A Simple Indoor Fall Control System for the Elderly Based on the Analysis of Object Bounding Box Parameters

Katsiaryna Kosarava
Dept. of System Programming and Computer Security
Yanka Kupala State University of Grodno
Grodno, Belarus
koluzaeva@gmail.com
ORCID 0000-0001-7326-5307

Boris Assanovich
Dept. of Information Systems and Technologies
Yanka Kupala State University of Grodno
Grodno, Belarus
bas@grsu.by

Abstract. **In this paper a human fall detection problem using video recordings features such as bounding box height to width ratio and speed of the bounding box movement is considered. We examine two datasets, the first of which, in addition to video recordings, also contains accelerometer readings. Support vector machine, decision tree and random forest were used as classification models. These models were built on secondary features generated with the use of the tsfresh library. The experimental results showed that the analized datasets are almost linearly separable according to some newly generated features. Thus, the use of the tsfresh library allows to apply simpler models and at the same time to provide a higher classification accuracy up to 1.0 in comparison with more complex LSTM models.**

Keywords: **human fall detection, machine learning, classification models, tsfresh, LSTM**

I. INTRODUCTION

In recent years, much attention has been paid to monitoring the activity of the elderly. One of the biggest problems of people over 70 consists in losing their balance and falling. Falls are especially dangerous for people who live alone, which reduces the ability to provide quick assistance, Giannakouris et al. [1]. This problem has led to an increase in the number of studies on remote detection of falls, allowing fast and qualified assistance to the elderly. Such systems often use sensor devices to register falls based on the collected data. Approaches, based on wearable sensors or mobile phones, require wearing a monitoring device. The problem here is that old people often forget to wear such devices. Alternative systems based on computer vision methods do not require the participation of the elderly [2]. In addition, today video cameras are becoming cheaper and do not require complex manipulations for their installation.

Kwolek and Kepski [3] collected data from a 3D accelerometer and image depth maps to detect falls. The measured acceleration value was exploited to control the defined threshold to extract a number of human characteristics and run the classifier to determine the presence or absence of a fall. Redmon et al. [4] studed a convolutional neural network (CNN) to analyze images, extracted from a video and get features using an optical flow method to detect motion between two consecutive video frames. Several cameras or additional devices can be exploited to detect movement, as was done for UP-Fall Detection dataset. The disadvantage in the considered cases is the use of a special camera and/or several cameras, as well as resource-intensive structures to obtain the required accuracy. Recently Lezzar et al [5] proposed a simple algorithm for determining fall, and common daily activities with the use of computer vision and deep learning, based on a 2D camera with occlusion recognition and control of the transition between states before and after a fall. By increasing the controlled states, the authors claim that they achieved an accuracy of 93.94% and recall 100% with SVM classifier.

In this study we describe a simple application of both linear (SVM, random forest, etc.) and nonlinear classifiers (LSTM) for detecting a person's fall, and also propose a transition to new features for time series analysis [6]. The article is organized as follows. The next section describes the algorithm for controlling a person's fall and the structure of the system. Further, experiments using various classifiers are presented and their effectiveness is compared. The last section contains the conclusion and prospects for the development of the methodology proposed.

II. FALL DETECTION MODELS BASED ON VIDEO AND ACCELEROMETER RECORDS

A. Dataset description

In this section, we compare the performance of fall detection models trained on a dataset [3]. This dataset contains 70 (30 falls + 40 activities of daily living (adl)) sequences. Fall events are recorded with 2

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

Microsoft Kinect cameras and corresponding accelerometric data. ADL events are recorded with only one device (camera 0) and accelerometer. Sensor data was collected using PS Move (60Hz) and x-IMU (256Hz) devices.

Dataset containes synchronized video camera records: frame number, time in milliseconds since sequence start and interpolated accelerometric data, corresponding to image frame. The cameras are recorded independently, so they are not strictly synchronized (synchronization based on nearest timestamp value). Raw accelerometric data contains time in milliseconds since sequence start and accelerometer data: SV_{total}, A_x, A_y, A_z. All accelerometer data are in gravity units (g). Total sum vector is calculated as follows: $SV_{total} = \sqrt{A_x^2 + A_y^2 + A_z^2}$.

The dataset also contains files with extracted features from depth maps stored in CSV format. Each row is one sample of data corresponding to one depth image with following features: *sequence name* ('fall-01','adl-01' etc.), *frame number*, *label* - '-1' means person is not lying, '1' means person is lying on the ground; '0' is temporary pose, when person "is falling", *HeightWidthRatio* - bounding box height to width ratio, *MajorMinorRatio* - major to minor axis ratio, computed from BLOB of segmented person, *BoundingBoxOccupancy* - ratio of how bounding box is occupied by person's pixels, *MaxStdXZ* - standard deviation of pixels from the centroid for the abscissa (X axis) and the depth (Z axis), respectively, *HHmaxRatio* - human height in frame to human height while standing ratio, *H* - actual height (in mm), *D* - distance of person center to the floor (in mm), *P*40 - ratio of the number of the point clouds belonging to the cuboid of 40 cm height and placed on the floor to the number of the point clouds belonging to the cuboid of height equal to person's height.

B. Experimental setup

In [3], the authors came to the conclusion that the fall detectors using only inertial sensors generate too much false alarms. This means that some daily activities are erroneously signaled as fall, which in turn leads to frustration of the users. Therefore, the authors train the SVM model to classify the position of a person in the frame (person is lyings / is'n lying) based on the extracted features of the depth maps to confirm the SV_{total} accelerometer indicators.

In this paper, we explore the possibility of combining accelerometer records and features extracted from depth maps.

The first approach to training the model was based on the assumption: if a person in the frame is in a horizontal position (i.e., person is lying), it is necessary to determine whether this is a consequence of a fall or not. To do this, we select frames from the video records on which a person is lying according to the *label* attribute, and the corresponding values of the remaining attributes. From the accelerometer records we take the SV_{total} value corresponding to the numbers of the selected frames for each video recording. We took only *HeightWidthRatio* from all the features of depth maps, since it is this indicator that allows us to determine the position of a person in the frame (lies / does not lie). The described sequence of actions was applied separately to the datasets for fall and for adl. Thus, we got a labeled dataset consisting of the following fields: SV_{total}, *HeightWidthRatio*, *label* (-1: not fall / 1: fall). After balancing the data, we got 903 records for each class. Visual analysis showed that with this approach to data generation, the classes are strongly mixed with each other. This is also confirmed by the modeling results: the best trained SVM model with 'rbf' kernel showed an accuracy of only 75% on 10-fold cross-validation.

The second approach to analyzing the dataset was to consider each feature of depth maps and accelerometer readings in the sequence in which they were measured, i.e. like time series. And construct new significant features for these time series. For this purpose, the library **tsfresh** [7] was used. It allows to extract over 1200 features from time series. Tsfreash contains a feature selection method that evaluates the importance of the different extracted features. To do so, for every feature the influence on the target is evaluated by an univariate tests. Those tests generate p-values that are then evaluated by the Benjamini Hochberg procedure to decide which features to keep and which to delete [8]. It was visually determined that the *HhmaxRatio*, *H*, *D*, *P*40 signs are practically identical for the two classes, so they were removed. For the remaining features Sv_{total}, *HeightWidthRatio*, *MajorMinorRatio*, *BoundingBoxOccupancy* and *MaxStdXZ*, 824 new significant features were generated using tsfresh. Thus, we got a dataset with a dimension of 70 by 824 (70 is the number of videos, 824 is the number of new significant features). On these data, models of SVM, decission tree and random forest were trained, with the selection of parameters according to 10-fold cross-validation, 1st Model Table I. The best result was shown by the SVM model with a 'linear' kernel (based on normalized data), the numbers of support vectors are 9 and 7 for two classes, respectively, 1st Model Table I. Moreover, the Decision Tree model almost accurately classified the data using only two features: the average, absolute value of consecutive changes of the series SV_{total} inside quantile corridor [0.0, 1.0] and the absolute value of zero coefficient of the one-dimensional discrete Fourier Transform for *MaxStdXZ* by fast fourier transformation algorithm [10].

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

If we leave only the SV_{total} and *HeightWidthRatio* features and generate new significant features with tsfresh, then we get data with dimensions 70 by 434. The best model was 'linear' SVM model, the numbers of support vectors are 8 and 7 for two classes, respectively, 2nd Model Table I.

TABLE I. EXPERIMENT RESULTS for FALL DATASET [3]

Trained Models	**Models Accuracy**			
	1st Model		**2nd Model**	
	Train acc.	Test acc.	Train acc.	Test acc.
SVM, 'linear' n_support	1.0 [9, 7]	1.0	1.0 [8, 7]	1.0
SVM, 'rbf' n_support	0.86 [17,16]	1.0	1.0 [16, 12]	1.0
Decision Tree	0.98	1.0	0.98	0.86
Random Forest	0.91	1.0	1.0	0.93

III. FALL DETECTION MODELS BASED ON VIDEO RECORDS

A. Dataset description and preprocessing

Obviously, using only video cameras is a more convenient and less expensive way of monitoring. As the second fall detection dataset we take one collected by the Imaging and Artificial Vision laboratory (ImViA) [9]. ImViA has been developing an architectures for specific vision and visualization systems and new applications in medical imaging unconventional imagery for robotics and scene reconstruction and analysis. ImViA fall detection dataset was collected in realistic video surveillance setting using a single camera. The frame rate is 25 frames/s and the resolution is 320×240 pixels. Video sequences contain variable illumination, and typical difficulties like occlusions or cluttered and textured background. The actors performed various normal daily activities and falls. The dataset contains 191 videos that were annotated, for evaluation purpose, with extra information representing the ground-truth of the fall position in the image sequence. Each frame of each video is annotated: the localization of the body is manually defined using bounding boxes. This annotation allows to evaluate the classification features independently from the automatic body detection.

Typically, they use several available datasets for fall detection, use the same location for detection and training. Thus, it does not allow the use of methods to change the position between the training data and the data for testing. To gauge this reliability, the dataset contains videos from different locations that allow multiple assessment protocols to be defined (Home, Coffee Shop, Office, and Lecture Hall).

It was visually determined that most of the videos are around 300 frames in length. Therefore, it was decided to select data with a length of no more than 350 frames and bring all records to a length of 300 values. Longer recordings were truncated to a size of 300, and shorter ones were supplemented with a repetition of the twentieth value from the end, to preserve the dynamics in the recording values.

It should be noted that out of 191 videos, only 112 were annotated and only 87 passed the validation test. Out of 87 records, 13 records are "without falling" records, the rest are records "with falling". After data transformations, there were 11 records "without falling". The number of records "with falling" and "without falling" had to be balanced, and after that there were 23 records left for training.

In annotation records, the first two lines indicate the presence or absence of a drop in the record. Thus, if the first number in the record is zero, then the record refers to records without falling. In other cases, the fall is present.

After determining the label of the record and bringing to a single size, all record values complement the object containing the values of all records. After performing labelling of all records, the object with all values is exported to a file for further work with data.

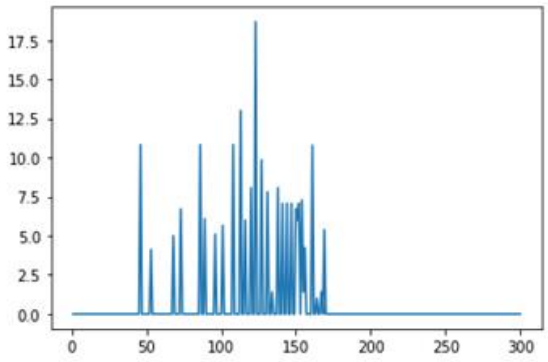

Fig. 1. The speed of the frame movement for each frame

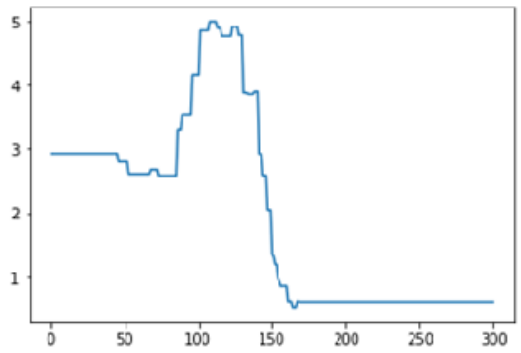

Fig. 2. *HeightWidthRatio* values interpolation

Further, for each record and each frame, two new features were calculated: *HeightWidthRatio* - the value of the aspect ratio of the bounding box,

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

converted from the lengths of the frame sides, and *BoundingBoxSpeed* - the speed of the frame movement for each frame, converted from the frame coordinates, Fig. 1. *HeightWidthRatio* values greater than 10 were replaced by 0 and later have been reconstructed using interpolation, Fig. 2.

B. Model setup

First, experiments were carried out using the LSTM neural network. The length of the LSTM input sequence was 300 elements, training was carried out on selected data (as described above) using the Adam optimization method, with control of the loss function based on binary cross-entropy. As a result, a trained neural network based on cross-validation data using batch normalization gave an average accuracy of 0.83.

Next, we used the second approach, described in the previous paragraph, to generate new significant features for the *HeightWidthRatio* and *BoundingBoxSpeed* time series using the tsfresh library. As a result, a dataset of 23 by 64 dimensions was obtained, where 23 is the number of records, 64 is the number of features, 1st Model Table II. Both 'linear' and 'rbf' SVM achieved train and test accuracy 1.0, but a 'linear' classifier is preferable, since the number of support vectors in both classes equals 1, which indicates a good generalizing ability of this model. The decision tree unmistakably classified the data according to only one feature X_0 - means the average, absolute value of consecutive changes of the series *HightWidthRatio* inside the quantile corridor [0.2, 0.8], Fig.3. Further, leaving only the X_0 feature, similar results were obtained, 2nd Model Table II.

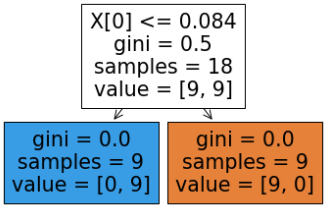

Fig. 3. Decision Tree classifier

TABLE II. EXPERIMENT RESULTS FOR FALL DATASET [9]

Trained Models	**Models Accuracy**			
	1st Model		**2nd Model**	
	Train acc.	Test acc.	Train acc.	Test acc.
SVM, 'linear' n_support	1.0 [2, 3]	1.0	1.0 [1, 1]	1.0
SVM, 'rbf' n_support	0.86 [5,3]	1.0	1.0 [3, 1]	1.0
Decision Tree	1.0	1.0	1.0	1.0
Random Forest	1.0	1.0	1.0	1.0

Analysis of more complex algorithms using the detection of key points of a person and training classifiers to fix the fall of the elderly [11] has not been carried out, since this technique requires rather high computational costs when processing video recordings.

IV. CONCLUSION

A method for controlling the fall of the elderly with the application of computer vision procedures has been proposed. The use of neural networks (LSTM structure) for the fall determination allowed to achieve 83% of the average classification accuracy.

The result obtained can be mainly explained by a rather small amount of the training data, remaining after the preparation phase. The approach to the fall control, using the full set of secondary parameters obtained with tsfresh library, allows to perform a linear separation of classes and achieve accuracy and precision close to 100%. The application of the tsfresh-based technology with only 2 secondary features associated with the ratio of bounding box parts, covering the person, and its movement speed allows to separate classes linearly with an accuracy of more than 93%. The efficiency of the tsfresh-based technology was confirmed on samples from 2 datasets: UP-Fall detection dataset and ImViA fall detection dataset. It is obvious that the method proposed has a lower computational complexity and achieves no less accuracy in comparison with known approaches. Further it can be improved by increasing the number of video cameras and/or additional sensors for the multimodal detection. It should also be noted that the analysis was performed on the abbreviated, normalized and balanced part of the ImVia dataset.

ACKNOWLEDGMENT

The work was partially supported by the COST Action CA16226 project. The authors thank Maxim Stupaktvich for participating in computational experiments.

REFERENCES

[1] K. Giannakouris, et all. "Ageing Characterises the Demographic Perspectives of the European Societies", EUROSTAT Statistics in focus, 2008. Available from: http://ec.europa.eu/eurostat.

[2] A. Ramachandran, R. Adarsh, P. Pahwa and K. Anupama, "Machine Learning-Based Fall Detection in Geriatric Healthcare Systems", 2018, IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Indore, India, 2018, pp. 1-6.

[3] B. Kwolek, M. Kepski, "Human Fall Detection on Embedded Platform Using Depth Maps and Wireless Accelerometer", Computer Methods and Programs in Biomedicine, 2014, 117(3), pp. 489-501.

[4] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection",

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779-788.

[5] F. Lezzar, D. Benmerzoug, I. Kitouni, "Camera-Based Fall Detection System for the Elderly With Occlusion Recognition", Applied Medical Informatics, Vol. 42, 2020, pp. 169-179.

[6] H. Ponce, L. Martínez-Villaseñor, J. Brieva, and E. Moya-Albor, "Challenges and Trends in Multimodal Fall Detection for Healthcare": Springer, 2020, 259 p.

[7] tsfresh,https://tsfresh.readthedocs.io/en/latest/text/introduction.html, Accessed 21 Jun 2021.

[8] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing", Journal of the Royal Statistical Society: Series B, 57, pp. 289–300.

[9] F. Franchetti and M. Püschel, "FFT (Fast Fourier Transform)", In: D. Padua (eds) Encyclopedia of Parallel Computing: Springer, Boston, MA, 2011, pp. 658-671.

[10] "Fall detection dataset", ImViA, https://imvia.u-bourgogne.fr/en/database/fall-detection-dataset-2.html, Accessed 21 Jun 2021.

[11] O. Seredin, A. Kopylov and E. Surkov, "The study of skeleton description reduction in the human fall-detection task", Computer Optics, Vol. 44, 2020, pp. 951-958.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.