# Person Detection and Head Tracking to Detect Falls in Depth Maps

Michal Kepski[2] and Bogdan Kwolek[1]

[1] AGH University of Science and Technology, 30 Mickiewicza Av.,
30-059 Krakow, Poland
bkw@agh.edu.pl
[2] University of Rzeszow, 16c Rejtana Av., 35-959 Rzeszów, Poland
mkepski@univ.rzeszow.pl

**Abstract.** We present a system for fall detection in which the fall hypothesis, generated on the basis of accelerometric data, is validated by k-NN based classifier operating on depth features. We show that validation of the alarms in such a way leads to lower ratio of false alarms. We demonstrate the detection performance of the system using publicly available data. We discuss algorithms for person detection in images acquired by both a static and an active depth sensor. The head is modeled in 3D by an ellipsoid that is matched to point clouds, and which is also projected into 2D, where it is matched to edges in the depth maps.

## 1 Introduction

Visual motion analysis aims to detect and track objects, and more generally, to understand their behaviors from image sequences. With the ubiquitous availability of low-cost cameras and their increasing importance in a wide range of real-world applications such as visual surveillance, human-machine interfaces, etc., it is becoming increasingly important to automatically analyze and understand human activities and behaviors [12]. The aim of human activity recognition is an automated analysis (or interpretation) of ongoing events and their context from video data. Its applications include surveillance systems [11], patient monitoring systems [7], and a variety of systems that involve interactions between persons and electronic devices such as human-computer interfaces [5].

Automatic recognition of anomalous human activities and falls in an indoor setting from video sequences is currently an active research topic [6]. Falls are a major cause of injury for older people and a significant obstacle in independent living of the seniors. They are one of the top causes of injury-related hospital admissions in people aged 65 years and over. Thus, significant attention has been devoted to develop an efficient system for human fall detection [6].

The most common method for fall detection consists in use of a tri-axial accelerometer. However, fall detectors using only accelerometers generate too much false alarms. In this work we demonstrate that accelerometer-based fall hypothesis can be authenticated reliably by a classifier operating on features representing a person in lying pose. We show that such authentication leads to better fall detection performance, particularly to lower ratio of false positive alarms. In a scenario with a static camera, the extraction of the person is achieved through dynamic background subtraction [2,10], whereas in scenarios with an active camera, he/she is extracted using depth region growing. Because in the course of performing common activities of daily living (ADLs), such as moving a chair, other scene elements of objects, e.g. in the considered example the chair, can be separated from the background apart from the person, we track also the head of the person undergoing monitoring. Thanks to the head tracking, a controller responsible for steering of the pan-tilt head with the camera is able to keep the person in the central or specified part of the image.

The head tracking techniques have already been applied in fall detection [8]. However, on the basis of monocular cameras it is not easy to achieve long-term head tracking. As demonstrated in [4,9], the head tracking supported by the depth maps is far more reliable. Thus, in this work the head is tracked in range maps, which indicate calibrated depth in the scene, rather than a measure of intensity or color. The depth maps were acquired by the Kinect sensor.

## 2 Overview of the System

A potential fall is indicated when the signal upper peak value (UPV) from the accelerometer, see inertial measurement unit (IMU) block on Fig. 1, is greater than 2.5 g. If such an event happen, the system determines the depth connected components on the depth images, delineates the person and computes the features, which are then fed to a classifier. Given the extracted in advance equation describing the floor, the distance between the person's gravity center and the floor is calculated by a procedure in a block called feature extraction, see Fig. 1. The person is extracted on the basis of a depth reference image, which is updated whenever the person moves or the scene changes. If two or more connected components of sufficient area are extracted, or alternatively the area of the connected component representing the person is too large according to its distance to the camera, the system starts tracking of the head. The tracking begins on the oldest frame in the circular buffer.

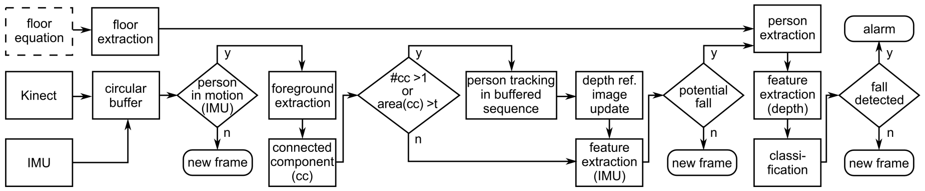

**Fig. 1.** Flowchart of the fall detection system.

# 3 Person Detection in Depth Maps

## 3.1 Human detection in depth maps from a stationary camera

The person is detected on the basis of a scene reference image, which is extracted in advance and then updated on-line. In the depth reference image each pixel assumes the median value of several pixels values from the past images. Given the sorted lists of pixels the depth reference image can be updated quickly by removing the oldest pixels and updating the sorted lists with the pixels from the current depth image and then extracting the median value. We found that for typical human motions, good results can be obtained using 13 consecutive depth images. For Kinect acquiring the images at 25 Hz we take every fifteenth image.

Figure 2 illustrates some images with the segmented person, which were obtained using the discussed technique. In the image #400 the person closed the door, which then appears on the binary image being a difference map between the current depth image and the depth reference image. As we can see, in frame 650, owing to adaptation of the depth reference image, the door disappears on the binary image. In frame 800 we can see a chair, which has been previously moved, and which disappears in frame 1000. As we can observe, the updated depth reference image allows us to extract the person's silhouette in the depth images. In order to eliminate small objects the depth connected components were extracted. Afterwards, small artifacts were eliminated. Otherwise, the depth images can be cleaned using morphological erosion. When the person does not move, the depth reference image is not updated.

In the detection mode the foreground objects are extracted through differencing the current image from such a depth reference map. Afterwards, the foreground object is determined through extracting the largest connected component in the thresholded difference map.

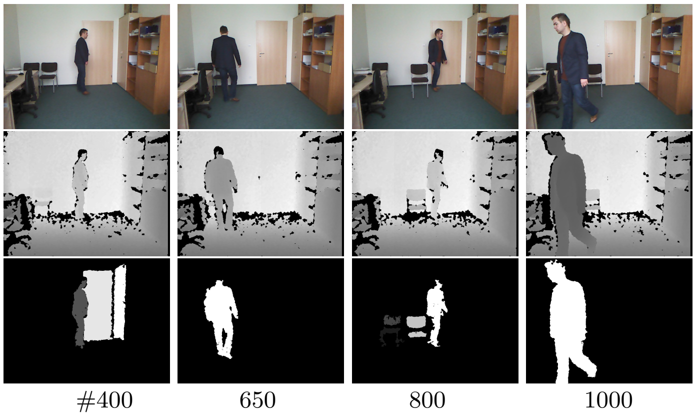

**Fig. 2.** Delineation of person using depth reference image. RGB images (upper row), depth (middle row) and binary images depicting the delineated person (bottom row).

### 3.2 Human detection in depth map sequences from active camera

In the considered scenario with an active camera, the person to be delineated in a scene moves around, while the pan-tilt unit rotates the camera to keep the subject in the central part of the depth map. Two degrees of freedom pan-tilt unit has been used to rotate the Kinect. The pursuing a moving subject is accomplished by a series of correcting saccades to the positions of the detected object in the depth maps. The object position is given as the centroid of the delineated area. The control goal of the pursuit is to hold the subject as close as possible to the central part of the depth map. This is achieved by two PID controllers, which were tuned manually to provide the assumed response time.

Region growing is a technique to separate objects or object segments from unordered pixels. The developed depth region growing starts with selecting a seed point in the current frame as an initial growing region of the whole depth region belonging to the person. Assuming that there is a common depth region between depth regions belonging to a person in two consecutive frames, such seed region is determined using the `and` operator between the previously delineated depth region belonging to person and the current depth map. The algorithm then seeks all neighboring pixels of the current region. The selected pixels are then sorted according to their depth similarities and stored in the list of candidate pixels. The depth similarity is the Euclidean distance between the depth values of a pixel from the candidate list and its closest pixel from the current region. The depth similarity is used in an examination whether a neighboring pixel around a region pixel is allowed to be included in the region. In our implementation the person region may not grow to a very large region away from a seed point. Given the location of the person in the scene as well as the distance of his/her head to the sensor we calculate the expected area of the person. The images in middle row of Fig. 4 depict some examples of the segmented person.

## 4 Head Tracking in Depth Maps

Since in the course of performing ADLs, other scene elements of objects can be separated from the background apart from the person, we track also the head of the person undergoing monitoring. As demonstrated in [8], the information about the head location can be very useful cue in fall detection.

The head is tracked using particle filtering (PF) [1]. Particle filters approximate the posterior distribution over the states with a set of particles $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)})$, where each of which has an associated importance weight $w^{(j)} \in \mathcal{R}^{+}$. In each time step, the particles are propagated further by the motion model $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. Subsequently, the importance weights are updated on the basis of the observation model $p(\mathbf{z}_k|\mathbf{x}_k)$.

The head is modeled by an ellipsoid. The observation model combines fitness score between the ellipsoid and point clouds, and fitness score between the projected ellipsoid and the edges on the depth map. A distance of a point $(x, y, z)$

to an ellipsoid, parameterized by axes of length $a, b, c$ is calculated as follows:

$$d = \sqrt{\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} + \frac{(z-z_0)^2}{c^2}} - 1 \tag{1}$$

The degree of membership of the point to the ellipsoid is determined as follows:

$$m = 1 - \frac{d}{t} \tag{2}$$

where $t$ is a threshold, which was determined experimentally. The ellipsoid fitness score is determined in the following manner:

$$f_1 = \sum_{(x,y,z)\in S} m(x,y,z) \tag{3}$$

where $S(x_0, y_0, z_0)$ denotes a set of points belonging to the head and its surround. The ellipsoid is then projected onto 2D plane using the (Kinect) camera model. For each pixel $(u, v)$ contained in the ellipse $E$ we calculate matching of the ellipse with the edges on the depth map as follows:

$$p = \begin{cases} D_e(u,v) \cdot (5 - D_d(u,v)), & D_d(u,v) < 5 \\ 0, & D_d(u,v) \geq 5 \end{cases} \tag{4}$$

where $D_e(u, v) \in 0, 1$ is pixel value in binary edge image and $D_d(u, v)$ is pixel value in an edge distance image. The ellipse fitness score is calculated as follows:

$$f_2 = \sum_{(u,v)\in E} p(u,v) \tag{5}$$

The head likelihood is calculated as follows:

$$p(\mathbf{z}_k^{(i)} | \mathbf{x}_k^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{f_1 \cdot f_2}{2\sigma^2}\right) \tag{6}$$

From time $k-1$ to $k$ all particles are propagated according to: $\mathbf{x}_k^{(i)} = \mathbf{x}_{k-1}^{(i)} + \delta^{(i)}$, where $\delta^{(i)} = \mathcal{N}(0, \Sigma)$. Figure 3 shows sample tracking results which were obtained by a particle filter consisting of 500 particles. The head tracking was

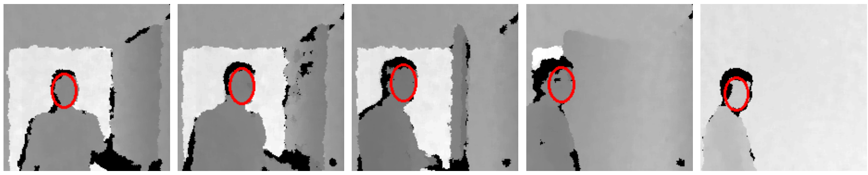

**Fig. 3.** Head tracking in depth maps acquired by a static camera, see Fig. 2.

done on the same sequence as in Fig. 2. The state vector consists of 3D location, and pitch and roll angles. As we can observe, owing to the tracking, the head can be delineated from the background, even if some foreground objects appear in the person segmented image, see frame #400 in Fig. 2.

Figure 4 shows sample tracking results, which were obtained on images acquired by the active camera. The segmented person by depth region growing is marked by green.

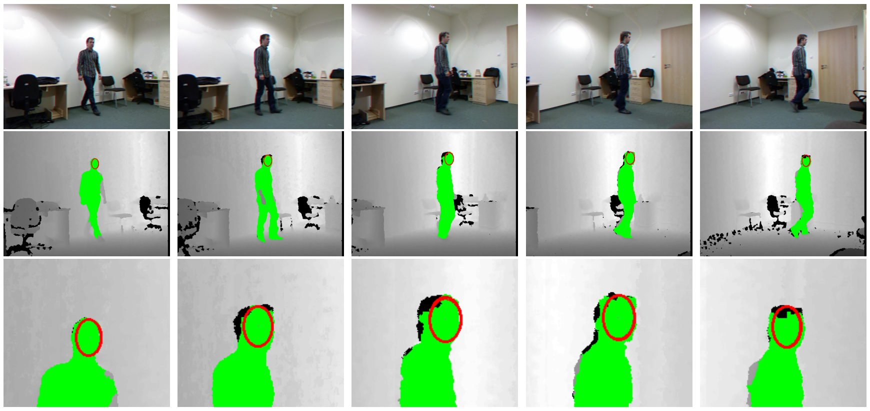

**Fig. 4.** Head tracking in depth maps acquired by an active camera. RGB images (upper row), segmented person (middle row), the ellipse overlaid on depth map (bottom row).

## 5 Lying Pose Recognition

A dataset consisting of images with normal activities like walking, sitting down, crouching down and lying has been composed in order to evaluate a k-NN classifier responsible for examination whether a person is lying on the floor. In total 312 images were selected from UR Fall Detection Dataset[1] and another image sequences, which were recorded in typical rooms, like office, classroom, etc. The discussed dataset consists of 202 images with typical ADLs, whereas 110 images depict a person lying on the floor. The following features were extracted from such a collection of depth images [3]:

- $h/w$ - a ratio of width to height of the person's bounding box
- $h/h_{max}$ - a ratio expressing the height of the person's surrounding box in the current frame to the person's height
- $D$ - the distance of the person centroid to the floor
- $max(\sigma_x, \sigma_z)$ - standard deviation from the centroid for the abscissa and the applicate, respectively.

[1] `http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html`

## 6 Experimental Results

To examine the classification performances we calculated the sensitivity, specificity, and classification accuracy. The sensitivity is the number of true positive (TP) responses divided by the number of actual positive cases. The specificity is the number of true negative (TN) decisions divided by the number of actual negative cases. The classification accuracy is the number of correct decisions divided by the total number of cases. At the beginning we evaluated the k-NN classifier on UR Fall Detection Dataset. As all lying poses were detected properly and all ADLs were classified correctly we obtained both sensitivity and specificity equal to 100%. That means that the probability of positive test is equal to 100%, given that a fall took place, and probability of negative test is also equal to 100%, given that an ADL has been performed.

Afterwards, the classification performance of the fall detector was evaluated on 10-fold cross-validation using dataset consisting of 402 negative examples and 210 positive examples. The results, which were obtained by the k-NN (for k=5) classifier are shown in Tab. 1. As we can observe, the specificity and precision of the classifier is equal to 100%.

**Table 1.** Performance of lying pose classification.

| | | | True | | |
|---|---|---|---|---|---|
| | | | Fall | No Fall | |
| Estimated | k-NN | Fall | 208 | 0 | Accuracy=99.67% |
| | | No fall | 2 | 402 | Precision=100.0% |
| | | | Sens.=99.05% | Spec.=100.0% | |

The system was tested with simulated-falls performed by young volunteers onto crash mats. The accelerometer was worn near the pelvis. Five volunteers attended in the tests and evaluations of our system. Intentional falls were performed in home towards a carpet with thickness of about 2 cm. Each individual performed ADLs like walking, sitting, crouching down, leaning down/picking up objects from the floor, lying on a bed. As expected, using only the accelerometer the number of false alarms was considerable. Experimental results demonstrated that most of them were successfully validated by depth image-based lying pose detector. In a scenario with the static camera, the verification of the fall can be achieved at low computational cost as the depth image processing is performed when the module processing the data from the accelerometer indicates a potential fall. Moreover, the accelerometer delivers information, which image should be processed to recognize the lying pose of the person. A comprehensive evaluation showed that the system has high accuracy of fall detection and very low level of false alarms.

## 7 Conclusions

In this paper we discussed a system for fall detection, which uses accelerometric and depth data. The fall hypothesis generated on the basis of accelerometric data is authenticated by depth map processing module. In maps acquired by a static camera, the person is delineated on the basis of depth reference image, whereas in images from an active camera, he/she is segmented using region growing. Since the foreground objects may appear in the segmented images we also perform head tracking. The head tracking contributes towards better localization of the person, particularly when the active camera is utilized. Given the segmented person, the algorithm extracts features, which are then used by k-NN classifier responsible for recognizing the lying pose. Owing to the use of k-NN classifier the ratio of false alarm is much lower in comparison to a fall detector using only accelerometric data. We demonstrated the detection performance of the classifier on publicly available dataset. It achieves 100% specificity and precision.

**Acknowledgments.** This study was supported by the National Science Center (NCN) within the research project N N516 483240.

## References

1. Isard, M., Blake, A.: CONDENSATION - conditional density propagation for visual tracking. Int. J. Computer Vision 29(1), 5–28 (1998)
2. Kepski, M., Kwolek, B., Austvoll, I.: Fuzzy inference-based reliable fall detection using Kinect and accelerometer. In: The 11th Int. Conf. on Artificial Intell. and Soft Comp., LNCS, vol. 7267, pp. 266–273. Springer (May 2012)
3. Kepski, M., Kwolek, B.: Unobtrusive fall detection at home using kinect sensor. In: Int. Conf. on CAIP. LNCS, vol. 8047, pp. 457–464. Springer (2013)
4. Kwolek, B.: Face tracking system based on color, stereovision and elliptical shape features. In: IEEE Conf. on Adv. Video and Signal Based Surveill. pp. 21–26 (2003)
5. Kwolek, B.: Visual system for tracking and interpreting selected human actions. Journal of WSCG 11(2), 274–281 (2003)
6. Mubashir, M., Shao, L., Seed, L.: A survey on fall detection: Principles and approaches. Neurocomputing 100, 144 – 152 (2013), special issue: Behaviours in video
7. Nait-Charif, H., McKenna, S.J.: Activity summarisation and fall detection in a supportive home environment. In: Int. Conf. on Pattern Rec. pp. 4:323–326 (2004)
8. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Monocular 3d head tracking to detect falls of elderly people. In: Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual Int. Conf. of the IEEE. pp. 6384–6387 (2006)
9. Russakoff, D.B., Herman, M.: Head tracking using stereo. Mach. Vision Appl. 13(3), 164–173 (2002)
10. Stone, E., Skubic, M.: Fall detection in homes of older adults using the Microsoft Kinect. IEEE J. of Biomedical and Health Informatics (2014)
11. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. 29, 983–1009 (2013)
12. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. CVIU 115(2), 224–241 (2011)