Fall Detection using Kinect Sensor and Fall Energy Image

Bogdan Kwolek[1] and Michal Kepski[2]

[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. One of the main reasons for low acceptance by seniors the available technology for automatic fall detection is that the existing devices generate too much false alarms. Additionally, the camera-based devices do not preserve the privacy adequately. In our approach an accelerometer is utilized to indicate a potential fall. A fall hypothesis is then verified in the second stage in which we employ a depth image, which was shot at the moment of the potential fall. A detector that was trained in advance on features extracted both from depth images and points cloud is responsible for verification whether a person is lying on the floor. After all, to reliably distinguish the fall from fall-like activities we perform final verification, in which we employ the proposed fall energy image. The fall energy image expresses the distribution of the person's motion in the set of images preceding the fall.

Keywords: Depth image and point cloud processing; fall detection.

1 Introduction

Falls are a major health risk and a significant obstacle to independent living of the seniors [13]. In response to the demand for fall detection technology, plenty of research has been done in the recent years to develop unobtrusive fall detection systems for enhancing the functional ability of the elderly and patients [16]. However, despite many efforts made to obtain reliable and unobtrusive person fall detection, current technology does not meet the seniors' needs. One of the main reasons for non-acceptance of the currently available technology by elderly is that the existing devices generate too much false alarms, which in turn lead to considerable frustration of the seniors. Additionally, the existing devices do not preserve the privacy and unobtrusiveness adequately.

In recent years, a lot of research has been done on detecting falls using a wide range of sensor types. Mubashir et al. [16] done a survey of methods used in the existing systems. Single CCD camera [18], multiple cameras [6], specialized omnidirectional ones [15] and stereo-pair cameras [9] are widely used in the vision systems for fall detection. Most of the currently available techniques for fall

detection are based on wearable sensors. Accelerometers or both accelerometers and gyroscopes are the most frequently used sensors in devices responsible for fall monitoring [17]. However, on the basis of inertial sensors it is not easy to separate real falls from fall-like activities [2] and in consequence the devices that are built on only such sensors typically trigger significant amount of false alarms. The reason is that the characteristic motion patterns of fall also exist in many actions. For instance, the crouch also demonstrates a rapid downward motion.

Recently, Kinect sensor was used in prototype systems for fall detection [10, 11, 14]. It is the world's first low-cost device that combines an RGB camera and a depth sensor. Unlike 2D cameras, it allows 3D tracking of the body movements. Thus, if only depth images are used it preserves the person's privacy. Because depth images are extracted with the support of an active light source, they are largely independent of external light conditions. Thanks to the use of the infrared light the Kinect is capable of extracting the depth images in dark rooms.

In this work we demonstrate an approach to reduce the number of false positives alarms in fall detection through the use of an accelerometer and the depth images. The accelerometer is utilized to indicate a potential fall. A fall hypothesis is then verified in the second stage in which we employ a depth image, which was shot at the time of the potential fall of the person. A detector that was trained in advance on features extracted both from depth images and points cloud is responsible for verification whether a person is lying on the floor. After all, to reliably distinguish the fall from fall-like activities we perform final verification, in which we employ the proposed fall energy image. The fall energy image expresses the distribution of the person's motion in a collection of the images, acquired in a certain period of time before the potential fall alert.

The contribution of this work is twofold: firstly we propose fall energy images (FEI) as an effective spatiotemporal representation of the human fall. Secondly, we show how to extract such energy fall images on the basis of the depth images and then how to utilize them to achieve reliable fall detection. Shape modeling using spatiotemporal features provides crucial information about human activities. In [7], a method for fall detection that is based on a combination of the eigenspace and integrated time motion images (ITMI) was developed. ITMI contain motion information and time stamps of motion occurrence. Multilayer perceptron neural network was utilized for classification of motions and detection of the fall event. In [19], a mobile human airbag system was designed for fall protection for the elderly. A Micro Inertial Measurement Unit consisting of three dimensional accelerometers, gyroscopes, a Bluetooth module and a Micro Controller Unit (MCU) is utilized to record human motion information. Through analysis of images acquired by a high-speed camera, a lateral fall can be determined on the basis of a gyro threshold. The classification of falls is performed by a support vector machine (SVM) classifier. The majority of vision based systems for fall detection do not take into account the motion information. In this work we demonstrate how to extract fall energy images using accelerometer and depth images as well as how to process them. The accelerometer helps us to extract the representative segment of the images as a representation of the fall event.

2 Person Detection in Depth Images

Depth is very useful cue to achieve reliable person detection because humans may not have consistent color and texture but have to occupy an integrated region in space. The Kinect combines structured light with two classic computer vision techniques, namely depth from focus and depth from stereo. It is equipped with infrared laser-based IR emitter, an infrared camera and a RGB camera. The IR camera and the IR projector compose a stereo pair with a baseline of approximately 75 mm. A known pattern of dots is projected from the IR laser emitter. These specs are captured by the IR camera and then compared to the known pattern. Since there is the distance between laser and sensor, the images correspond to different camera positions, and that in turn allows to use stereo triangulation to calculate each spec depth. The field of view of the system is 57° horizontally and 43° vertically, the minimum measurement range is about 0.6 m, whereas the maximum range is somewhere between 4-5 m. It captures the depth and color images simultaneously at a frame rate of about 30 fps. The default RGB video stream has size 640×480 and 8-bit for each channel, whereas the depth stream is 640×480 resolution and with 11-bit depth.

The software called NITE from PrimeSense offers skeleton tracking on the basis of depth images. However, this software is targeted for supporting the human-computer interaction, and not for detecting the person fall. Thus, in many circumstances it can have difficulties in extracting and tracking the person's skeleton. Therefore, we employ a person detection method [11], which reliably extracts the subject including situations when he/she is lying on the floor.

The person was delineated on the basis of a scene reference image, which was 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. In the setup phase we collect a number of the depth images, and for each pixel we assemble a list of the pixel values from the former images, which is then sorted in order to extract the median. 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 depth images [11]. For Kinect acquiring the images at 25 Hz we take every fifteenth image.

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

3 V-disparity Based Ground Plane Extraction

Given a depth map provided by the Kinect sensor, the disparity d can be determined in the following manner:

$$d = \frac{b \cdot f}{z} \tag{1}$$

where z is the depth (in meters), b is the horizontal baseline between the cameras (in meters), f is the (common) focal length of the cameras (in pixels). The IR camera and the IR projector form a stereo pair with a baseline of approximately $b = 7.5$ cm, whereas the focal length f is equal to 580 pixels.

Let H be a function of the disparities d such that $H(d) = I_d$. The I_d is the v-disparity image and H accumulates the pixels with the same disparity from a given line of the disparity image. Thus, in the v-disparity image each point in the line i represents the number of points with the same disparity occurring in the i-th line of the disparity image. In [12] the v-disparity maps between two stereo images were used to achieve reliable obstacle detection. In our work the v-disparity maps are extracted using depth images determined by Kinect.

The line corresponding to the floor pixels in the v-disparity map was extracted using the Hough transform operating on v-disparity values and a predefined range of parameters. The accumulator was incremented by v-disparity values [11]. Assuming that the Kinect is placed at height about 1 m from the floor, the line representing the floor should begin in the disparities ranging from 21 to 25 depending on the tilt angle of the sensor. Given the extracted line in such a way, the pixels belonging to the floor areas were determined [11]. Due to the measurement inaccuracies, pixels falling into some disparity extent d_t were also considered as belonging to the ground. Assuming that d_y is a disparity in the line y, which represents the pixels belonging to the ground plane, we take into account the disparities from the range $d \in (d_y - d_t, d_y + d_t)$ as a representation of the ground plane.

After the transformation of the pixels representing the floor to the 3D points cloud, the plane described by the equation $ax+by+cx+d$ has been recovered [11]. The parameters a, b, c and d were estimated using the RANSAC algorithm. The distance to the ground plane from the 3D centroid of points cloud corresponding to the segmented person was determined on the basis of the following equation:

$$D = \frac{|aX_c + bY_c + cZ_c + d|}{\sqrt{a^2 + b^2 + c^2}} \tag{2}$$

where X_c, Y_c, Z_c stand for the coordinates of the person's centroid.

4 Lying Pose Recognition

The recognition of lying pose was achieved using a classier trained on features representing the extracted person both in depth images and in point clouds. A data-set consisting of images with normal activities like walking, sitting down, crouching down and lying has been composed in order to train a classifier responsible for testing whether a person is lying on the floor and to evaluate its performance. Thirty five volunteers with age under 28 years attended in preparation of the data-set. The image sequences were recorded using two Kinect devices. The first Kinect was placed at a height of about one meter to the floor, whereas the second one was placed at a ceiling corner of the room. Figure 1 shows example depth images seen from such two different views.

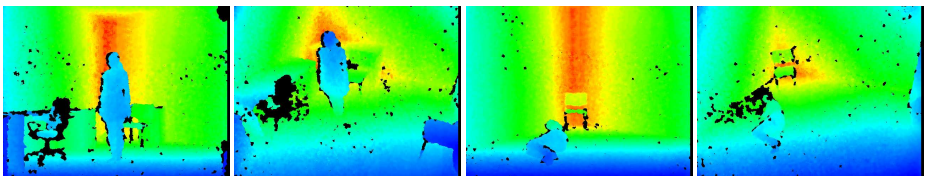

Fig. 1: Person in depth images seen from two different views.

In total 312 images representing typical human actions were selected and then utilized to extract the following features:

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

Figure 2 depicts a scatterplot matrix for the employed attributes, in which a collection of scatterplots is organized in a two-dimensional matrix simultaneously to provide correlation information among the attributes. In a single scatterplot two

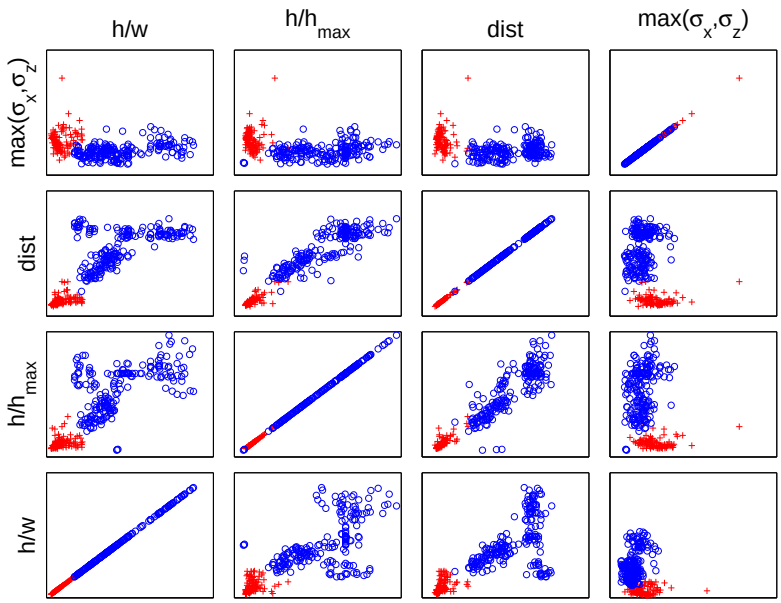

Fig. 2: Multivariate classification scatter plot for features used in lying pose recognition.

attributes are projected along the x-y axes of the Cartesian coordinates. As we can observe, the overlaps in the attribute space are not too significant. We considered also another attributes, for instance, a filling ratio of the rectangles making up the person's bounding box. The worth of the considered features was evaluated on the basis of the information gain [4], which measures the dependence between the feature and the class label. In the assessment of the discrimination power of the considered features and selecting the most discriminative ones we utilized the `InfoGainAttributeEval` procedure from the Weka [5], which is a collection of machine learning algorithms.

5 Fall Energy Image

Several motion features have been proposed until now to represent people activities, such as Motion History Image (MHI) [1]. Usually, the MHI is generated on the basis of binary images, where the person silhouette sequence is condensed into gray scale images as a weighted combination of all motion images. The result of such a motion condensation is a scalar-valued image in which more recently moving pixels are brighter. One of the advantages of the MHI representation is that a range of action images may be encoded in a single motion-shape. Typically, in action recognition phase such a static shape pattern is compared with pre-stored action prototypes.

The Fall Energy Image is an average of all silhouette images of a single fall. Such a spatiotemporal energy map spans the time scale of person fall. The energy map is calculated using a number of binary silhouette images before the fall. The images are scaled according to the distance of the person to the camera. We assume that a fall occurs if the signal upper peak value from the accelerometer is greater than $3g$. Figure 3 illustrates example fall energy images with the corresponding plots of signal upper peak value (UPV) vs. time. As we can observe, both actions have quite similar characteristics in the acceleration domain, but totally different fall energy maps.

The weighted average (moment) of the fall energy expressed by pixel intensities was computed using moments as follows:

$$\begin{aligned} x_c &= \frac{\sum_x \sum_y xP(x,y)}{\sum_x \sum_y P(x,y)} \\ y_c &= \frac{\sum_x \sum_y yP(x,y)}{\sum_x \sum_y P(x,y)} \end{aligned} \tag{3}$$

where x, y are pixel coordinates. The major length and width (eigenvalues) of the fall energy has been calculated in the following manner [8]:

$$\begin{aligned} l &= 0.707\sqrt{(a+c) + \sqrt{b^2 + (a-c)^2}} \\ w &= 0.707\sqrt{(a+c) - \sqrt{b^2 + (a-c)^2}} \end{aligned} \tag{4}$$

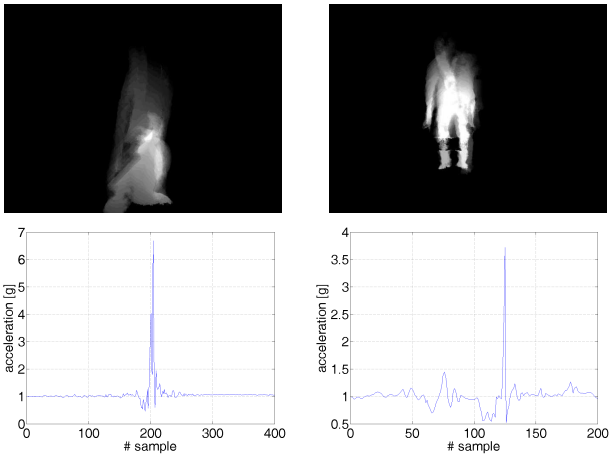

Fig. 3: Fall energy images for a forward fall (left) and sitting on a chair (right) with corresponding plots of signal upper peak value vs. time (bottom row).

where

$$a = \frac{M_{20}}{M_{00}} - x_c^2, \;\; b = 2(\frac{M_{11}}{M_{00}} - x_c y_c), \;\; c = \frac{M_{02}}{M_{00}} - y_c^2, M_{00} = \sum_x \sum_y P(x,y),$$
$$M_{11} = \sum_x \sum_y xyP(x,y), \; M_{20} = \sum_x \sum_y x^2 P(x,y), \;\; M_{02} = \sum_x \sum_y y^2 P(x,y).$$

We calculated also the average fall energy, i.e. the mean value of non-zero pixel values in the fall energy image $P(x,y)$ as well as the Euclidean distance d_E between the weighted average location of the fall energy (y_c, x_c) and the geometrical centroid of the thresholded energy map. Figure 4 depicts the scatter plot matrix for such energy features. The features were extracted on the basis of 30 image sequences in which half of them contained person falls. The remaining sequences contained person activities, which were very similar to fall. The activities were performed close to the floor and contained actions consisting in sitting on the floor, laying down on the floor, for instance to raise an object, etc. The features were extracted on the basis of 30 depth images just before the human fall, which in turn was signaled by a procedure processing data from the accelerometer. That means that the FEI image expresses the fall energy in about 1 sec. As we can observe, on the basis of such a set of features the person fall can be distinguished from the non-fall activities. We considered also energy features extracted on the basis of the bank of Log-Gabor filters. Their worth was evaluated on the basis of the information gain and then compared to the discrimination power of the above discussed features. The experimental results showed that their worth is not worse in comparison to Gabor filter based energy features and therefore we decided to use them in the evaluation of the whole system. It is worth to note that they can be extracted in considerably shorter time.

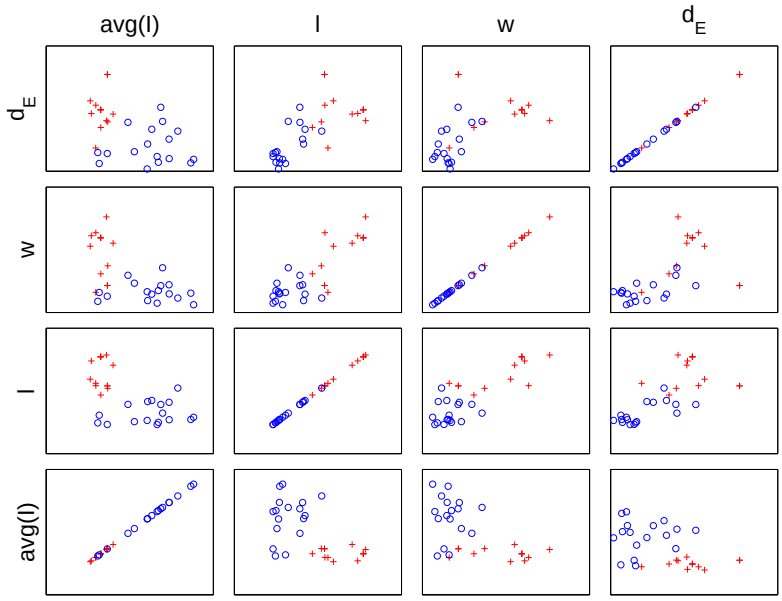

Fig. 4: Multivariate classification scatter plot for features extracted on fall energy images.

6 Experimental Results

Thirty five young health volunteers with age under 28 years attended in preparation of the data-sets and in the evaluation of the fall detection accuracy. To show the resistance of the system to the placement of the camera the images were acquired by two Kinect devices. The motion data were acquired by a wearable smart device (Sony PlayStation Move) containing accelerometer and gyroscope sensors. Data from the device were transmitted wirelessly via Bluetooth and received by a laptop computer. In all, 312 images acquired by two Kinect devices were selected and then used to evaluate the k-NN classifier responsible for checking whether the person is lying on the floor. The number of images with a fall was equal to 110. We evaluated also KStar [3], SVM and multilayer perceptron (MLP) classifiers. The KStar and MLP classified all falls correctly, whereas the remaining algorithms incorrectly classified two instances. A k-NN based motion classifier was trained on 30 image sequences of which 15 contained fall events. Its accuracy was evaluated in 10-fold cross-validation and one fall was classified incorrectly. The SVM and KStar classified all falls correctly.

The complete system for fall detection was tested with simulated-falls performed by young volunteers under supervised conditions 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 walk-

ing, 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 can be ignored owing to the use of our recognition module of the lying pose. This operation is done at low computational cost as the verification of the fall is performed if the module processing the data from the accelerometer triggers the alarm. Moreover, on the basis of the accelerometer based alarm the system obtains information which image should be processed to decide if an event consisting in person lying on the floor takes place. All person activities that have been considered in the previous work [10] were classified correctly. During the evaluation of the system the volunteers found several fall-like actions, which were not considered in the previous work and for which the two-stage algorithm triggered false alarms. The experimental results obtained on the system with three modules, i.e. accelerometer, lying pose recognition and fall energy analysis demonstrated that the fall energy features are very useful in further reduction of the false alarm ratio. A comprehensive evaluation showed that the system has high accuracy of fall detection and very low level of false alarms. It demonstrated that the placement of the cameras does not have an influence on the classification accuracy.

The depth images were acquired by the Kinect sensors using OpenNI. The system was implemented in C/C++ and runs at 25 fps on 2.4 GHz I7 notebook. The most computationally demanding operation is extraction of the depth reference image of the scene. For images of size 640×480 the computation time needed for extraction of the depth reference image is about 9 milliseconds. In order to reduce the computational overload the depth reference images were only updated if on the image acquired in the moment of the fall, two or more blobs had been detected. In practice, we examined the thresholded difference between the current depth map and the reference map in terms of the number of blobs.

7 Conclusions

In this work we demonstrated how to achieve reliable fall detection with low false positives number. Given the alarm trigger obtained on the basis of data from wireless accelerometer, the system extracts the person features from the corresponding depth image and point clouds. The system uses them in a k-NN classifier to examine if the person is lying on the floor. In order to further reduce the false alarm ratio the system extracts fall energy images from a sequence of images up to the fall and then employs the energy features in a k-NN classifier. Experimental results demonstrated that this leads to considerable reduction of false alarms and high detection ratio. The system preserves the privacy of the user and works in poor lighting conditions.

Acknowledgment

This work has been supported by the National Science Centre (NCN) within the project N N516 483240.

References

1. Ahad, M.A.R., Tan, J.K., Kim, H., Ishikawa, S.: Motion history image: its variants and applications. Mach. Vision Appl. 23(2), 255–281 (Mar 2012)
2. Bourke, A., O'Brien, J., Lyons, G.: Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. Gait & Posture 26(2), 194–199 (2007)
3. Cleary, J., Trigg, L.: An instance-based learner using an entropic distance measure. In: Int. Conf. on Machine Learning. pp. 108–114 (1995)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley (1992)
5. Cover, T.M., Thomas, J.A.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd edn. (2005)
6. Cucchiara, R., Prati, A., Vezzani, R.: A multi-camera vision system for fall detection and alarm generation. Expert Systems 24(5), 334–345 (2007)
7. Foroughi, H., Naseri, A., Saberi, A., Yazdi, H.: An eigenspace-based approach for human fall detection using integrated time motion image and neural network. In: 9th Int. Conf. on Signal Processing. pp. 1499–1503 (2008)
8. Horn, B.: Robot Vision. The MIT Press, Cambridge, MA (1986)
9. Jansen, B., Deklerck, R.: Context aware inactivity recognition for visual fall detection. In: Proc. IEEE Pervasive Health Conference and Workshops. pp. 1–4 (2006)
10. Kepski, M., Kwolek, B., Austvoll, I.: Fuzzy inference-based reliable fall detection using Kinect and accelerometer. In: The 11th Int. Conf. on Artificial Intelligence and Soft Computing. pp. 266–273. LNCS, vol. 7267, Springer-Verlag (May 2012)
11. Kepski, M., Kwolek, B.: Human fall detection using Kinect sensor. In: Proc. of the 8th Int. Conf. on Computer Recognition Systems, Advances in Intelligent Systems and Computing, vol. 226, pp. 743–752. Springer (2013)
12. Labayrade, R., Aubert, D., Tarel, J.P.: Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In: Intelligent Vehicle Symposium, 2002. IEEE. vol. 2, pp. 646 – 651 vol. 2 (june 2002)
13. Marshall, S.W., Runyan, C.W., Yang, J., Coyne-Beasley, T., Waller, A.E., Johnson, R.M., Perkis, D.: Prevalence of selected risk and protective factors for falls in the home. American J. of Preventive Medicine 8(1), 95–101 (2005)
14. Mastorakis, G., Makris, D.: Fall detection system using Kinect's infrared sensor. J. of Real-Time Image Processing pp. 1–12 (2012)
15. Miaou, S.G., Sung, P.H., Huang, C.Y.: A customized human fall detection system using omni-camera images and personal information. Distributed Diagnosis and Home Healthcare pp. 39–42 (2006)
16. Mubashir, M., Shao, L., Seed, L.: A survey on fall detection: Principles and approaches. Neurocomputing 100, 144 – 152 (2013), special issue: Behaviours in video
17. Noury, N., Fleury, A., Rumeau, P., Bourke, A., ÓLaighin, G., Rialle, V., Lundy, J.: Fall detection - principles and methods. In: Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. pp. 1663–1666 (2007)
18. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Monocular 3D head tracking to detect falls of elderly people. In: Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. pp. 6384–6387 (2006)
19. Shi, G., Chan, C.S., Li, W.J., Leung, K.S., Zou, Y., Jin, Y.: Mobile human airbag system for fall protection using MEMS sensors and embedded SVM classifier. Sensors Journal, IEEE 9(5), 495–503 (2009)