A SKELETON FEATURES-BASED FALL DETECTION USING MICROSOFT KINECT V2 WITH ONE CLASS-CLASSIFIER OUTLIER REMOVAL

O.S. Seredin [1, *], A.V. Kopylov [1], S.-C. Huang [2], D.S. Rodionov [1]

[1] Tula State University, Institute of Applied Mathematics and Computer Science, 300012 Tula, Russia – oseredin@yandex.ru, and.kopylov@gmail.com
[2] National Taipei University of Technology, Department of Electronic Engineering, Taipei 106, Taiwan – schuang@ntut.edu.tw

Commission VI, WG VI/4

KEY WORDS: Fall Detection, Movement Analysis, Skeleton Description, RGB-D Camera, Privacy Preserving Elderly People Care

ABSTRACT:

The real-time and robust fall detection is one of the key components of elderly people care and monitoring systems. Depth sensors, as they became more available, occupy an increasing place in event recognition systems. Some of them can directly produce a skeletal description of the human figure for compact representation of a person's posture. Skeleton description makes the output of source video or detailed information about the depth outside the system unnecessary and raises the privacy of the entire system. Based on a comparative study of different RGB-D cameras, the most promising model for further development was chosen - Microsoft Kinect v2. The TST Fall Detection Dataset v2 is used here as a base for experiments. The proposed algorithm is based on the skeleton features encoding on the sequence of neighboring frames and support vector machine classifier. A version of a cumulative sum method is applied for combining the individual decisions on the consecutive frames. It is offered to use the one-class classifier for detection of low-quality skeletons. The 0.958 accuracy of our fall detection procedure was obtained in the cross-validation procedure based on the removal of records of a particular person from the database (Leave-one-Person-out).

1. INTRODUCTION

An Activity Monitoring, Identification, and Tracking system for elderly people care, besides technical characteristics of accuracy, speed of response and reliability, must possess two crucial properties to be used in everyday practice: privacy and unobtrusiveness. According to research (Wild et al., 2008) elderly people are more likely to adopt in-home surveillance technologies if they are private and unobtrusive, namely, if they do not bring discomfort in everyday life, do not require to wear and maintain any device and to attain new technical skills and, that is extremely important, they do not capture any video images. High-resolution camera-based monitoring systems demonstrate a promising event recognition accuracy, and although they are not utilizing wearable devices, the capturing and processing of video data, do not make the system completely privacy preserving. That's why many researchers avoid RGB cameras and prefer other types of sensors like infrared sensors, radio signal strength etc.

At the same time studies show that privacy-preserving data representations, such as silhouettes (Demiris et al., 2009) or skeletons (Vemulapalli et al., 2014) could reduce the concern of elderly people relative to image and video-based surveillance systems.

The module of real-time and robust fall detection is one of the key components of elderly people care and monitoring systems. From the biomechanical point of view, a fall is determined by the human body posture and dynamics of movement of its parts.

RGB-D cameras take a significant place in the domain of event recognition as recently they have become more accessible, they can provide reliable depth information of the scene and can be easily adapted to the new environment. It is possible to design activity recognition systems using depth sensors, like Microsoft Kinect or similar devices, exploiting depth maps, which are less affected by environment illumination changes, more other they can provide body shape and skeleton, and simplify the problem of human figure detection and segmentation. For example, Leone et al. proposed an elderly people monitoring system, based specifically on critical events detection by 3D range camera (Leone et al., 2011). Analysis of depth information instead of initial video frames increases the privacy-preserving properties of the technique. The availability of skeleton points extracted from the depth data allows to have a compact representation of the human pose and prevents of the detailed depth image or initial video contemplation, so rises the privacy of a system to the next level (Cippitelli et al., 2016).

We propose here a wearable device-free and non-privacy invasive fall detection system which relies on just the skeleton data from RGB-D sensor and does not need to transmit and store private data outside the elderly person house.

The basic overall architecture of the proposed system is shown in Fig. 1. It should be noticed, that in the case of registered fall, the alert message is sent first to the elderly person for confirmation. If he or she is able to react to the message, elderly people can stop the succeeding alert message propagation to prevent a false alarm. Otherwise, the relative or social worker will receive an emergency call. We propose several privacy-based access levels. At the lowest level, only the information in a message itself is available outside the monitoring house. At the second level, actual RGB video of a person's fall is substituted by the credible animated pictogram-based on skeleton data. The highest level is intended to the detailed analysis of the situation using actual video record of the detected fall and the current state of things. Although falls have a very serious risk for an elderly person, this is not the sole danger. Thus, we propose an extensible system based on modular architecture to easily add new functionality using other

detectors type e.g. fire and smoke detectors, intrusion detectors, etc.

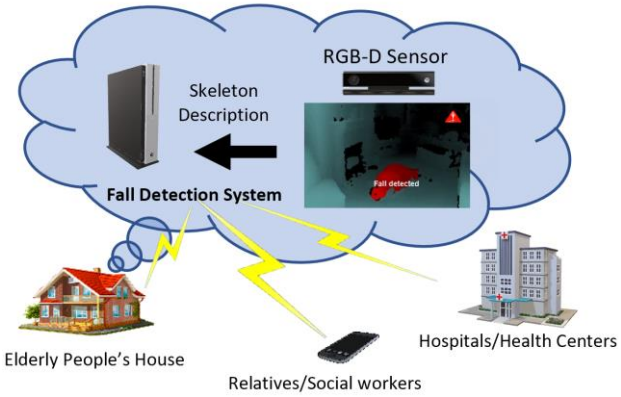

Figure 1. The basic overall architecture of the proposed system

The core of the system is based on machine learning principles, and object representation plays an important role for development of fall detection algorithms. In this work, one of the most natural and rich enough descriptors of the pose of a human body, a skeleton, is used. A skeleton could be obtained directly from modern RGB-D sensors or constructed based on multiple views from common RGB cameras. Based on a comparative study of different RGB-D cameras, the most promising model for further development, Microsoft Kinect v2, is chosen. Statistical results indicate that Kinect replaced the camera and became the most popular type of sensor used in fall detection systems after 2014 (Xu et al., 2018).

Three groups of methods of human figure representation on the basis of skeleton description can be found in the literature on fall detection.

The first group of methods uses the general geometrical characteristics of a skeleton like bounding rectangle, geometric moments and their invariants, positions or distances from the particular skeleton point, e.g. the point, corresponding to the head or center of mass, from the floor, etc. These methods are less sensitive to the skeleton estimation defects but do not have enough flexibility to operate well in the complex or changing environment. For example, the method (Mastorakis and Makris, 2012) apply RGB-D camera mounted on a tripod and use characteristics of the bounding rectangle for fall detection. By measuring the rate of decrease or increase in width, height and depth of the bounding rectangle, it becomes possible to detect falls accurately and in real time, excluding false positives in normal human activities (such as lying on the bed or on the floor. This technique does not require knowledge about the scene, such as the equation of the plane of the floor. The algorithm (Pathak and Bhosale, 2017) uses an RGB-D camera located under the ceiling and directed vertically. Such an arrangement allows to avoid practically overlapping of the person with the objects which are in the room. The fall is determined by the threshold of 0.4 m from the floor of the person's coordinates. This algorithm uses only 3D information from the camera, which makes it invariant to ambient light, and also increases the privacy of the system, as using only the information from the 3D camera is impossible to detect a person's face. The disadvantages of this method include a limited field of view at this location of the camera and the problematic detection of everyday activity, as sitting and lying. Method (Bevilacqua et al., 2014) also uses a bounding rectangle, the first derivative in height and the first derivative in the width-depth composition are calculated. But these parameters are subject to noise because of the low accuracy of the sensor. The method involves the use of Kalman filter for a more accurate estimate of the rate of change of height and the composition of width and depth. To exclude false positives when performing normal actions, the following idea is used: the y-coordinate of the upper left corner of the bounding rectangle is tracked since it is close to the y-coordinate of the centroid of the head. If this coordinate is less than the required threshold, the fall is fixed.

Method (A. Mundher and Jiaofei, 2014), based on finding the distance from the skeleton points to the floor. For this method requires mandatory presence the floor in the surveillance area. A fall is defined when the points of the skeleton (head, the center of shoulders, the center of the pelvis, right and left ankle) is below the threshold level.

The second group of methods utilizes the correspondence between the skeleton and human body parts (Vemulapalli et al., 2014), (Bian et al., 2015).

The third group relies on the skeleton vertices positions in 3D space, roughly corresponding to the joints location. Pairwise relative positions (Wang et al., 2012) or skeleton covariance matrices (Hussein et al., 2013) are often used for relative pose description. However, the joint positions itself are not sufficient to accurate fall detection and additional spatiotemporal features could be applied.

We propose here an extended version of Euclidian Distance Matrix (Wang et al., 2012) to take into account the dynamics of a human posture by interframe speed and acceleration features. Therefore, the SVM classifier is trained under the data from several frames. It is offered to use the one-class classifier for outlier removal to exclude low-quality skeletons. A version of a cumulative sum method is applied then for combining the individual decisions on the consecutive frames. The 0.95 accuracy of our fall detection procedure was obtained in the cross-validation procedure based on the removal of records of a particular person from the database (Leave-one-Person-out).

2. DEPTH SENSORS AND DATASETS

There are different depth sensors (see Fig. 2) like Microsoft Kinect (Zhang, 2012), Intel Real Sense Depth Cameras [https://click.intel.com/realsense.html], Asus Xtion Pro Live [https://www.asus.com/3D-Sensor/Xtion_PRO/], Asus Xtion 2 [http://xtionprolive.com/asus-xtion2], and more. Comparison of cameras main features is present in Table 1. Kinect is one of the Microsoft productions for capturing color and depth images, building the body skeleton. We used Kinect version 2 in this research (just for capturing skeleton data).

The most recent and exhaustive analysis of databases used for Kinect-based data techniques presented in (Cai et al., 2017). Authors study RGB-D benchmark datasets, most of them created in a time range from 2011 to 2014. From the point of interest for our research, we concern on the human activity analysis.

The University of Rzeszow created UR Fall Detection dataset in 2014 (Kepski and Kwolek, 2012), which devotes to detecting and recognizing human falls. In this dataset, the video

sequences are recorded by two Kinect cameras. One is mounted at the height of approximate 2.5m such that it is able to cover the whole room (5.5 m2).

Figure 2. RGB-D cameras: (a) Microsoft Kinect v1, (b) Microsoft Kinect v2, (c) Asus Xtion PRO LIVE, (d) Intel RealSense D435

The other one is supposed to be parallel to the floor with a distance about 1m from the ground. In this dataset, there are totally 60 sequences that record 66 falls when conducting common daily activities, such as walking, taking or putting an object from the floor, bending right or left to lift an object, sitting, tying laces, crouching down and lying. Meanwhile, corresponding accelerometer data are also collected using an elastic belt attached to the volunteer. Unfortunately, this dataset has no information about skeletons structure.

	Microsoft Kinect v1	Microsoft Kinect v2	Asus Xtion PRO LIVE	Intel RealSence D435
RGB Sensor Resolution and Frame Rate	640x480 30 fps	1920x1080 30 fps	1280x1024	1920x1080 30 fps
Depth Sensor Resolution and Frame Rate	320x240 30 fps	512x424 30 fps	640x480 30 fps	up to 1280x720 up to 90 fps
Maximum distance of use	4.5 m	4.5 m	3.5 m	up to 10 m
Horizontal Field of View	57	70	58	86
Vertical Field of View	43	60	45	57
Skeletal tracking	Yes	Yes	NO	NO

Table 1. Main features of RGB-D cameras

TST Fall Detection Dataset v2 (Gasparrini et al., 2016) [IEEE DataPort TST Fall Detection Dataset v2, URL: https://ieee-dataport.org/documents/tst-fall-detection-dataset-v2] contains depth frames and skeleton joints collected using Microsoft Kinect v2 and acceleration samples provided by an IMU during the simulation of ADLs and falls.

The dataset is composed by ADLs (Activity of Daily Living) and falls simulated by 11 young actors.

The following actions are part of the **ADL category**:
- sit, the actor sits on a chair;
- grasp, the actor walks and grasps an object from the floor;
- walk, the actor walks back and forth;
- lay, the actor lies down on the floor;

The following actions are part of the **FALL category**:
- front, the actor falls from the floor and ends up lying;
- back, the actor falls backward and ends up lying;
- side, the actor falls to the side and ends up lying;
- UpSit, the actor falls backward and ends up sitting.

The total amount of records in The TST Fall Detection Dataset v2 is 264, and the total number of frames is 46 418. The framerate of records is 30 fps. The shortest record lasts 2.5 sec. (75 frames) and the longest one lasts 15.4 (463 frames) sec. The average record time is 5.8 sec (175 frames).

The TST Fall Detection Dataset v2 is one of the most recent datasets which can provide rich enough skeleton descriptions for human actions, and it is this dataset is selected for experimental evaluations. Although all records in the dataset are divided according to the action category and type, not all frames in each record actually correspond to the action. For example, fall records often include preliminary walking and lying then the fall is already finished. That's why frames in the dataset need to be labeled as belonging to the action or not.

We made the manual labeling of frames for all records in the fall category. The number of frames assigned to the fall category by experts is 8 306. The shortest, average and longest falling time are, correspondingly, 0.9 sec. (27 frames), 2 sec (62 frames) and 3.2 sec (97 frames). Fig. 3 shows the distribution of actual fall fragments (red) over all records in the dataset. Each row in Fig. 2 corresponds to one record in the dataset. Quite debatable is the question of the exact definition of the beginning and end of the fall. It is difficult to specify particular frames associated with the fall action segment of a record. We used an agreed solution from at least two human labelers.

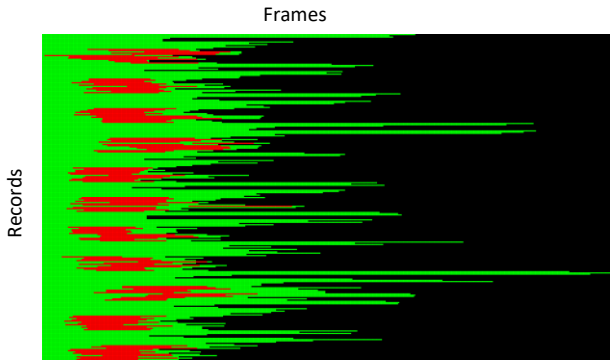

Figure 3. Representation of fall action segments (red), obtained as a result of manual labelling of records (each raw in figure) of all 11 persons in TST Fall Detection dataset v2

Such labeling, besides training the classifier, gives the additional possibility for evaluation of fall detection algorithms.

3. SKELETON BASED FEATURES

A skeleton description of the human body possesses the increasing popularity in the action recognition due to its compactness and availability. Modern RGB-D cameras like Microsoft Kinect allows to immediately obtain the skeleton points in a real time, see Fig. 4. Moreover, the accuracy of representation improves with each new camera version, opening a wide perspective for future applications. In comparison to other forms of visual data, like scene radiance in infrared or visible ranges and depth maps, the skeleton data is easier to

transmit while preserving the high level of privacy in the surveillance system.

Following the observation in (Wang et al., 2012) that the pairwise relative positions of the joints provide more discriminative features than 3D joint positions, we use the Euclidian Distance Matrix (EDM) of skeleton points, normalized by the height of the observed human figure to represent the posture.

Although, the rank of EDM of the 3D points is no more than five, and, in theory, we don't need to keep all distances to completely represent the point set relative configuration, distances between all significant skeleton points are kept in due of robustness increasing. As the position of fingers and feet are omissible for fall detection, we use only 17 from 25 points of the skeleton, provided by Kinect v2 as shown in Fig. 4.

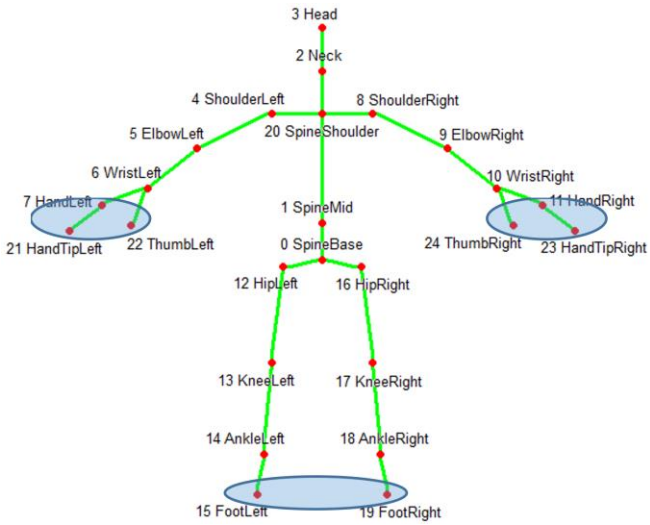

Figure 4. Skeleton provided by Microsoft Kinect v2 and points excluded from analysis (grey ovals)

So, we obtained 136 EDM features. Additionally, we use the heights of chosen 17 points and so the number of posture features is equal to 153. The dynamics of human activity is described by taking into account skeleton descriptions of a human body on several neighboring frames. Differences in posture features in neighboring frames produce extra 153 features (interframe speed). Finally, we add extra 153 attributes as a change of interframe speed features between neighboring frames –analog of acceleration. So, for each frame (starting from the third one) of the record we calculate 459 features. Finally, the preliminary data matrix of frames from the whole database contains 45809 objects of two classes – FALL (8306) and ADL (37584) in 459 feature space.

4. OUTLIER SKELETON DETECTION AND REMOVAL BASED ON THE ONE-CLASS CLASSIFIER

Unfortunately, sometimes skeletal representation from Kinect subjects to failure. Fig. 5 shows the examples of such errors. This aspect frustrates our idea to obtain relevant features from skeleton points. A large amount of such collapsed skeletons makes it hard to manually label them. We suggest using outlier detection technique based on SVM one-class classification (Schölkopf et al., 2001).

Following the idea to sift out near 3% of frames we used following parameters of Gaussian RBF based classifier: *gamma*=0.0001, *nu*=0.03.

Below is a statistic for one class classifier trained on all dataset applied for only records with falls: number of outliers in FALL class – 803, number of outliers in ADL class – 151. For all records number of outliers in ADL class – 391.

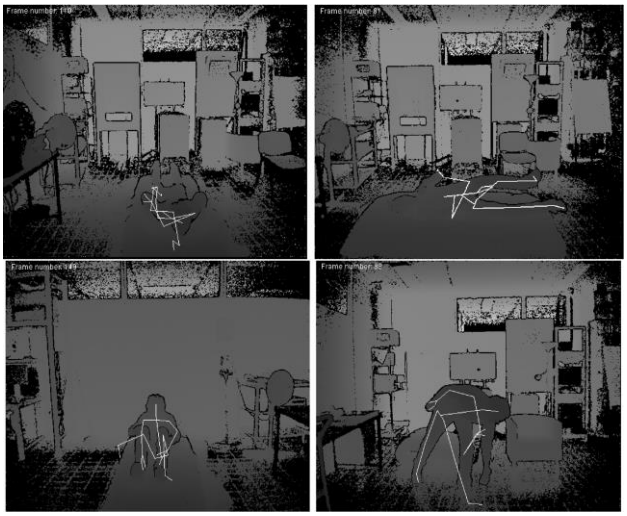

Figure 5. Examples of skeleton collapse in TST Fall Detection dataset v2

It is not possible to make a direct decision about class membership for frames marked as outliers. In the next Section, we suggest the technique for using information from previous frames to handle this obstacle.

5. COMBINING THE INDIVIDUAL DECISIONS ON THE CONSECUTIVE FRAMES

We used here two-class SVM classifier with Gaussian RBF kernel for separating Fall and ADL classes. According to the decision function as the distance from the optimal separating hyperplane the results are quite promising, see Fig. 6 in comprising with Fig. 3. We see that procedure perceives the tendency of falls. However, the 10-fold cross-validation procedure for two-class classification problem after outliers filtering gives quite pure statistical results shown in Fig. 7 as confusion matrix. The average accuracy is 0.874 and F1-measure just 0.783.

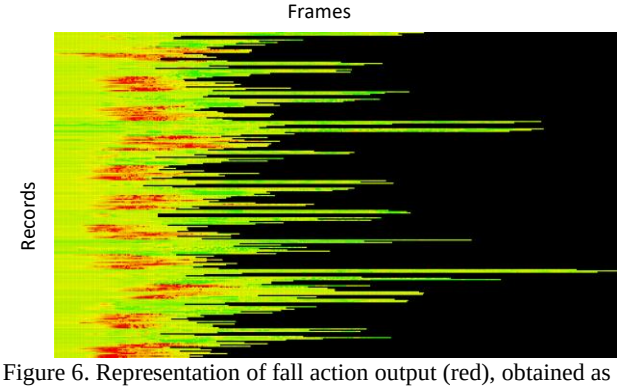

Figure 6. Representation of fall action output (red), obtained as a result of SVM classification of records of 11 persons in TST Fall Detection dataset v2

It is obvious that recognition based on the individual frame in a time-process task without taking in to account the interrelations among decisions is weak and individual frame decisions need to be combined to provide more robust and accurate solution.

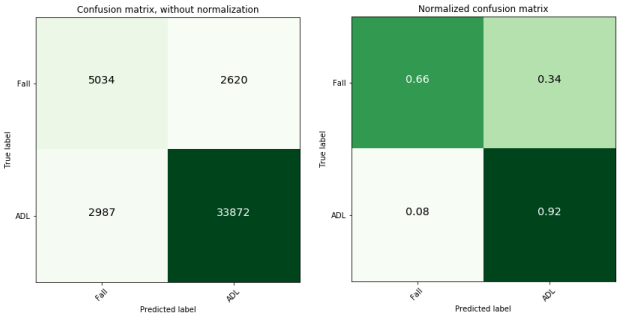

Figure 7. Confusion matrices for 10-fold cross-validation procedure for two-class classification problem under the frames.

In this Section, we will suggest the simple and effective procedure for significant improvement of overall decision. We will follow here the idea of pattern recognition in interrelated data described in (Dvoenko et al., 2004). In the framework of this approach, the sequence of decisions for frames can be represented as a Markov chain with two stages and the final decision is based on the posterior distribution of classes given the feature description of all frames up to the current. The nature of the actions like fall leads to the relatively long frame sequences with a constant stage. In combination with the online character of the fall-detection system, this allows applying a sequential analysis technique developed by E. S. Page (Page, 1954). The method is based on the calculation of the cumulative sum of a statistical characteristic, that has the meaning of the conditional likelihood-ratio. In this paper, we use a distance from a hyperplane as such a characteristic.

The procedure is as follows. In the initial frame the cumulative sum $L(0)=0$. The cumulative sum recurrently calculates for the subsequent frames according to the following rule:

$$L(t)=\begin{cases} b_1, L(t-1)+d(t)>b_1, \\ L(t-1)+d(t), b_2<L(t-1)+d(t)<b_1, \\ b_2, L(t-1)+d(t)<b_2, \end{cases} \quad (1)$$

where b_1 and b_2 = up and bottom thresholds,

$d(t)$ = the distance from the separating hyperplane for the skeleton on the frame at the time instance t.

The change in the class label is assumed to be detected at the time point t^*, then the cumulative sum $L(t)$ reaches one of the thresholds. The particular time point of the class change corresponds to the moment of the latest break from the one threshold before reaching the second one. When determining the values of the thresholds b_1 and b_2, the necessary sensitivity of the procedure must be set. The more their absolute values are, the less sensitive this procedure will be not only to noises but also to short changes in distance from the hyperplane.

The outliers do not participate in two-class training procedure and on the classification stage the distance from hyperplane $d(t)$ becomes senseless. Nevertheless, the procedure (1) formally requires some value of $d(t)$. The cumulativeness requirement implies to keep $L(t)=L(t-1)$ for the outlier, which is equivalent to set $d(t)=0$.

The procedure meets the requirements of computational simplicity and online processing. The classification algorithm based on this procedure is a real-time algorithm and is actually a simplified implementation of the dynamic programming principle.

So, the final procedure we proposed in this paper consists in seven major steps.

Cumulative sum initialization.

For each frame in a sequence perform the following operations:

- acquiring skeleton descriptions by Microsoft Kinect v2;
- calculation of EDM and dynamic features;
- outlier removal based on one-class classifier;
- calculation of the distance from a hyperplane, obtained by pretrained two-class SVM classifier.

Updating the cumulative sum and threshold touch checking.

FALL or ADL detection.

6. EXPERIMENTAL RESULTS

For evaluation of the quality of the proposed method, we use the data from TST Fall Detection dataset v2 in cross-validation procedure. However, the standard cross-validation procedure should be corrected to remove not just one frame or portion of frames as a test batch, but the whole information from one person. We will call such kind of validation as Leave-one-Person-out procedure. Thereby, while the procedure is rather unfavorable from the point of the final score, it simulates the real-world situation of applying the classifier to the unknown scene with a new person in it.

The overall experimental study of the proposed procedure consists of several stages.

1. Dividing database on 2 portions: ten persons for training set (240 records) and one person for test set (24 records).
2. For the training set outlier filtering is performing based on one-class classifier.
3. Two class SVM classifier with Gaussian RBF kernel is trained with 10-fold cross-validation parameters (gamma and C) tuning on frames.
4. Adjusting the parameters of the cumulative sums procedure according to the minimum difference between labeling and output of the procedure (1).
5. All 24 records for one test person are classified using the outlier filter and SVM model under the cumulative sum procedure with binary output. Three types of statistics are calculating: a) the accuracy of classification for the whole record (is fall present inside record or not) and b) accuracy for frames which give us a quantitative estimation of coincidence in position and duration of fall segments determined by the procedure and labeled by experts; c) average delay of fall start position determined by suggested classifier.
6. Repeating pp. 1-5 for all persons in the database.

Tables 2-4 show the results of SVM training on frames with different parameters on the training set (10 persons) and applying decision function to the data set for one-person.

gamma\C	1000	10000	100000
0.001	0.922	0.912	0.905
0.0001	0.930	0.930	0.924
0.00001	0.909	0.922	0.931
0.000001	0.891	0.903	0.913

Table 2. Persons 1-10 SVM parameters tuning in cross-validation procedure on frames

gamma\C	1000	10000	100000
0.001	0.863	0.839	0.837
0.0001	0.889	0.884	0.819
0.00001	0.903	0.896	0.885
0.000001	0.892	0.903	0.896

Table 3. Person ID=11 SVM recognition accuracy

After choosing the best parameters of SVM classifier the tuning of cumulative sums procedure is performed. Table 4 reports a log for parameters of b_1 (lower) and b_2 (upper) tuning for the training dataset (10 persons) and applying them to the test person ID=11 (see Table 5). Green highlighting means the absence of mistakes in fall detection inside records.

b_1/b_2	0	1	2	3	4	5
0	0.981	0.987	0.987	0.987	0.987	0.987
-1	0.987	0.987	0.987	0.987	0.987	0.986
-2	0.987	0.987	0.987	0.987	0.986	0.986
-3	0.987	0.987	0.987	0.986	0.986	0.986
-4	0.987	0.987	0.986	0.986	0.986	0.985
-5	0.987	0.986	0.986	0.986	0.985	0.985

Table 4. Persons 1-10 cumulative sums procedure parameters tuning for coincidence in position and duration of fall segments determined by the procedure and labeled by experts

b_1/b_2	0	1	2	3	4	5
0	0.855	0.881	0.878	0.882	0.878	0.878
-1	0.881	0.878	**0.882**	0.878	0.878	0.880
-2	0.878	0.882	0.878	0.878	0.880	0.878
-3	0.882	0.878	0.878	0.880	0.878	0.870
-4	0.878	0.878	0.880	0.878	0.870	0.870
-5	0.878	0.880	0.878	0.870	0.870	0.868

Table 5. Coincidence in position and duration of fall segments determined by the procedure and labeled by experts for Person ID=11

Example of fall detection record of tested person Data11\Fall\side\3 presented in Fig. 8. Red graph – expert labeling, blue curve – SVM output as distance to optimal separating hyperplane (fall is positive), green graph – cumulative sums procedure output, orange – frames with outliers.

The TST Fall Detection Dataset v2 is rather recent and just a few publications reported the results on it. Unfortunately, researchers use the different experimental protocol in their studies and direct comparison gives a flimsy view of their quality. Nevertheless, considering, that the most of experimental evaluations use different modifications of cross-validation scheme, we provide a Table 6 with fall detection accuracies of several methods, tested on TST Fall Detection Dataset v2.

For our approach experimental evaluation, we apply probably the most fastidious protocol of Leave-One-Person-Out, described above.

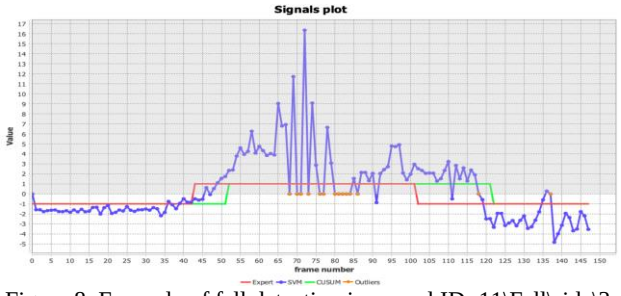

Figure 8. Example of fall detection in record ID=11\Fall\side\3, red – expert labeling (+1 – FALL, -1 – ADL), blue – SVM output as distance to optimal separating hyperplane (FALL is positive), green – cumulative sums procedure output, orange – frames with outliers.

For example, the best result for the Leave-one-Person-out procedure for the person with ID=11 contains just one mistake – one ADL record (lay) assumed to FALL class. So, the accuracy of classification for the whole record (is FALL present inside record or not) is 0.958. The accuracy of coincidence in position and duration of fall segments determined by the procedure and labeled by experts is 0.882.

Method	Source data	Classifier	Evaluation Scheme	Accuracy
(Gasparrini et al., 2016)	Skeleton joint position; accelerometer data	Empirical thresholding rule	Not described	0.99
(Fakhrulddin et al., 2018)	Two accelerometers time series data	CNN	Random 90% and 10% splitting and averaging	0.923
(Hwang et al., 2017)	Depth map	3D-CNN + data augmentation	5 random trials from 240 and 24 records splitting and averaging	0.942
(Min et al., 2018)	Skeleton joints information	SVM	Not clear	0.9205
Our	Skeleton joints information	SVM + One-class classifier + CUSUM	Leave-One-Person-Out	0.917

Table 6. The accuracy of fall detection algorithms on TST Fall Detection Dataset v2

The full exhaustive search with consistent exception of each person from database shows accuracy of classification for the whole record as 0.917. The accuracy of coincidence in position and duration of fall segments determined by the procedure and labeled by experts is 0.879. The average delay of fall start position determined by suggested classifier on test records is 6.52 frames.

7. CONCLUSION

Experimental study shows that the proposed skeleton features in combination with SVM classifier and CUSUM procedure, can provide reliable and accurate fall detention in real time. Since Microsoft Kinect v2 has been investigated for quite a long time, there are a few open datasets to perform an evaluation of fall defection algorithms based on its data. Moreover, the dataset we used for experiments required the additional manual labeling. For accurate estimation of productivity of our fall detection method, a special Leave-One-Person-Out was proposed.

Despite the stronger testing conditions, the proposed method shows its the competitive equality with methods using additional accelerometer data. In addition, we tested not only the fact of detecting a fall in the entire record but the position of the beginning of the fall and obtain rather small average delay.

ACKNOWLEDGEMENTS

This work was supported by the Russian Fund for Basic Research, under Grant Nos. 16-57-52042, 18-07-00942 and Ministry of Science and Technology, Taiwan, under Grant Nos. MOST 105-2923-E-027-001-MY3.

REFERENCES

A. Mundher, Z., Jiaofei, Z., 2014. A Real-Time Fall Detection System in Elderly Care Using Mobile Robot and Kinect Sensor. Int. J. Mater. Mech. Manuf. 2, 133–138. doi:10.7763/IJMMM.2014.V2.115

Bevilacqua, V., Nuzzolese, N., Barone, D., Pantaleo, M., Suma, M., D’Ambruoso, D., Volpe, A., Loconsole, C., Stroppa, F., 2014. Fall detection in indoor environment with kinect sensor. 2014 IEEE Int. Symp. Innov. Intell. Syst. Appl. Proc. 319–324. doi:10.1109/INISTA.2014.6873638

Bian, Z.-P., Hou, J., Chau, L.-P., Magnenat-Thalmann, N., 2015. Fall Detection Based on Body Part Tracking Using a Depth Camera. IEEE J. Biomed. Heal. Informatics 19, 430–439. doi:10.1109/JBHI.2014.2319372

Cai, Z., Han, J., Liu, L., Shao, L., 2017. RGB-D datasets using microsoft kinect or similar sensors: a survey. Multimed. Tools Appl. 76, 4313–4355. doi:10.1007/s11042-016-3374-6

Cippitelli, E., Gasparrini, S., Gambi, E., Spinsante, S., 2016. A Human Activity Recognition System Using Skeleton Data from RGBD Sensors. Comput. Intell. Neurosci. 2016. doi:10.1155/2016/4351435

Demiris, G., Oliver, D.P., Giger, J., Skubic, M., Rantz, M., 2009. Older adults’ privacy considerations for vision based recognition methods of eldercare applications. Technol. Heal. Care 17, 41–48. doi:10.3233/THC-2009-0530

Dvoenko, S.D.D., Kopylov, A.V. V., Mottl, V.V. V., Dvoenko, S.D.D., Kopylov, A.V. V., 2004. The Problem of Pattern Recognition in Arrays of Interconnected Objects. Statement of the Recognition Problem and Basic Assumptions. Autom. Remote Control 65, 127–141. doi:10.1023/B:AURC.0000011696.31008.5a

Fakhrulddin, A.H., Fei, X., Li, H., 2018. Convolutional neural networks (CNN) based human fall detection on Body Sensor Networks (BSN) sensor data. 2017 4th Int. Conf. Syst. Informatics, ICSAI 2017 2018–Janua, 1461–1465. doi:10.1109/ICSAI.2017.8248516

Gasparrini, S., Cippitelli, E., Gambi, E., Spinsante, S., Wåhslén, J., Orhan, I., Lindh, T., 2016. Proposal and Experimental Evaluation of Fall Detection Solution Based on Wearable and Depth Data Fusion, in: Loshkovska, S., S., K. (Eds.), ICT Innovations 2015. Advances in Intelligent Systems and Computing. Springer Cham, pp. 99–108. doi:10.1007/978-3-319-25733-4_11

Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M., 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. IJCAI Int. Jt. Conf. Artif. Intell. 2466–2472.

Hwang, S., Ahn, D., Park, H., Park, T., 2017. Maximizing Accuracy of Fall Detection and Alert Systems Based on 3D Convolutional Neural Network. Proc. Second Int. Conf. Internet-of-Things Des. Implement. - IoTDI ’17 343–344. doi:10.1145/3054977.3057314

Kepski, M., Kwolek, B., 2012. Fall Detection on Embedded Platform Using Kinect and Wireless Accelerometer 407–414. doi:10.1007/978-3-642-31534-3_60

Leone, A., Diraco, G., Siciliano, P., 2011. Detecting falls with 3D range camera in ambient assisted living applications: A preliminary study. Med. Eng. Phys. 33, 770–781. doi:10.1016/j.medengphy.2011.02.001

Mastorakis, G., Makris, D., 2012. Fall detection system using Kinect’s infrared sensor. J. Real-Time Image Process. 9, 635–646. doi:10.1007/s11554-012-0246-9

Min, W., Yao, L., Lin, Z., Liu, L., 2018. Support vector machine approach to fall recognition based on simplified expression of human skeleton action and fast detection of start key frame using torso angle. IET Comput. Vis. doi:10.1049/iet-cvi.2018.5324

Page, E.S., 1954. Continuous inspection schemes. Biometrika 41, 100–115.

Pathak, D., Bhosale, V.K., 2017. Fall Detection for Elderly People in Homes using Kinect Sensor. Int. J. Innov. Res. Comput. Commun. Eng. 5, 1468–1474. doi:10.15680/IJIRCCE.2017

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, a J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471. doi:10.1162/089976601750264965

Vemulapalli, R., Arrate, F., Chellappa, R., 2014. Human action recognition by representing 3D skeletons as points in a lie group. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 588–595. doi:10.1109/CVPR.2014.82

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1290–1297. doi:10.1109/CVPR.2012.6247813

Wild, K., Boise, L., Lundell, J., Foucek, A., 2008. Unobtrusive In-Home Monitoring of Cognitive and Physical Health: Reactions and Perceptions of Older Adults. J Appl Gerontol. 27, 181–200. doi:10.1080/10810730902873927.Testing

Xu, T., Zhou, Y., Zhu, J., 2018. New Advances and Challenges of Fall Detection Systems: A Survey. Appl. Sci. 8, 418. doi:10.3390/app8030418

Zhang, Z., 2012. Microsoft kinect sensor and its effect. IEEE Multimed. 19, 4–10. doi:10.1109/MMUL.2012.24