A New Large-scale Video Dataset for Human Fall Detection

Anahita Shojaei-Hashemi
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC
anahitas@ece.ubc.ca

Panos Nasiopoulos
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC
panosn@ece.ubc.ca

Mahsa T. Pourazad
TELUS Communications Inc.
Vancouver, BC
mahsa.pourazad@gmail.com

Abstract

Falling is a common cause of domestic injuries with the potential of fatality. It is also a major obstacle against independent living and aging in place. A system to automatically detect fall incidents in residences can alleviate these problems. Video cameras have been proven to be efficient in this regard. However, lack of data is the bottleneck in training video-based models, especially when deep learning is employed. The public large datasets that are currently available on human fall have shortcomings in simulating real life scenarios, which adversely affect the performance of the trained models in practice. To address these issues, we have captured a new dataset, which will be released soon. We have designed a deep-learning-based model and trained it on our dataset. The trained model has successfully passed the tests assessing various aspects of the dataset.

1 Introduction

Remote health and wellness monitoring of people at home is of great interest in the field of public healthcare [1]. It improves the occupants' quality of life, while significantly decreases the cost imposed on healthcare systems. An important element of a remote health monitoring system is fall detection. Falls are the cause of a major part of domestic injuries. They may also be the sequel to a serious health condition, such as heart attack or stroke. While immediate treatment after a fall can significantly decrease its adverse health impacts, if the incident is not discovered right away, it can result in critical health complications and even death. Therefore, an automatic fall detection system is of great importance in the wellbeing and independence of the residents and in supporting aging in place.

Different means can be used to detect falls, including wearable devices, ambient sensors, and video cameras. Wearable devices constantly measure changes in the height, orientation, and velocity [2] of the person. As the person needs to always wear the device, it is inconvenient and intrusive. Ambient

NeurIPS 2020 Workshop on Dataset Curation and Security

sensors integrated into mats, mattress, or couch do not have this problem. However, they are not sufficiently accurate and have high false-positive rates. They are not easily affordable, either, since plenty of them are required for a reasonable coverage of the residence. Like ambient sensors, video cameras are not intrusive. Affordable cameras with reasonable resolutions contain a great amount of information. This makes designing high accuracy fall detection models feasible. It also enables identifying the occupants to fine-tune the model for better performance and to provide personalized healthcare. These features make video cameras a popular means of fall detection.

After years of using classic machine learning and handcrafted features for video-based fall detection [3, 4, 5, 6, 7, 8, 9], researchers started to develop deep-learning-based models three years ago, and it quickly turned into the dominant approach in the field [10, 11, 12]. However, it requires a huge amount of data, which is not easy to obtain. Publicly available real-life footages of fall incidents are scarce, so simulated data is left as the only feasible alternative. As performing the act of falling in a natural way while sticking to safety measures is difficult, most of the simulated public datasets are short on falls. To the best of our knowledge there are only two public video datasets that contain relatively large numbers of falls: NTU RGB+D Action Recognition Dataset [13] and SDU Fall Dataset [14]. NTU RGB+D dataset is composed of 56,880 videos of 60 actions, including 948 videos of falling. Each action is performed twice by 40 subjects and is captured with three synchronous Microsoft Kinect v2 cameras in RGB, depth map, and infrared (IR) modalities. Each video contains one action and has the average length of two seconds. There is an extension to this dataset called NTU RGB+D 120 Action Recognition Dataset [15]. It consists of 114,480 videos of 120 actions, where the 60 new actions are performed by 106 subjects. SDU dataset is comprised of 1200 videos of six actions, including falling. 20 subjects perform each action ten times, and one Kinect camera records the scene in RGB and depth map modalities. Like NTU RGB+D dataset, each video contains one action with the average length of two seconds.

Although these datasets, especially SDU, are used to benchmark deep learning approaches for fall detection, they have two major downsides. They consist of very short videos, i.e. less than 10 seconds, with each video encompassing only one action, whereas in practice, there is a video stream of consecutive actions. Another drawback is the lack of diversity in falling scenarios. These datasets include only one scenario, where a sudden falling happens almost immediately after the subject enters the scene. In real life, however, a fall incident can happen in different ways, depending on for instance the cause. To address these issues, we have created a large dataset of long videos for human fall detection. It contains plenty of imitated fall events, covering various real-life scenarios.

To evaluate our dataset, we have proposed a deep learning model for human fall detection, which is composed of a two-dimensional (2D) convolutional neural network (CNN) and a long short-term memory (LSTM) neural network. We have trained the model on our dataset and have tested it both offline and in real time. Based on the results, our model has high performance in practice and under conditions close to real life.

This paper is structured as follows. Section 2 explains the proposed model, and section 3 introduces the dataset in detail. The evaluation procedure and results are discussed in section 4, and section 5 concludes the paper.

Table 1: Actions and corresponding labels

Label	Description
1	Walking, Standing, Picking from floor, etc.
2	Falling (Beginning of Downwards Motion)
3	Fallen (Second Part of Body Touches Ground)
4	Recovering (Beginning to Get Up)
5	Recovered (On Two Feet)
6	Sitting, Doing sit-ups, etc.

Table 2: The CNN architecture

Convolution (32×3×3) ReLU, MaxPooling (2×2)
Convolution (64×3×3) ReLU, MaxPooling (2×2)
Convolution (128×3×3) ReLU
Fully Connected (128) ReLU
Fully Connected (6) Softmax

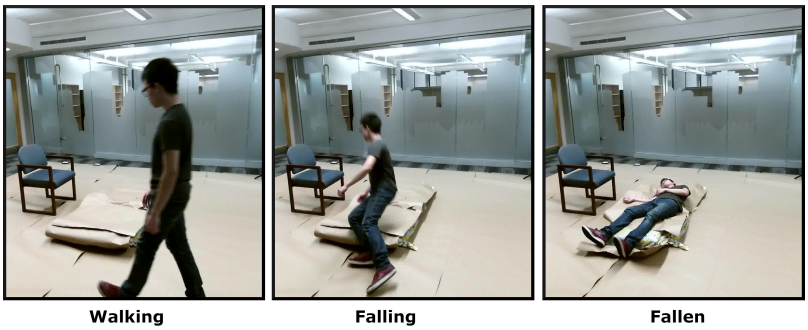

Figure 1: Sample labeled video frames

2 Proposed Model

Our proposed method detects fall incidents through RGB video, using Long-Term Recurrent Convolutional Network (LRCN) [16]. The model consists of an LSTM neural network on top of a 2D-CNN to address both the temporal and spatial aspects of video. At each time step, the current frame together with a window of preceding frames are given to the model. The CNN takes each frame and encodes its spatial information into a vector of features. The feature vectors of the frames within the window are fed to the LSTM network, and it determines whether a falling is taking place, or another action is going on.
The CNN is pre-trained using the videos of different actions, so that it can classify each video frame into one of the six classes of actions listed in Table 1. We have employed a custom CNN, the structure of which is described by 2.The last layer of the CNN, which is the classification layer of the CNN, is removed when the LSTM neural network is attached, so that the spatial features, i.e. the outputs of the layer before the last, can directly flow into the LSTM network. The CNN parameters are kept frozen during the training of the LSTM network, which is composed of a single layer with 512 hidden units and has six classes.

Table 3: Captured scenarios

No.	Scenario	Description	Actions
1	Sitting	The subject sits on the floor, mat, or a chair.	6
2	Sitting Exercises	The subject does sit ups, stretches on the floor, or does another seated or lying exercise.	6
3	Sitting to Fall	The subject sits on a chair and falls to the mat.	6 ,2, 3
4	Standing Exercises	The subject does jumping jacks, stretches, or other standing exercises.	1
5	Walking	The subject walks around the room, around the mat, or over the mat.	1
6	Cleaning	The subject mimics cleaning, using a broom, mop, or handle.	1
7	Dropping and Picking Up an Item	The subject drops an item, e.g. a pen, bends down, picks it up, and goes back to standing position.	1
8	Forward Fall, Side Fall, Backward Fall	The subject falls on the mat. To have a variety of falls, different points of the body, e.g. knees, side, back, hands, or arms, contact the floor first. The falls are also done at varying speeds.	2, 3
9	Falling and Recovering	The subject recovers after falling. It is done slowly.	2, 3, 4, 5
10	Heart Attack to Fall	The subject clutches their chest, mimicking a heart attack, and then falls.	1, 2, 3
11	Stroke to Fall	The subject touches their head, mimicking a stroke, and then falls.	1, 2, 3
12	Trip to Fall	The subject mimics tripping over an object, and then falls.	1, 2, 3

Table 4: Model performance at the threshold of 0.5

TPR (Recall)	FPR	Precision	Accuracy
0.73	0.08	0.81	0.85

3 Data Capturing and Labeling

The data capturing was fulfilled throughout several sessions, where RGB videos were recorded at 30 frames per second using four Kinect V2 cameras. Each camera captured the scene from a different viewpoint. This helps prevent bias in the dataset and creates more samples in one shot. It also addresses the needs of multi-view approaches. There were 20 participants, performing 12 scenarios demonstrated by Table 3. Each scenario is composed of one or a combination of the actions listed in Table 1. Around 400 videos were captured, in total. The captured videos are labeled frame by frame. Each frame is assigned a label between 1 and 6 according to Table 1. Label 1 represents actions performed while standing on two feet. Labels 2 to 5 include actions taking place in the course of falling to recovering. Finally, label 6 contains actions executed while sitting or lying on the floor. Figure 1 depicts a few frames of a sample captured video with their corresponding labels.

4 Experiments and Results

The captured videos were split into a training and a test set with the ratio of 75 to 25. The two modules of the model, i.e. the CNN and the LSTM neural network, were trained separately. First, the CNN was trained for action recognition at frame level, so that it could discriminate the six categories of actions. With dropout and L2 regularization to prevent overfitting, it achieved 79% accuracy. Then, all the video frames were encoded into feature vectors by the trained CNN, and these vectors were used to train the LSTM network. In this way, the CNN parameters were kept frozen while training the LSTM network. A sliding window of length 16 and stride 1 was employed to feed the data to the LSTM neural network. The ground truth label for each window was determined based on the labels of the frames within the last quarter of the window. Dropout and L2 regularization were applied in training the LSTM network, too.
In training both modules, weighted loss was used to compensate for the class imbalance at frame and window levels, where the weights were inversely proportional to the number of samples in the corresponding classes. Also, the dropout ratio and the regularizer λ were set as respectively 0.5 and 0.01. Table 4 shows the performance of the whole model on the test data.

5 Conclusion

In this paper we introduced a new large dataset of videos for human fall detection. The dataset consists of relatively long videos and encompasses various falling scenarios to address the needs of fall detection models for real-life application. We also proposed a deep neural network for video-based fall detection. It takes RGB videos and detects fall incidents in real time. We trained and tested the model on our dataset as a showcase. With 85% accuracy, 0.81 precision, and 0.74 recall, the model showed promising performance.

References

[1] V. Gruessner. The history of remote monitoring, telemedicine technology. `https://mhealthintelligence.com/news/the-history-of-remote-monitoring-telemedicine-technology`.

[2] Liming Chen and Ismail Khalil. Activity recognition: Approaches, practices and trends. In *Activity Recognition in Pervasive Intelligent Environments*, pages 1–31. Springer, 2011.

[3] Derek Anderson, James M Keller, Marjorie Skubic, Xi Chen, and Zhihai He. Recognizing falls from silhouettes. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6388–6391. IEEE, 2006.

[4] Nicolas Thome and Serge Miguet. A hhmm-based approach for robust fall detection. In *2006 9th International Conference on Control, Automation, Robotics and Vision*, pages 1–8. IEEE, 2006.

[5] Adam Williams, Deepak Ganesan, and Allen Hanson. Aging in place: fall detection and localization in a distributed smart camera network. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 892–901, 2007.

[6] Chia-Wen Lin and Zhi-Hong Ling. Automatic fall incident detection in compressed video for intelligent homecare. In *2007 16th International Conference on Computer Communications and Networks*, pages 1172–1177. IEEE, 2007.

[7] Xinguo Yu. Approaches and principles of fall detection for elderly and patient. In *HealthCom 2008-10th International Conference on e-health Networking, Applications and Services*, pages 42–47. IEEE, 2008.

[8] Hammadi Nait-Charif and Stephen J McKenna. Activity summarisation and fall detection in a supportive home environment. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 323–326. IEEE, 2004.

[9] Bart Jansen and Rudi Deklerck. Context aware inactivity recognition for visual fall detection. In *2006 Pervasive Health Conference and Workshops*, pages 1–4. IEEE, 2006.

[10] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Vision-based fall detection with convolutional neural networks. *Wireless communications and mobile computing*, 2017, 2017.

[11] Anahita Shojaei-Hashemi, Panos Nasiopoulos, James J Little, and Mahsa T Pourazad. Video-based human fall detection in smart homes using deep learning. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2018.

[12] Jacob Nogas, Shehroz S Khan, and Alex Mihailidis. Deepfall: Non-invasive fall detection with deep spatio-temporal convolutional autoencoders. *Journal of Healthcare Informatics Research*, 4(1):50–70, 2020.

[13] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[14] Xin Ma, Haibo Wang, Bingxia Xue, Mingang Zhou, Bing Ji, and Yibin Li. Depth-based human fall detection via shape features and improved extreme learning machine. *IEEE journal of biomedical and health informatics*, 18(6):1915–1922, 2014.

[15] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.