Fall Detection with a Single Commodity RGB Camera Based-on 2D Pose Estimation

Chatchai Wangwiwattana

Faculty of Computer Science

University of the Thai Chamber of Commerce, Thailand

Email: chatchai_wan@utcc.ac.th

Abstract

This study is a joint gesture estimation algorithm with LSTM to train the device to recognize a drop with a single non-modified RGB camera. Additionally, we try to detect as soon as the person falls. (Not when he / she is already lying on the ground) This work focuses on optimizing the modeling while maintaining computational resources for the real-time application.In older adults, falls can cause severe injuries. Even right before dropping, reaction time is important, so the near-by device can respond on time.

This article introduces a method for early detection of falls that is non-invasive. Our method only uses a single RGB Commodity Camera. With CNN-based, we estimate human-position, then use the pose to recognize falls without any calibration and additional equipment, processing join data and predict with LSTM model. The model achieves 97% in accuracy. It can be extended to a smart home. Information obtained testing the system, and many supports for the course of the study. Will benefit fall Detection with a Single RGB Camera.

Keywords : Computer Vision; Fall detection; human pose estimation

Introduction

Rise in the world's elderly population, systems aligned with elderly health care are rapidly in demand A potential alternative is to equip the smartness of the devices around us to act as a surveillance and assistance tool for the elderly. There is a great interest in fall detection and assistance platforms. For elderly people, falls are one of the most dangerous. Home C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau,2006 ; D. Wild, U. S. Nayak, and B. Isaacs,1981 , V. Scott, L. Wagar, and S. Elliott ,2010) is the leading cause of injury. Therefore, it is important to drop it as soon as possible. Fifty percent of those who do not seek treatment after the accident are more likely to die after six months of bronchopneumonia, dehydration, or hypothermia, even though they do not have physical damage. In addition, there is a high chance that a person with experience falling once would be 200% more likely to fall (Jennifer L O'Loughlin, Yvonne Robitaille, Jean-Francois Boivin, and Samy Suissa. 1993.) again. The Kellogg International Working Group on the Prevention of Falls in Elderly describes fall as "unintentionally coming to ground, or some lower level not as a consequence of sustaining a violent blow, loss of consciousness, sudden onset of paralysis as in stroke or an epileptic seizure" (M J Gibson, R O Andres, T E Kennedy, L C Coppard, and Others. 1987.) Despite several characteristics of falls, one is a rapid decrease in the height of a person's head and a person's head about the ground plane for a longer period. Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. (2006), Xinguo Yu.(2008), Ildoonet, Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, Rached Tourki, Ulrich Lindemann, A Hock, M Stuber, W Keck, Clemens Becker, Caroline Rougier, Jean Meunier, Alain St-Arnaud, Jacqueline Rousseau, V Scott, L Wagar, S Elliott, Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick, Adam Williams, Deepak Ganesan, Allen Hanson, Miao Yu, Adel Rhuma, Syed Mohsen Naqvi, Liang Wang, Jonathon Chambers, Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Xinguo Yu, V Vaidehi, Kirupa Ganapathy, K Mohan, A Aldrin, K Nirmal, Deidre Wild, U S Nayak, B Isaacs, M J Gibson, R O Andres, T E Kennedy, L C Coppard, Others, Caroline Rougier, Jean Meunier, Alain St-Arnaud, Jacqueline Rousseau, Lykele Hazelhoff, Jungong Han, Others, Markus D Solbach, John K Tsotsos, Tong Zhang, Jue Wang, Liang Xu, Ping Liu, Jennifer L O'Loughlin, Yvonne Robitaille, Jean-Francois Boivin, and Samy Suissa. (2008)

Falling tracking systems can loosely be divided into two groups: wearable systems and remote systems. Wearable devices use data stream accelerometers to track signal irregularities. You may simply set the threshold for vertical location and velocity. If these features hit a certain stage, it can be considered falls (Tong Zhang, Jue Wang, Liang Xu, and Ping Liu. (2006),Ulrich Lindemann, A Hock, M Stuber, W Keck, and Clemens Becker. (2005)) More advanced recognition algorithms use machine learning to identify falls based on dropping patterns in signals. At present, fall detection with wearable sensors is reliable. If the machine fails to detect a fall, users may manually trigger the feature. In case of no answer within pre-set time limits, the device may be automatically triggered. The downside to this strategy is that people always need to wear gadgets. They may experience pain or forget to wear it. Some systems require a periodic fee, and high maintenance cost such as periodically charge.

External fall detection systems make use of mounted equipment such as web cameras, CCTVs, or portable receiver devices. This category of systems solves the weakness that wearable systems have. Users do not expect the gadgets to be worn; thus, they are less burdened. When the device is mounted, it operates for everyone in the display field of the camera and does not have battery limitations. The drawbacks to this strategy are not stable compared to wearable devices. Occlusion and illumination variance are still important considerations for the robustness of the device. The cost of the installment may be high. In the case of a video-based device, it is of great significance regarding privacy in an undisclosed location, such as inside a bathroom. However, these drawbacks could be minimized by overtime. The cost of installment is lowered over time because high-resolution cameras and CCTV are becoming cheaper. Many households have already installed such home control devices. In the case of CCTV, it is more likely to be built with a light source that is adequate for most video-based algorithms. For privacy reasons, these data may be collected internally and processed within a chip. There is an increasing demand for devices capable of vision processing. Until then, the device implementer would enact data protection policies on data streaming, such as processing offline data within the unit, automatically blur face, encrypted data, and never transfer data to third parties. In addition, the framework could make it easier for users to manually manage their own data. However, the use of video monitoring is growing and more cost-effective over time. Drop detection system can be one of the many functions that this stream uses. Considering all the advantages, much of this weakness would be less of a problem in the future. Remote devices are ideal for hospitals, robotics, live residents, and smart houses. Our previous work presents a video-based framework for this analysis. The main novelty of this work is to attempt to track human fall with a single unmodified RGB camera. Since time to fall is important, we consider both velocity and location to detect fall before humans reach the floor. By taking advantage of the state-of-the-art deep learning 2D human pose which is robust agent lighting variance, occlusion, and camera angle (Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. (2017). The model performs up to 93% in accuracy (Chatchai Wangwiwattana, Engsak Sathienpaisal, Torsak Chaiwattanaphong, Puchpimon Singhapong, Oranuj Janrathitikarn, and Suwaroj Akrawutpornpat. 2019)

Objectives of Research

We extended our previous work by using a pose estimation algorithm in cooperated with LSTM to train a machine to recognize falling with a single unmodified RGB camera. In addition, we try to detect as soon as the person falls (not when he/she is already lying on the ground). This work focuses on increasing modeling performance while maintaining computational resources for real-time application.

Research Methodology

We split the process into three steps: pre-processing, processing, and model development. The summary of the algorithm is shown in Figure 1. When a frame is obtained from a camera, the size of the image has been scaled down to 320x244 pixels to reduce the computing cost of the pose estimation, while also being able to preserve acceptable precision. Then, we extract the positions of all joint locations in further processing through a deep learning-based pose estimator.

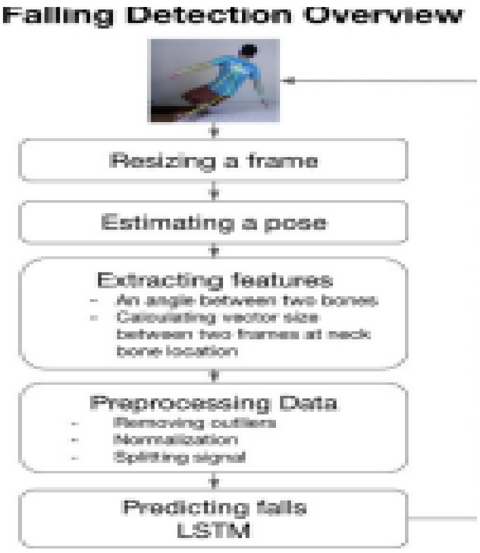


Figure 1: overview of the approach

1. Extracting Features of Falling Detection

Before training the model, we extracted and preprocess features from human pose key points. Therefore, we selected a two primary features for detecting falls as follows:

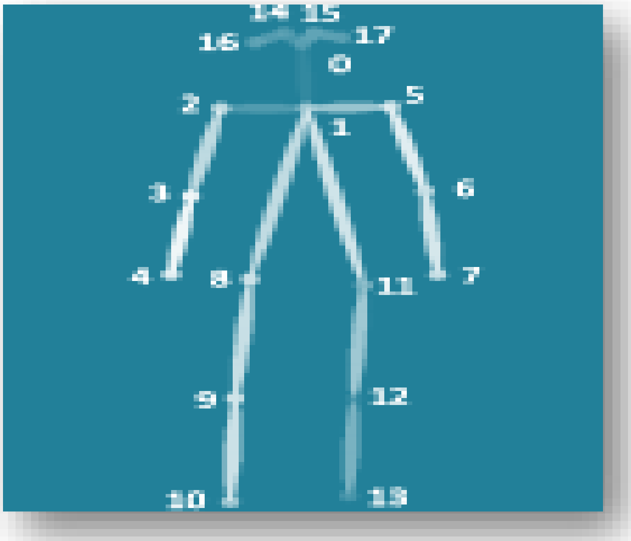


Figure 3 An image of pose key points

· An angle hip and neck key points (*angle*)

· A velocity size of vector of the neck (key point 1 in Figure 2) related to previous frame (*vf*)

The reason we select the neck as a reference point is that it is the most stable part of the body. The neck part is more likely to be in a camera frame, and the pose estimating algorithm will recognize the point in most poses from multiple camera angles and any occlusion. Second, there is less likely to be a shift in the different tasks involving hands and legs, such as gathering items. In addition, the location of the neck can shift as an individual moves as well as falls. This is critical for the fall detection. In addition to the neck key point, the hip points (key points 8th and 11th in Figure 2) are both excellent reference points for similar reasons. Therefore, we use them to measure velocity and angles.

A velocity vector is calculated with equation 1.

$$\vec{v} = \frac{(x_c - x_p, y_c - y_p)}{t_f}$$

Where x_c stands for x of current frame, x_p is x in previous frame. The same goes with y_c and y_p. Then it is divided by a frame time. This not only help to normalized frame time, but it is also able to estimate the velocity when the pose estimator failed to recognize poses in some frames. The last feature, the angle between two hip and neck key points

$$\Theta = \arccos(\frac{\vec{u} \cdot \vec{v}}{\|u\| \times \|v\|})$$

Then we remove noise and

Outliers by applying moving average with five-point window size. We only select frames around falling action (about 1-second video in Error! Reference source not found.) to machine learning models.

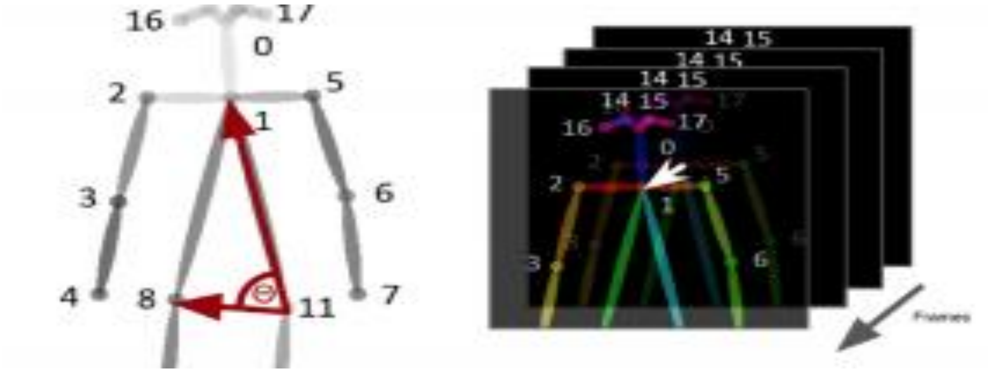


Figure 4 (Left) Hip and Neck Angle Vectors, (Right) Neck Key Points For Between Frame Velocity

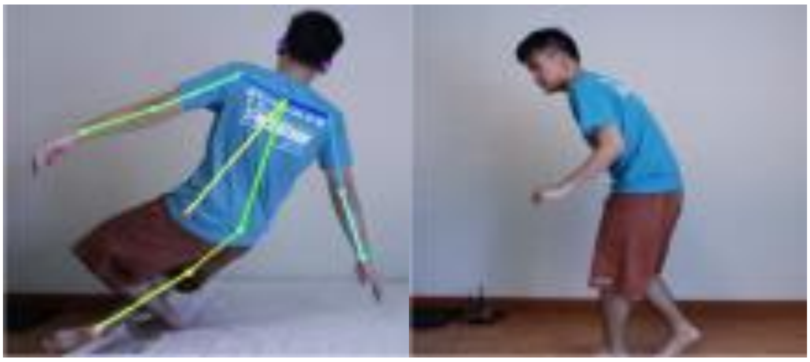


Figure 5 An example of an input frame

Our previous work shows that the vector size between the two frames (*vf*) and the angle between the neck key point and the hip key points (*angle*) are effective predictors as the vector size indicates the strength of the fall, and the angle between the neck and the hip key points gives additional details to the model [11]. In this study we use them together as a key feature set.

1.1. Training a Model

We extracted features from the previous section from our own dataset. The dataset consists of 20 videos of a fallen action in various camera angles (front, sides, and back) and 20 videos of not falling such as walking, crouching in various camera angles. All videos are recorded in a well-lit room for two seconds each. They are recorded in 50FPS, and resolution of 1024 by 720.

We used the LSTM (Long short-term memory) network since the network works well to recall the sequence of events. The network architecture used in this study consists of 10 hidden layers of LSTM, a 50 percent dropout layer, and 1 dense layer. We used sigmoid function as an activation function, binary cross-entropy as a loss function, and Adam optimizer (Figure 6). According to our test, the Sigmoid activation function performs better than the Tanh activation function in all cases. The 50 percent drop-out layer prevents the model from being over-fit, resulting in a better performance. Finally, Adam allows the model to converge faster and more robust in our test. We performed the grid search algorithm as the hyperparameter optimization technique.

To prepare the LSTM network, we broke the data into several chunks. We trained a model with different window sizes, ranging from 5 to 21 points, to find the most robust models. The window moves 1 point to the right at each time step (Figure 7).

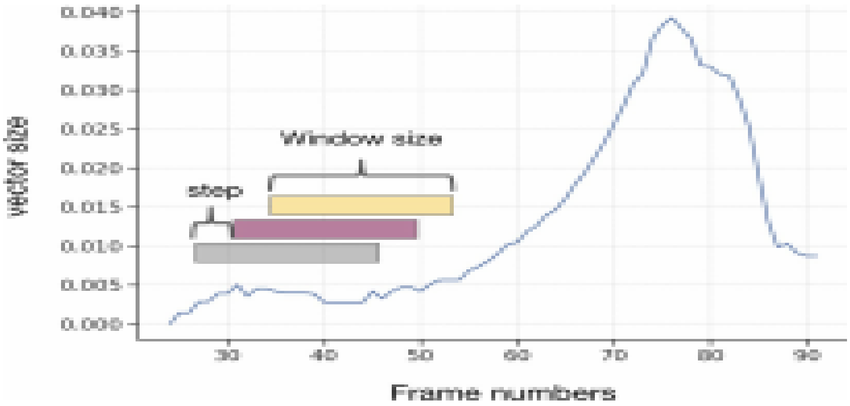
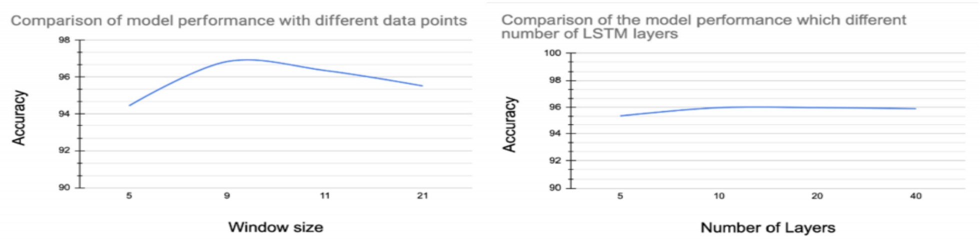


Figure 7 An image of splitting signals

We have trained the model for 150 iterations and the batch size is 32 data points, since the experiment indicates that the loss value is stable about 100 iterations. Approximately 2,000 data points for each mark are labelled *falling* and *not falling* to each frame manually. Then we break the data to 80 percent for training and 20 percent for testing. The model is trained on leave-one-out cross validation.

Research Results

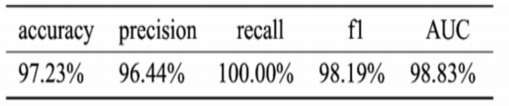
We conducted the same training operation with a different configuration of the hyper parameter. The result in Figure 8a shows that the optimum window size is 9 points. The model becomes more overfitted as the size of the windows increases. Considering the 50 FPS video source, 9 frames are 1/5 of a second. On the other hand, the number of hidden layers of LSTM reaches a diminishing return for about 10 layers as shown in Figure 8. This may mean that the set features are simple (only two features); thus, the model requires less memory.



(a) window sizes (b) LSTM layers

Figure 8 The model performance

Table 1 Precision, recall, f1, and AUC



accuracy	precision	recall	f1	AUC
97.23%	96.44%	100.00%	98.19%	98.83%

We use the best hyperparameter configuration above to perform additional measurements (precision, recall, f1, and AUC). The finding shown in Table 1 is positive. The overall accuracy of the model is 97%. A high recall value shows the model is highly effective at catching falls. On the other hand, the precision value means that the model is susceptible to falling, even though the actions do not. Given all measurements, the model is successful in detecting falls in our case.

Discussion of Research Results

We presented a single commodity camera to detect falls based on a 2D posing estimator. This approach works against camera orientation and lighting conditions. Since the model uses part of the time series signal to predict the fall, it can be detected before and during the fall. Thus, it does not have to wait until the user has done fallen. It can also be used in combination with devices that can respond quickly to mitigate the risk of injury like an airbag. While the framework performs premise outcomes, a lot of work needs to be done before it is incorporated into real-world devices. First, 97% precision is not sufficient for life saving applications. The device needs to be trained and checked with a broader range of CCTV cameras with varying lighting conditions, occlusions, and camera angles. The lower sampling rate is still unclear as to the predictive performance effect. However, this method is a starting point for any researcher to continue towards a more powerful single-camera, out-of-box, fall detection system.

Suggestions

1. The model uses part of the time series signal to predict the fall, it can be detected before and during the fall. this method is a starting point for any researcher to continue towards a more powerful single-camera, out-of-box, fall detection system. In the next study, it should be a developmental research study to create safety innovations for people in the network who need technology to help create safety

References

Andriluka M., Pishchulin L., Gehler P., and Schiele. B. (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Cao Z, et al..(2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7291–7299, 2017.

Charfi I., Miteran J., Dubois J., Atri M., and Tourki. R. (2012). Definition and performance evaluation of a robust SVM based fall detection solution. In 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, pages 218–224. IEEE, 2012.

Gibson M J, et al..(1987). Kellogg International Work Group on the Prevention of Falls by the Elderly. The prevention of falls in later life Danish Medical Bulletin, 34(4):1–24, 1987.

Hazelhoff J.., et al. (2008). Video-based fall detection in the home using principal component analysis. In International Conference on Advanced Concepts for Intelligent Vision Systems, pages 298–309. Springer, 2008.

Loughlin J.L.O, Robitaille Y., Boivin J F., and Suissa. S. (1993). Incidence of and risk factors for falls and injurious falls among the community-dwelling elderly. American journal of epidemiology, 137(3):342–354, 1993.

Lu N., Wu Y., Feng L., and Song. J. (2019). Deep learning for fall detection: Three dimensional CNN Combined with LSTM on video kinematic data. IEEE Journal of Biomedical and Health Informatics, 23(1):314–323, jan 2019.

Lin T., Maire M.,et al. (2014). Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

Lindemann U.,et al .(2005). Evaluation of a fall detector based on accelerometers: A pilot study. Medical and Biological engineering and computing, 43(5):548–551, 2005.

Núñez-Marcos A , Azkune G., and Arganda-Carreras. I. (2017). Vision-based fall detection with convolutional neural networks. Wireless Communications and Mobile Computing, 2017.

Rougier C, Meunier J, St-Arnaud A., and Rousseau. J. (2006). Monocular 3D head tracking to detect falls of elderly people. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, pages 6384–6387. IEEE, 2006.

Rougier C., Meunier J., St-Arnaud A., and Rousseau. J. (2011). Robust video surveillance for fall detection based on human shape deformation. IEEE Transactions on circuits and systems for video Technology, 21(5):611–622, 2011.

Solbach M. D. and Tsotsos. J. K. (2017). Vision-based fallen person detection for the elderly. In Proceedings of the IEEE International Conference on Computer Vision, pages 1433–1442, 2017.

V Vaidehi, et al. (2011). Video based automatic fall detection in indoor environment. In 2011 International Conference on Recent Trends in Information Technology (ICRTIT), pages 1016–1020. IEEE, 2011.

V Scott, L Wagar, and S Elliott. (2010). Falls and related injuries among older Canadians: Fall related hospitalizations and intervention initiatives. Prepared on behalf of the Public Health Agency of Canada, Division of Aging and Seniors. Victoria, BC: Victoria Scott Consulting, 2010.

Wangwiwattana C. , Sathienpaisal E., Chaiwattanaphong T., Singhapong P., Janrathitikarn O., and Akrawutpornpat. S. (2019). Falling Detection with a Single RGB Camera Based-on 2D Pose Estimation (in Progress). International Multidisciplinary Academic Conference, pages 75–81, 2019.

Wild D., Nayak U S., and Isaacs. B. (1981). How dangerous are falls in old people at home? Br Med J (Clin Res Ed), 282(6260):266–268, 1981.

Williams A., Ganesan D, and Hanson. A. (2007). 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. ACM, 2007.

Yu. X. (2008). 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.

Zhang T.,et al. (2006). Using wearable sensor and NMF algorithm to realize ambulatory fall detection. In International conference on natural computation, pages 488–491. Springer, 2006.