallaro [38] analyses these topics, along with distributed camera networks, multi-camera topologies and optimal camera placement. How-



(a) Bullet type camera [34]

(b) Pan-tilt-zoom camera [35]

(c) Image from a night vision camera [36]

(d) Image from a wearable camera

(e) Image from a thermal camera [37]

(f) Image from an omnidirectional camera

Fig. 1 These figures show respectively different types of cameras and images.

ever, in AAL scenarios, like smart homes or care centres, other environmental sensors are also employed, which typically rely on a middleware, *i.e.* the interplatform service-oriented software that integrates sensor and actuator protocols of different manufacturers [39]. As a consequence, the system architecture will also constrain how a multi-camera network can be deployed and where the image streams can be analysed.

In [40], several recent assistive smart home projects are reviewed and it can be observed that RGB cameras are used widely in AAL for applications as activity recognition and fall detection.

2.2.2 Image pre-processing

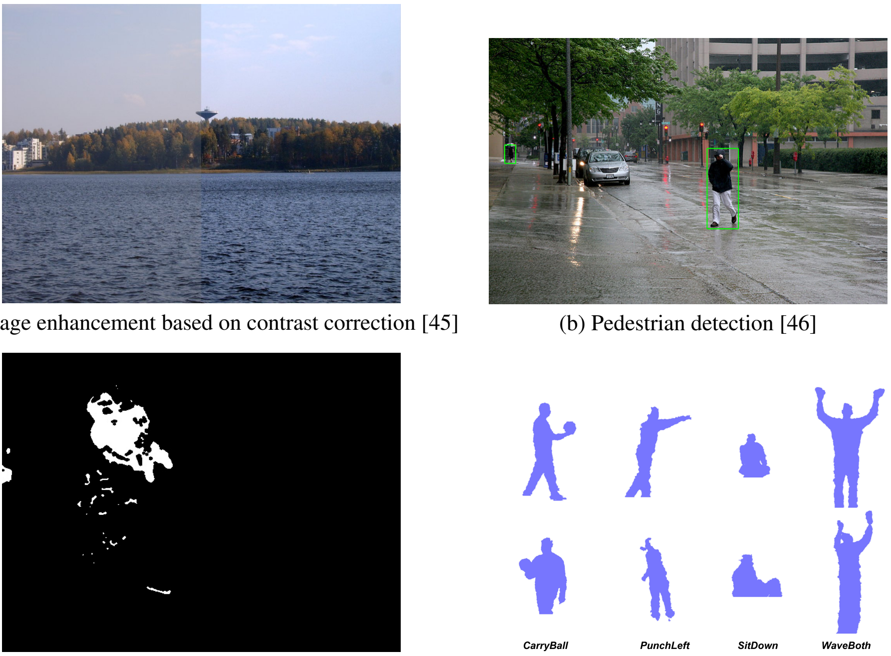
Since for the applications mentioned in Section 2.1 the main interest is focused on the recorded people, the part of the image that contains the human silhouette, *i.e.* the region of interest (ROI), has to be extracted. Blob detection techniques make it possible to identify these ROI based on colour, brightness, shape or texture. In this stage, sensor-specific image pre-processing methods can be applied to filter noise, increase contrast or enhance colours. In order to separate the ROI from the rest of the image, most frequently motion segmentation techniques are applied, which rely on the fact that the people in the image are in motion whereas the background is rather static. Image segmentation techniques as codebook representation [41], Gaussian mixture learning [42] and *GrabCut* [43] are frequently used among the state of the art. However, alternative approaches can be found too. For example, in [44] silhouettes are obtained based on contour saliency combining both colour and thermal images.

After the foreground pixels have been identified, blob detection techniques group neighbouring pixels based on different criteria such as connectivity, colour, shape and width and height ratios, and identify the image regions that should be considered as a single object (namely a *blob*). Once a region of interest is obtained, the containing pixels have to be described and normalised in a suitable manner in order to apply pattern recognition techniques or learn and classify them with machine learning algorithms. Additionally, a dimensionality reduction is usually desirable, since the increasing spatial and temporal resolution of video data would otherwise make real-time methods infeasible. Figure 2 shows examples of different pre-processing techniques that are typically performed before the feature extraction and recognition stages can be initiated.

For example, in [47] a view-invariant fall detection system is developed relying on view-invariant human pose estimation. The video stream provided by a monocular camera is first downsampled (or upsampled if necessary) to 15 fps to ensure stable real-time execution and then converted to 8-bit greyscale images. As part of the pre-processing, foreground extraction is performed based on the work of the W^4 system [48] using a non-parametric background model that learns greyscale levels and variations on a pixel-by-pixel basis assuming an empty background. The foreground is then detected based on the deviation of the learned model. An erosion filter is employed to delete noise and blobs are obtained based on connectivity and a minimum size. Additionally, a temporal segmentation based on motion energy boundaries is performed to segment the continuous stream in individual sequences that can be analysed in isolation. This processing then allows to continue with pose modelling and recognition.

2.2.3 Feature extraction and recognition

Once the necessary pre-processing stages have been executed, image representations based either on the whole image or on the detected regions of interest are



(a) Image enhancement based on contrast correction [45]

(b) Pedestrian detection [46]

(c) Person segmentation obtained from Figure 1(f)

(d) Human silhouettes corresponding to different activities.

Fig. 2 Result examples of image pre-processing methods are shown respectively for noise reduction, blob detection, background segmentation and silhouette extraction techniques.

generated in order to obtain the characteristic information that defines the event to be detected. These are the so-called visual features. Image and video features can be distinguished as dense features, which represent the data with a global (also known as *holistic*) descriptor, or sparse features, which use a set of local representations of the region of interest or even of the whole image.

A very popular feature for human detection are histograms of oriented gradients (HOG). Dalal and Triggs [49] proposed to evaluate normalised local histograms of image gradient orientations in a dense grid on a gamma and colour normalized image. In [50], the authors combined this approach with a similar feature for oriented optical flow (histogram of oriented flow –HOF–) in order to capture both shape and motion, leading to the state of the art standard for human detection. In [51], this method is used in addition to holistic features extracted from raw trajectory cues for recognition of ADL of healthy subjects and people with dementia using the URADL dataset [52].

Another well-established image representation are motion history images (MHI) [53], where in this case both shape and motion are captured in a single bidimensional feature. First, background segmentation is applied to a sequence of images, which are typically downsampled in both size and frame rate. The segmented foreground of each image of the sequence is then combined by assigning to each coordinate of the feature vector a value that represents the recentness of the motion in that pixel. These values can then be mapped to greyscale intensities, where pixels with more recent motion appear brighter. This allows to encode the temporal evolution of the motion as well as its spatial location. This feature is used for example in [20] to detect different types of falls and activities.

SIFT [54], SURF [55], and other local descriptors are widely used to detect and characterise objects and persons. SIFT is a gradient-based algorithm that detects and describes local features in images; it is invariant to image translation, scaling, and rotation. This are usually clustered into different classes, named visual words, building a codebook. Then, an image can be characterised with a bag of words [56] (BoW), a vector counting the occurrence or frequency of each visual word.

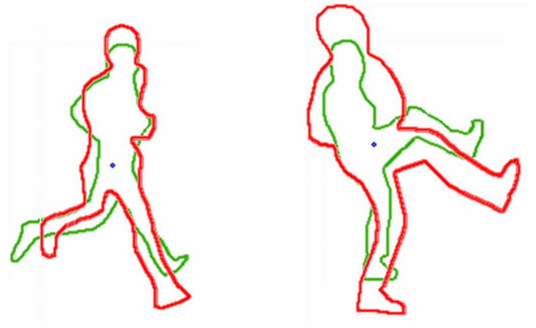


Fig. 3 Centred silhouettes from two viewpoints (respectively in red and green) used for contour-based multi-view pose representation in [57].

Since in the aforementioned pre-processing stage can include silhouette extraction, it is also common to build a holistic descriptor of the individual's silhouette. Seeing that the shape of the silhouette is defined by its boundary, contour representations can lead to very summarised and descriptive features. In [58], such a representation is proposed for the recognition of human actions. The feature is built using the distances between the contour points and the centroid of the silhouette in order to obtain location invariance. The vector of distances is then downsampled to a fixed size and normalized to unit sum to obtain also scale invariance. This feature has been used successfully in [57] combining also multiple view points by means of feature concatenation (see Figure 3). It has been further improved in [59], where the silhouette is divided into radial sectors, and the statistical range of distances to the contour is used as characteristic value, leading to further dimensionality reduction and reduced noise sensitivity. Finally, the work is applied to a visual monitoring system that enables multi-view recognition of human actions to provide care and safety services, such as detection of home accidents and telerehabilitation [10].

Sparse features, such as key points have also been used extensively, as in [52], where a proposal is made based on velocity histories of tracked key points. The obtained key points are tracked with a KLT tracker and their velocity histories are computed. The features are also augmented with additional data, including the absolute initial and final positions of the key points and the relative positions with respect to the position of an unambiguously detected face if present. The local appearance information is encoded relying on horizontal and vertical gradients and PCA-SIFT [60], and also colour information of the same area is encoded based on PCA-reduced colour histograms. Both texture and colour information are not used directly, instead a codebook is obtained using K-means clustering and the nearest neighbour element is assigned to represent each key point.

These features are then classified with different recognition techniques based on learning and data analysis methods, such as the aforementioned BoW technique, or others. These can be classified in two categories, data-driven approaches, which learn generative or discriminative models from user data and infer the corresponding class, or knowledge-driven approaches, which take advantage of domain-specific knowledge and rules. Analysing these is out of scope in this chapter, however it is covered extensively in the present book and we refer the reader to consult the corresponding chapters.

3 Using depth sensors

Advances in the development of 3D sensors motivate their use instead of cameras or wearable sensors, since they provide advantages like privacy protection and improved robustness when it comes to behaviour modelling, gesture recognition or activity recognition. In contrast to RGB based analysis, depth based approaches do not process RGB colour images, but depth or range images measuring distances from objects to the sensor. Figure 4 shows a RGB camera image together with its corresponding depth image: depth images do not visualize the scene with colours, but the grey level indicates the distance of the objects and its surrounding to the sensor. The darker the colour, the closer the object is to the sensor. On the other hand, brighter colours are used for objects at a higher distance. Black holes within the depth image indicate reflecting or absorbing areas where no valid depth measurement is available and are caused due to the functionality of the sensor. In contrast to RGB based approaches, only the silhouette is detected and thus no conclusions whether the person is wearing clothes or the emotional state can be obtained since neither the clothes, nor the face are visible. Hence, the appearance of the person is fully protected. However, this is only the case if processing is solely based on depth images. RGB-D based approaches combine depth information together with appearance information in order to obtain and combine more details. Although these approaches reduce the privacy of elderly, colour information can be taken into consideration, thus allowing to perform a more in depth analysis.

The use of depth sensors became popular due to the introduction of the Microsoft Kinect in 2010 as an add-on for the Xbox console. It was the first low-cost 3D sensor and thus received a lot of attention from the research community [61]. The first version of the Kinect consists of a RGB camera, an infrared projector as well

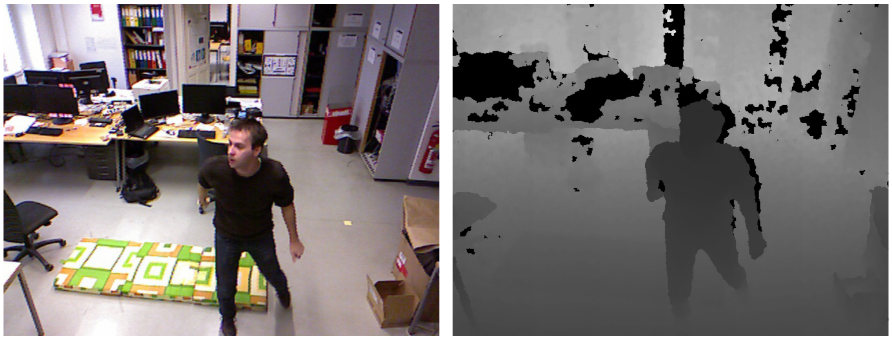


Fig. 4 RGB camera image and its corresponding depth image.

as an infrared camera. The functionality of the Kinect is based on structured light imaging, where the projector emits a pre-defined infrared light pattern to the scene [62, 63]. Due to the spatial arrangement of the pattern and its varying sizes, as well as distortions depending on the distance to the camera, the depth camera captures the light pattern and an on-board chip calculates a depth map. In contrast to the first versions of the Kinect, the functionality of the Kinect for Windows v2 is not based on structured light, but on Time-of-Flight (ToF) in order to achieve more accurate results[1].

The main advantages of depth based sensors, especially within the context of AAL, can be summarized as follows:

- *No additional light source needed:* due to the use of infrared light, sensors also work during the night (*e.g.*, when falls of elderly people occur).
- *Sensor is robust to changing lighting conditions:* switching the lights on and off does not affect the results of the depth images. However, direct sunlight interferes with the projected infrared pattern and thus, no depth value can be calculated. This restricts the use of the sensors to indoor environments only.
- *No calibrated camera setup is needed:* in contrast to the use of a calibrated multiple camera setup in order to calculate a 3D reconstruction, no calibration is needed.
- *Standard algorithms can be applied to depth information:* standard algorithms from computer vision (*e.g.* foreground/background segmentation, tracking) can be applied to depth data directly.
- *Protection of privacy:* if only depth information is processed, privacy is protected since the appearance of the person is not recognized in depth images. However, if a combined analysis of RGB and depth images is performed, privacy is not protected.

Typical applications of RGB-D approaches within the context of AAL are:

- Fall detection

[1] http://blogs.microsoft.com/blog/2013/10/02/collaboration-expertise-produce-enhanced-sensing-in-xbox-one/, accessed 09-September-2015

- Rehabilitation
- Serious Gaming
- Pose analysis
- Gesture based interfaces
- Robotics

Behavior modeling or activity recognition of humans depends on the correct detection of humans and their pose in the scene. Person detection and tracking is a complex problem due to the non-rigid nature of persons, complex motion, and occlusions, among others [64]. Kinect and other depth sensors allow for a partial reconstruction of the scene geometry, which facilitates the problem. This section lists a selection of recent methods designed for or applicable to Kinect sensors, categorized based on the data representation they operate on (Skeletal, depth maps, point clouds, or plan-view maps).

3.1 Skeletal

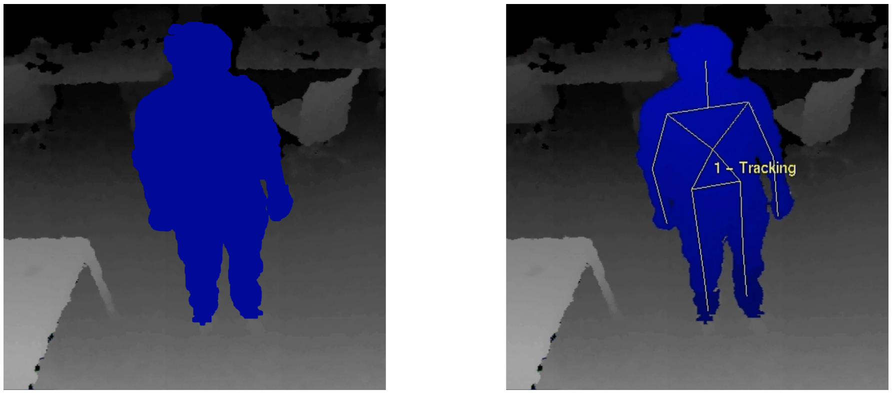


Fig. 5 Skeleton tracking.

Human pose estimation is proposed by Shotton et al. [65], allowing to extract body parts and skeleton joints based on depth images. Figure 5 depicts the pipeline for the skeleton joint calculation [66]: foreground detection is performed at the beginning in order to separate humans from the background. The use of a randomized forest allows to label different body parts in the depth image. Clustering of pixels hypothesize body joint positions, which are refined using a skeleton model exploiting temporal and kinematic constraints.

The skeleton tracking algorithm is optimized for people facing the sensor, either standing or sitting in front of the sensor. Hence, it can fail, if the person is not directly facing the sensor. The number and type of detected joints depends on the

SDK to be used - two SDK are mainly used together with the Kinect [63]: the official SDK provided by Microsoft, being able to track 20 respectively 25 skeleton joints (Kinect v1 and Kinect v2) and the OpenNI [2] provided by PrimeSense, allowing to track 15 different skeleton joints [63]. PrimeSense developed the first sensor of the Kinect for Microsoft and introduced, in cooperation with Asus, their own 3D sensor [3]: Asus Xtion pro, offering almost the same hardware specification as the Microsoft Kinect, but built in a smaller case. In contrast to the Microsoft Kinect, the Asus Xtion pro does not contain a RGB camera, but only a depth sensor and thus allows to fully respect the privacy within the context of AAL, since it is technically not possible to obtain a RGB image from this sensor.

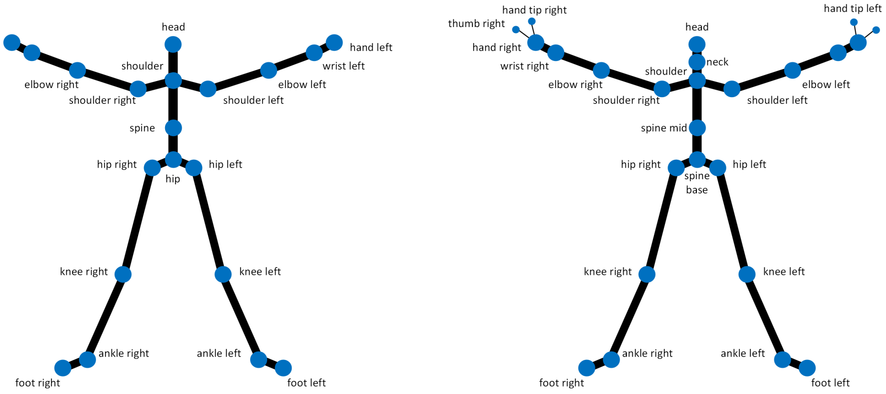


Fig. 6 Skeleton joints of Kinect v1 (left) and Kinect v2 (right).

Figure 6 depicts an overview of the estimated skeleton joints for both versions of the Kinect. In the first version, 20 skeleton joints are estimated whereas in the second version of the Kinect, 25 skeleton joints are estimated. Five skeleton joints were added and hands are modeled in more detail in order to allow a more accurate gesture control.

3.2 Depth maps

Methods operating on depth maps frequently utilize histogram-based features and supervised learning for person detection [67–69]. Conceptually, these methods are similar to seminal work by Dalal and Triggs [49] that introduced histograms of ori-

[2] http://www.openni.org, accessed 10-April-2014

[3] https://www.asus.com/de/Multimedia/Xtion_PRO/, accessed 04-August-2015

[3] Image Source: https://msdn.microsoft.com/en-us/library/jj131025.aspx, https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx; accessed 03-August-2015

ented gradients as powerful features for person detection in greyscale images. These features model the local appearance of objects by means of gradient distribution and, when consolidated to larger blocks, constitute powerful feature descriptors. These descriptors are then classified as (not) representing a person using a support vector machine. Spinello and Arras [67] as well as Wu et al. [69] proposed similar descriptors for depth maps, which model the local distribution depth gradient orientations.

3.3 Point Clouds

Another approach is to first reproject depth map pixels to world coordinates and to operate on the resulting point cloud. Kelly et al. [70] follow this approach, clustering the point cloud using an iterative top-down approach based on 3D proximity tests. Thresholds are computed dynamically from the observed maximum height and the golden ratio, which allows for estimating the proportion of persons based on their height. Subsequently the obtained clusters are analysed with respect to under- and over-segmentation via ellipse fitting on shoulder and head height, respectively. This method is able to cope with significant occlusions but is prohibitively slow (less than 1fps including tracking).

Hegger et al. [71] compute surface normals from subsampled point clouds and cluster points using an efficient top-down method that results in small clusters of adjacent points.

3.4 Plan-View Maps

A reason why methods operating on point clouds are slow is the large number of points (up to 307200) and the fact that clustering is a complex task. Hegger et al. [71] alleviate this problem by subsampling the point cloud. This is accomplished by discretising the continuous space into cubic cells, which significantly reduces the number of points at the expense of resolution. A similar yet more extreme approach to data reduction is to utilize so-called plan-view maps, two-dimensional representations of the scene as viewed from the top and under orthographic projection [72].

3.5 Accuracy

The resolution of a measuring device describes the smallest details it is able to resolve. With regard to depth sensors it is desirable to distinguish between depth resolution and spatial resolution. In this text, the term depth resolution denotes the smallest difference in distance the sensor can distinguish, while spatial resolution refers to the minimum size of reliably detectable objects.

The depth resolution of the sensor restricts detectable scene changes. If the depth resolution is too low, fallen persons or parts thereof may not be distinguishable from the floor. Kinect sensors can distinguish between two object distances only if their difference is large enough. This is because the sensor can derive disparities with only limited accuracy. In fact, the sensor distinguishes between 1024 distinct disparities and thus distances [73], while the measuring range spans approximately 10m.

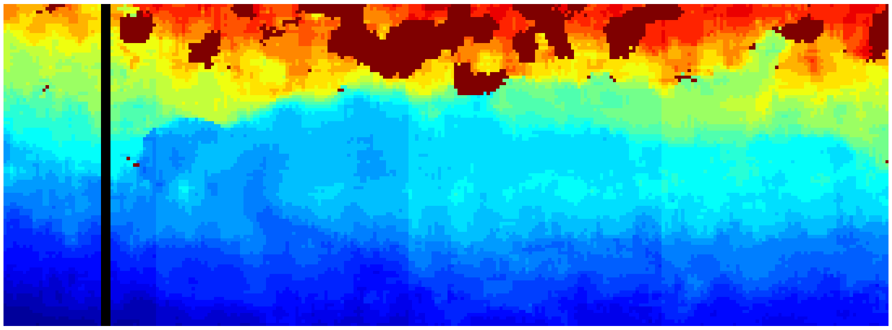


Fig. 7 Illustration of the effect of limited depth resolution on fall detection

Figure 7 depicts a person lying flat on the floor at a distance of approximately 6m from the sensor. Distances are colour-coded, blue colours represent closer distances. Due to the limited depth resolution, limbs are only partially distinguishable from the floor. This is illustrated by the fact that colours in regions of arms or legs do not always differ from those of the floor beneath them (left side of the vertical line). The minimum size objects must have in order to be reliably (i.e. continuously) detectable from a certain distance depends on the spatial resolution of the sensor. According to a datasheet by the developer of the depth sensing technology of the Kinect, the spatial resolution is 3.4mm at an object distance of 2m.

The precision describes the variability between multiple measurements of the same object under stable conditions [74]. With regard to the Kinect, the precision indicates the closeness of repeated distance measurements. This criteria is particularly important with respect to background subtraction algorithms, as their performance depends on data stability. Knowledge of the sensor precision allows for estimating the performance of such algorithms and aids in proper configuration.

Smisek et al. [61] evaluated the accuracy of the Kinect sensor (first version), by analyzing its depth resolution. The depth resolution of the Kinect is within the range between 0.65 mm at a distance of 0.5 meters and up to 685 mm at a distance of 15.7 m. These results indicate that the use of the Kinect for indoor environments is feasible, although the accuracy decreases with higher distance. Results show that the depth resolution within a range up to ten meters is below 300 mm. Moreover, the performance of the Kinect is compared to the performance of a stereo-camera (two Nikon D60 SLR) as well as a ToF (SwissRanger SR-4000) system. Smisek et al. [61]

showed that the Kinect performs similar to a stereo system with medium resolution and outperforms the ToF system in terms of accuracy and costs. Stoyanov et al. [75] evaluated the Kinect and two ToF sensors based on ground truth data obtained by a laser sensor. For short distances, the Kinect slightly outperformed both ToF sensors and performed similarly to the accuracy of the laser sensor. However, no sensor achieved the accuracy of the laser sensor on longer distances.

Ditta [76] compared the skeleton tracking of the Kinect with a marker based system from Vicon and showed, that the errors of the Kinect are approximately 5 mm within a range of 1-3 meters in comparison to the Vicon system. Galna et al. [25] evaluated the use of the Kinect for the detection of movement symptoms for people with Parkinson's disease and compared their results with a Vicon system. Normal actions (standing, walking and reaching) are combined with actions from the Unified Parkinson's disease Scale and include, amongst others, hand clasping and finger tapping. The timing of movement is measured accurately as well as extensive movements are detected accurately. Only when monitoring fine movement, the Kinect is not able to obtain accurate results and thus is outperformed by the Vicon system. Overall, Galna et al. [25] conclude that the Kinect can accurately detect most movements related to Parkinson's Disease. The accuracy of the Kinect within exergames is evaluated and compared to a Vicon system by van Diest et al. [77]. The outcome of their evaluation shows that the Kinect accurately detects movements of the trunk, but does not detect the movement of hands and feet accurately, resulting in a difference of up to 30% in comparison to the Vicon system. The reasons for the lower accuracy of the Kinect is the reduced resolution (640x480) in comparison to the Vicon system (4704x3456) and the low and irregular sampling frequency [77].

Plantard et al. [78] propose a framework to simulate 500 000 poses at a workplace in combination with different positions of the Kinect, in order to perform an automatic large scale evaluation of the accuracy of the Kinect. The results show that the accuracy depends on the specific pose as well as the position of the Kinect. Although most results are accurate (*e.g.*. error of the shoulder position is 2.5 cm), positions with partial occlusions results in the failure of the skeleton tracking algorithm.

Moreover, the performance of the skeleton tracking system during six different exercises is evaluated by Obdrzalek et al. [79]. Again, a marker based tracking system provides ground truth data and it is shown that the Kinect has a great potential. However, since exercises are performed either sitting or while touching a chair, the skeleton tracking algorithm fails when body parts are occluded or a chair is presented. Although problems with the skeleton tracker are reported, Obdrzalek et al. [79] state that within a more controlled environment, tracking results are better. They conclude that during general postures a variability of about 10 cm can be observed in comparison to the marker based tracking system. Nevertheless, these results show that the use of a Kinect is feasible, but depends on the application and context. Due to its low costs in comparison to other 3D sensors and its accuracy, the Kinect is used in computer vision for achieving different and diverse tasks: approaches using the Kinect for object tracking and recognition, human activity analysis, hand gesture analysis and 3D mapping of indoor environments are just few of them [63].

4 Conclusion

Over the course of the last years, vision-based solutions gained of interest within the field of AAL. However, since AAL solutions are used in private spaces (*e.g.*. bedrooms, bathrooms), privacy issues need to be considered. The use of cameras in private spaces without any privacy protection mechanisms yields in low acceptance by the end user. Hence, privacy needs to be protected either by design, or by explicitly taking appropriate measures in order to prevent subjects from carrying out camera-avoiding techniques to sabotage the monitoring [80]. Although using depth sensors results in 3D data where the person cannot be identified, the location of the sensor provides information about the person itself. Especially when being used in homes for older adults, the connection between the depth image (where the person cannot be identified) and the room number (where the system is placed) allows to implicitly identify the person on the depth image. Again, appropriate protection need to be considered in order to ensure that only authorized personal has access to this personal data. Moreover, when using data from different sources, the aggregation of data might result in private information, violating the privacy of the user. Especially if a person is being tracked over multiple cameras/sensors, his or her whole trajectory is available and thus information about the behaviour of the person is gathered.

Although vision-based systems need to be designed carefully in order to protect the privacy, their flexibility and adaptability is one of their biggest advantages. The application of AAL systems can be extended easily, since only software needs to be updated in order to provide additional features and no installation of additional sensors is needed. Thus allows to install vision-based systems according to the specific needs of the current context and the possibility to adopt the system to changing needs later on. Moreover, using vision-based systems results in a big amount of information to be processed, allowing to analyse complex scenarios and offering possibilities for various applications within the field of AAL.

Acknowledgements

Final version published in *Active and Assisted Living: Technologies and Applications* by Dr. Francisco Florez-Revuelta & Dr. Alexandros Andre Chaaraoui (Eds.), IET, 2016. ISBN 978-1-84919-987-2.

References

[1] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition—a review," *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 103 – 135, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314204000451

[2] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000042934.15159.49

[3] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1–2, pp. 4 – 18, 2007, special Issue on Vision for Human-Computer Interaction. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314206002293

[4] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 3, pp. 311–324, May 2007.

[5] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2–3, pp. 90 – 126, 2006, special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314206001263

[6] F. Cardinaux, D. Bhowmik, C. Abhayaratne, and M. S. Hawley, "Video based technology for ambient assisted living: A review of the literature," *J. Ambient Intell. Smart Environ.*, vol. 3, no. 3, pp. 253–269, Aug. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2010465.2010468

[7] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 873 – 10 888, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412004757

[8] T. Winkler and B. Rinner, "Security and privacy protection in visual sensor networks: A survey," *ACM Comput. Surv.*, vol. 47, no. 1, pp. 2:1–2:42, May 2014. [Online]. Available: http://doi.acm.org/10.1145/2545883

[9] A. Bamis, D. Lymberopoulos, T. Teixeira, and A. Savvides, "The behaviorscope framework for enabling ambient assisted living," *Personal and Ubiquitous Computing*, vol. 14, no. 6, pp. 473–487, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00779-010-0282-z

[10] A. A. Chaaraoui, J. R. Padilla-López, F. J. Ferrández-Pastor, M. Nieto-Hidalgo, and F. Flórez-Revuelta, "A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context," *Sensors*, vol. 14, no. 5, pp. 8895–8925, 2014.

[11] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428 – 440,

1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314298907445

[12] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231 – 268, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107731420090897X

[13] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976 – 990, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609002704

[14] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1922649.1922653

[15] B. Kröse, T. van Oosterhout, and T. van Kasteren, "Activity monitoring systems in health care," in *Computer Analysis of Human Behavior*, A. A. Salah and T. Gevers, Eds. Springer London, 2011, pp. 325–346. [Online]. Available: http://dx.doi.org/10.1007/978-0-85729-994-9_12

[16] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Mégret, J. Pinquier, R. André-Obrecht, Y. Gaëstel, and J.-F. Dartigues, "Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia," *Multimedia tools and applications*, vol. 69, no. 3, pp. 743–771, 2014.

[17] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2847–2854.

[18] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, "Recognition of activities of daily living with egocentric vision: A review," *Sensors*, vol. 16, no. 1, p. 72, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/1/72

[19] H. Nait-Charif and S. McKenna, "Activity summarisation and fall detection in a supportive home environment," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4, Aug 2004, pp. 323–326 Vol.4.

[20] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Fall detection from human shape and motion history using video surveillance," in *Advanced Information Networking and Applications Workshops, 2007, AINAW '07. 21st International Conference on*, vol. 2, May 2007, pp. 875–880.

[21] R. Cucchiara, A. Prati, and R. Vezzani, "A multi-camera vision system for fall detection and alarm generation," *Expert Systems*, vol. 24, no. 5, pp. 334–345, 2007.

[22] H. Aghajan, J. C. Augusto, C. Wu, P. McCullagh, and J.-A. Walkden, "Distributed vision-based accident management for assisted living," in *Pervasive Computing for Quality of Life Enhancement*. Springer, 2007, pp. 196–205.

[23] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, pp. 144 – 152, 2013, special issue: Behaviours in video. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231212003153

[24] D. Webster and O. Celik, “Systematic review of kinect applications in elderly care and stroke rehabilitation,” *Journal of NeuroEngineering and Rehabilitation*, vol. 11, no. 1, p. 108, 2014. [Online]. Available: http://www.jneuroengrehab.com/content/11/1/108

[25] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, “Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson’s disease,” *Gait & Posture*, vol. 39, no. 4, pp. 1062 – 1068, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966636214000241

[26] L. Wang, T. Tan, H. Ning, and W. Hu, “Silhouette analysis-based gait recognition for human identification,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1505–1518, Dec 2003.

[27] V. J. Verlinden, J. N. van der Geest, A. Hofman, and M. A. Ikram, “Cognition and gait show a distinct pattern of association in the general population,” *Alzheimer’s & Dementia*, vol. 10, no. 3, pp. 328 – 335, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1552526013001362

[28] D. Lim, C. Kim, H. Jung, D. Jung, and K. J. Chun, “Use of the microsoft kinect system to characterize balance ability during balance training,” *Clinical interventions in aging*, vol. 10, p. 1077, 2015.

[29] A. Dubois and F. Charpillet, “A gait analysis method based on a depth camera for fall prevention,” in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, Aug 2014, pp. 4515–4518.

[30] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, “Eulerian video magnification for revealing subtle changes in the world,” *ACM Trans. Graph. (Proceedings SIGGRAPH 2012)*, vol. 31, no. 4, 2012.

[31] F. Tahavori, M. Alnowami, and K. Wells, “Marker-less respiratory motion modeling using the microsoft kinect for windows,” pp. 90 360K–90 360K–10, 2014. [Online]. Available: http://dx.doi.org/10.1117/12.2043569

[32] S. Colantonio, G. Coppini, D. Germanese, D. Giorgi, M. Magrini, P. Marraccini, M. Martinelli, M. A. Morales, M. A. Pascali, G. Raccichini *et al.*, “A smart mirror to promote a healthy lifestyle,” *Biosystems Engineering*, vol. 138, pp. 33–43, 2015.

[33] M.-L. Wang, C.-C. Huang, and H.-Y. Lin, “An intelligent surveillance system based on an omnidirectional vision sensor,” in *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, June 2006, pp. 1–6.

[34] “A photo of the lumenera Li045C intelligent IP camera with on-camera video analytics, released in 2006,” By Lumihaychuk [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons, last accessed 2015-09-26.

[35] “Axis model 214 pan-tilt-zoom IP camera,” By Kkmurray (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons, last accessed 2015-09-26.

[36] "Two american soldiers pictured during the 2003 Iraq war seen through an image intensifier," By w:en:User:AlexPlank (w:en:Image:Nightvision) [Public domain], via Wikimedia Commons, last accessed 2015-09-26.

[37] "A false color image of two people taken in long-wavelength infrared (body-temperature thermal) light," By Cody.pope [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons, last accessed 2016-05-03.

[38] H. Aghajan and A. Cavallaro, *Multi-camera networks: principles and applications*. Academic press, 2009.

[39] H. Nakashima, H. Aghajan, and J. C. Augusto, *Handbook of ambient intelligence and smart environments*. Springer Science & Business Media, 2009.

[40] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 3, pp. 579–590, May 2013.

[41] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172 – 185, 2005, special Issue on Video Object Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077201405000057

[42] D.-S. Lee, "Effective gaussian mixture learning for video background subtraction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 5, pp. 827–832, May 2005.

[43] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004. [Online]. Available: http://doi.acm.org/10.1145/1015706.1015720

[44] J. Davis and V. Sharma, "Fusion-based background-subtraction using contour saliency," in *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, June 2005, pp. 11–11.

[45] "Example of photo editing: Contrast correction. left side of the image is unmodified, right side has been touched with gimp," By Toniht at en.wikipedia [Public domain], from Wikimedia Commons, last accessed 2015-09-26.

[46] "Pedestrian detection example," By Milwaukee (WIS) N 5th St "Tree, Rain, Wind" Pedestrian 1.jpg: vincent desjardins derivative work: Indif [CC BY 2.0 (http://creativecommons.org/licenses/by/2.0)], via Wikimedia Commons, last accessed 2015-09-26.

[47] Z. Htike, S. Egerton, and K. Y. Chow, "A monocular view-invariant fall detection system for the elderly in assisted home environments," in *Intelligent Environments (IE), 2011 7th International Conference on*, July 2011, pp. 40–46.

[48] I. Haritaoglu, D. Harwood, and L. Davis, "W4: real-time surveillance of people and their activities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 809–830, Aug 2000.

[49] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 886–893 vol. 1.

[50] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance,” in *Computer Vision–ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin Heidelberg, 2006, pp. 428–441.

[51] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris, “Activity detection and recognition of daily living events,” in *Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, ser. MIIRH ’13. New York, NY, USA: ACM, 2013, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/2505323.2505327

[52] R. Messing, C. Pal, and H. Kautz, “Activity recognition using the velocity histories of tracked keypoints,” in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 104–111.

[53] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[54] M. Brown and D. G. Lowe, “Invariant features from interest point groups,” in *British Machine Vision Conference (BMVC)*, 2002.

[55] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (surf),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[56] J. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008, cited By 712. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-45049084813&partnerID=40&md5=501fd1c88bae3a8f2aabe7deafe5cb72

[57] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, “Silhouette-based human action recognition using sequences of key poses,” *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799 – 1807, 2013, smart Approaches for Human Action Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513000342

[58] Y. Dedeoğlu, B. U. Töreyin, U. Güdükbay, and A. E. Çetin, “Silhouette-based method for object classification and human action recognition in video,” in *Computer Vision in Human-Computer Interaction*. Springer, pp. 64–77.

[59] A. A. Chaaraoui and F. Flórez-Revuelta, “A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views,” *International Scholarly Research Notices*, vol. 2014, 2014.

[60] Y. Ke and R. Sukthankar, “Pca-sift: a more distinctive representation for local image descriptors,” in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, June 2004, pp. II–506–II–513 Vol.2.

[61] J. Smisek, M. Jancosek, and T. Pajdla, “3D with Kinect,” in *Proceedings of the International Conference on Computer Vision Workshops (ICCV Workshops)*.

Barcelona, Spain: IEEE, Nov. 2011, pp. 1154–1160. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6130380

[62] D. Fofi, T. Sliwa, and Y. Voisin, "A comparative survey on invisible structured light," in *Electronic Imaging 2004*, J. R. Price and F. Meriaudeau, Eds. SPIE, May 2004, pp. 90–98. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=837538

[63] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: a review." *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23807480

[64] A. Yilmaz, O. Javed, and M. Shah, "Object Tracking: A Survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.

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

[66] P. Kohli and J. Shotton, *Consumer Depth Cameras for Computer Vision*, A. Fossati, J. Gall, H. Grabner, X. Ren, and K. Konolige, Eds. London: Springer London, 2013. [Online]. Available: http://www.springerlink.com/index/10.1007/978-1-4471-4640-7http://link.springer.com/10.1007/978-1-4471-4640-7

[67] L. Spinello and K. Arras, "People detection in RGB-D data," in *Proc. Int. Conf. Intelligent Robots and Systems*, 2011, pp. 3838–3843.

[68] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features," in *Proc. Asian Conf. Computer Vision*, 2011, pp. 25–38.

[69] S. Wu, S. Yu, and W. Chen, "An attempt to pedestrian detection in depth images," in *Proc. Chinese Conf. on Intelligent Visual Surveillance*, 2011, pp. 97–100.

[70] P. Kelly, N. E. O'Connor, and A. F. Smeaton, "Robust pedestrian detection and tracking in crowded scenes," *Image and Vision Computing*, vol. 27, no. 10, pp. 1445–1458, 2009.

[71] F. Hegger, N. Hochgeschwender, G. Kraetzschmar, and P. Ploeger, "People Detection in 3D Point Clouds Using Local Surface Normals," in *Robot Soccer World Cup 2012*, ser. Lecture Notes in Computer Science. Springer, 2013, vol. 7500, pp. 154–165.

[72] D. Beymer, "Person counting using stereo," in *Proc. Workshop on Human Motion*, 2000, pp. 127–133.

[73] K. Khoshelham and S. Elberink, "Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[74] B. Taylor, C. Kuyatt, and J. Lyons, *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. Diane Publishing, 1994.

[75] T. Stoyanov, A. Louloudi, H. Andreasson, and A. J. Lilienthal, “Comparative evaluation of range sensor accuracy in indoor environments,” in *Proceedings of the European Conference on Mobile Robots (ECMR)*, Örebro, Sweden, 2011, pp. 19–24. [Online]. Available: http://www.aass.oru.se/Research/Learning/publications/2011/Stoyanov_etal_2011-ECMR11-Comparative_Evaluation_of_Range_Sensor_Accuracy_in_Indoor_Environments.pdf$\delimiter”026E30F$nhttp://oru.diva-portal.org/smash/record.jsf?pid=diva2:540987

[76] T. Dutta, “Evaluation of the Kinect sensor for 3-D kinematic measurement in the workplace.” *Applied Ergonomics*, vol. 43, no. 4, pp. 645–649, Jul. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003687011001529

[77] M. van Diest, J. Stegenga, H. J. Wörtche, K. Postema, G. J. Verkerke, and C. J. Lamoth, “Suitability of kinect for measuring whole body movement patterns during exergaming,” *Journal of Biomechanics*, vol. 47, no. 12, pp. 2925 – 2932, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0021929014003984

[78] P. Plantard, E. Auvinet, A.-S. Pierres, and F. Multon, “Pose Estimation with a Kinect for Ergonomic Studies: Evaluation of the Accuracy Using a Virtual Mannequin,” *Sensors*, vol. 15, pp. 1785–1803, 2015. [Online]. Available: http://www.mdpi.com/1424-8220/15/1/1785/

[79] S. Obdrzalek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel, “Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population,” in *Proceedings of the International Conference in Medicine and Biology Society (EMBS)*. IEEE, Aug. 2012, pp. 1188–1193. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6346149

[80] K. Caine, S. Šabanović, and M. Carter, “The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults,” in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, March 2012, pp. 343–350.