Fall Detection Based on YOLOv7

Ye Wei, Gu Xusheng, Ye Yuanxiang, Xiong Zhentao

Institute of Artificial Intelligence, Xiamen University
Student ID:36920221153139,36920221153079,36920221153140,36920221153131

Abstract

In the information age, computers are widely used in all fields. Object detection is one of the core problems in computer vision. At the same time, the problem of an aging population is getting worse. In the daily life of the elderly, falls are the most frequent events and the most serious threats to the health of the elderly. Because the monitoring equipment is widely used in daily life, the fall detection algorithm based on computer vision has more research significance and value. YOLO(You only look once) is the most popular deep learning network in the field of object detection. YOLO redefines object detection as a regression problem. It applies a single convolutional neural network to the whole image, divides the image into grids, and predicts the class probability and bounding box for each grid. Based on the existing YOLOv5 network, efficient aggregation network and reparameterized convolution are used to accelerate the network under the condition of ensuring the performance of the model. The experimental results show that our method improves mAP by 5.4% while maintaining the same inference speed as the previous method. In addition, we try to combine the Coordination Attention module which can improve the effect of objection detection task, hoping to further improve the performance of the network. The experimental results show that the addition of CA(coordination attention) module leads to the decline of network performance.

Introduction

Fall detection has a wide range of applications. For example, fall detection is quite necessary for elderly people or children who are alone at home. If they can be detected when they fall, then they will get more timely treatment. Through fall detection, fall victims can be detected and treated in time, which will greatly reduce injuries caused by accidental falls. Therefore, it is a meaningful study to monitor the fall behavior of the elderly and make emergency measures in time. In addition, accuracy and recall need to be considered.

Thanks to recent progress in deep neural network and image processing technology, more and more new task can be solved easily and old problems also meet new solutions, such as fall detection. Furthermore, newer advances in equipment and detection algorithms have also allowed the problem to be better solved. HD(High Definition) cameras can capture more information for computer analysis. At present, the proposed fall detection algorithms include wearable based (Karantonis et al. 2006; Lee, Robinovitch, and Park 2014) and environment based (Fang et al. 2006; Zhuang et al. 2009). Pedestrian fall detection based on computer vision (Foroughi, Aski, and Pourreza 2008; Rougier et al. 2011; Thome and Miguet 2006; Rougier et al. 2006) has many advantages, such as deploying monitoring equipment.

HD cameras have been able to detect and monitor in complex environment. Monitors are able to perceive unknown environments and detect human posture autonomously. However, detection in environments with many people, especially with different postures, is still a challenging problem. For example, monitor have difficulty identifying falls and squats. Moreover, the time to analyze the human posture needs to be taken into account because the task is in real time.

To address the above issues, this paper proposes a fall detection pipeline based on the YOLOv7, including human detection, posture analysis and prediction, and fall alert. YOLO is a deep neural network-based object recognition and localization algorithm, whose most important feature is that it runs very fast and can be used in real-time systems. The YOLO algorithm uses a CNN with direct regression function to complete the whole process of target detection. In this work, the model first completes the detection of the person and marks the position of the person. Then the pose of the person is analyzed by extracting features. Finally determine whether the human body is in a state of falling, and if so, issue a warning.

We perform extensive quantitative and qualitative experiments on some challenging scenarios to validate our fall detection and warming framework, which provide a new application based on YOLOv7.

Related Work

To the best of our knowledge, the research on pedestrian fall detection can be roughly divided into the following three types: method based on wearable, environment and computer vision. Dean et al. evaluated the fall of the elderly through the acceleration vectors of different axes (Karantonis et al. 2006), but the rate appears to be high; Since then, Lee et al. proposed an approach based on the vertical velocity component (Lee, Robinovitch, and Park 2015). Mostapha et al. considered the hardware embedded into the sole to reduce intrusiveness constraints (Zitouni et al. 2019), which improved the acceleration and timeliness. Hussain et al. proposed a wearable sensor-based continuous fall monitoring system capable of detecting falls and identifying fall patterns and activities associated with fall events (Hussain et al. 2019). However, the above wearable method relies too much on equipment, especially a big inconvenient for the elderly.

Environment based feels the surroundings through infrared sensors, which greatly reduces the necessity of wearing. Zhuang et al. detected the falling behavior by extracting the fluctuation of sound signal from audio equipment in the environment (Zhuang et al. 2009); Mazurek et al. extract kinematic features and mel-cepstrum-related features for classification to assess their utility using data from infrared depth sensors (Mazurek, Wagner, and Morawski 2018), which further improved the accuracy. Although that reduces the trouble of wearing, the detection is easily affected by noise around. Guto et al. consider deep learning for fall detection in IoT and fog computing environments. They propose a convolutional neural network consisting of three convolutional layers, two max pooling and three fully connected layers as our deep learning model (Santos et al. 2019). Furthermore, the detection devices costs too much, making it an unrealistic way in real life.

Computer vision based ways collect crowd behavior information through video, and recognizes human posture according to the detection algorithm. Feng et al. fit the contour of the target into an ellipse, which geometric and motion features are extracted to form a new feature by SVM(Support Vector Machine) (Feng, Liu, and Zhu 2014). Min et al. represent pedestrian by a rectangular box, and explaine the posture of the pedestrian by the length width ratio of the rectangular box, so as to detect the fall (Min et al. 2018). Espinosa et al. proposed a fall detection system based on a 2D CNN inference method and multiple cameras. This method analyzes images in a fixed time window and uses optical flow extraction features to obtain information about relative motion between two consecutive images. (Espinosa et al. 2019) Although the above are portable and does not cost a lot, they use predetermined models which perform not well in complex and changing environments, such as strong illumination change, dynamic background interference, occlusion problems and so on.

The method based on deep learning can improve this deficiency by virtue of its network learning ability of nonlinear mapping, and also has a good performance in the detection task. The main content of this paper is pedestrian fall detection based on YOLO algorithm.

Proposed Solution

In our work, we use YOLOv7 to detect the fall behavior of pedestrians. We show that YOLOv7-based fall detection has higher speed and accuracy than YOLOv5-based. The flexible insertion of the Coordinate Attention module into YOLOv7 allows the network to perform better in the detection task while adding little computational overhead. In this task, we use the YOLOv7 network with Coordinate Attention module to implement pedestrian fall detection.

Input

The input to the network is an image or video frame. YOLOv7 was trained and tested on relatively large images like 640*640 and 1280*1280 at the beginning, so the input to this work is also a larger image. In order to facilitate the experiment, the dataset of our experiment uniformly converted the image into a resolution of 640*640.

Backbone Network

The backbone of YOLOv7 is shown in Fig.1. there are 50 layers in total in the backbone of YOLOv7. Firstly, it goes through 4 convolutional layers, and the CBS is mainly composed of Conv + BN + SiLU. After 4 CBSs, the feature map becomes 160 * 160 * 128 in size. After that, we pass through the ELAN module proposed in the paper, which is composed of multiple CBSs, the input and output feature sizes remain the same, the number of channels changes in the first two CBSs, the next few input channels are consistent with the output channels, and the last CBS is output as the desired channel. Overall, the backbone, after 4 CBS, access to, for example, an ELAN, and then the back is the output of three MP + ELAN, corresponding to the output of C3/C4/C5, the size of 80 * 80 * 512, 40 * 40 * 1024, 20 * 20 * 1024. Each MP by 5 layers, ELAN has 8 layers, so the entire backbone layer is 4 + 8 + 13 * 3 = 51 layers, starting from 0, the last layer is the 50th layer.

Head

YOLOv7 head is actually a pafpn structure, just like the previous YOLOv4 and YOLOv5. First, for the backbone final output 32 times downsampling feature map C5, then after SPPCSP, the number of channels is changed from 1024 to 512. first fuse with C4 and C3 according to top down, to get P3, P4 and P5; then fuse with P4 and P5 according to bottom-up. The difference is that the CSP module in YOLOv5 is replaced by a module similar to ELAN, which is slightly different from ELAN in backbone in that the number YOLOv7 head is actually a pafpn structure, just like the previous YOLOv4 and YOLOv5. First, for the backbone final output 32 times downsampling feature map C5, then after SPPCSP, the number of channels is changed from 1024 to 512. first fuse with C4 and C3 according to top down, to get P3, P4 and P5; then fuse with P4 and P5 according to bottom-up. The difference is that the CSP module in YOLOv5 is replaced by a module similar to ELAN, which is slightly different from ELAN in backbone in that the number of cat is different. Also the downsampling is changed to MP2 layer.

Coordinate Attention

Channel attentions have significant effect on improving model performance, but they usually ignore location information, which is very important for generating spatially selective attention maps. Therefore embedding location information into channel attention. Unlike channel attention, which converts the feature tensor into a single feature vector by 2-dimensional global pooling, coordinate attention decomposes channel attention into two 1-dimensional feature

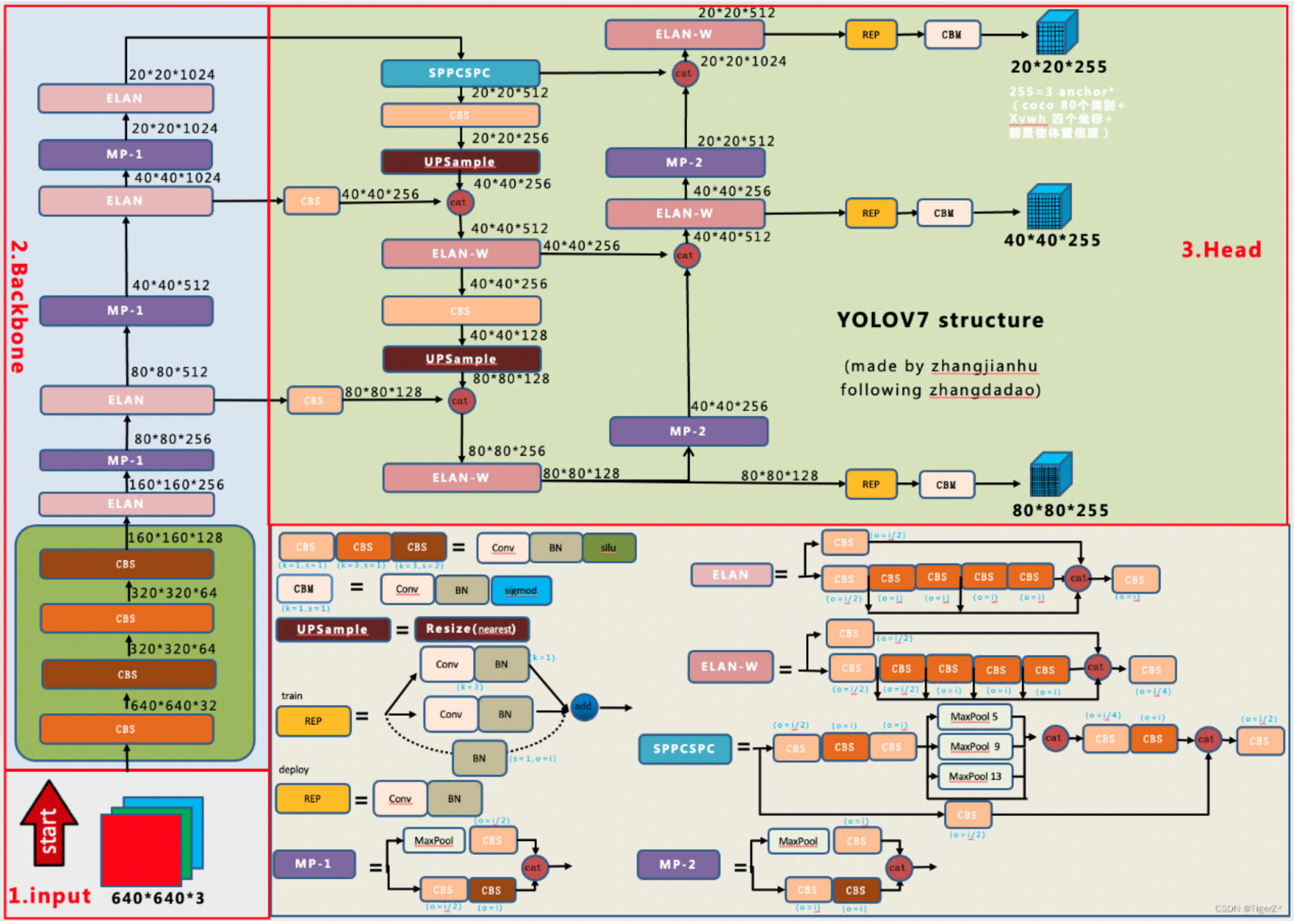

Figure 1: The network structure

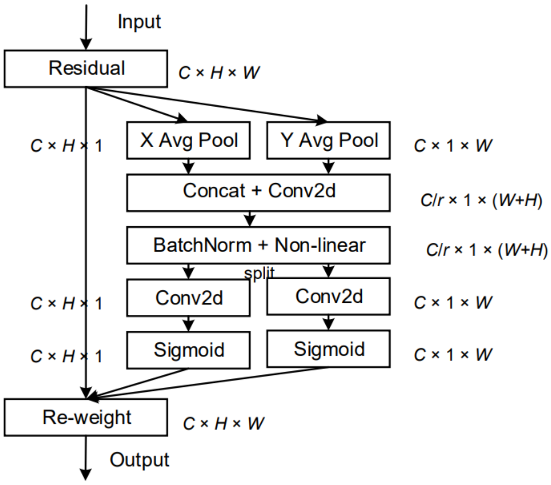

Figure 2: Coordinate Attention Block

encoding processes that aggregate features along 2 spatial directions, respectively. In this way, remote dependencies can be captured along one spatial direction, while accurate location information can be retained along the other spatial direction. The generated feature maps are then encoded as a pair of direction-aware and position-sensitive attenton maps, respectively, which can be applied complementarily to the input feature maps to enhance the representation of the object of attention.

Experiments

Datasets and Experimental Settings

The data set of this experiment was collected through multiple methods, and 1428 pictures of falls were obtained. Among the 1428 images, 1142 images are randomly selected as the training set, and the remaining 286 images are used as the test set. The YOLO main structure is used as the baseline, and the number of iterations is set to 100, among which the frozen trunk feature extraction network is trained 50 times, and trained 50 times after unfreezing to speed up the training process. The input image size is normalized to 640*640 to reduce the demand for GPU memory. The backbone network passes through 4 CBSs, connects to 1 ELAN, and then passes through 3 MP+ELAN outputs, corresponding to the output of C3/C4/C5, the sizes are 80*80*512, 40*40*1024, 20*20*1024. And the output module adopts YOLO head structure.

To improve the detection capability of the target detection model, the diversity of fall poses in the collected images is ensured as much as possible. In addition, the image set also includes small targets, target stacking, and occlusion. The input image size of the model is 640×640, the number of channels is 3, and the images in the dataset are original images without any pre-processing such as clarification.

Experimental Environment and Protocol Design

The experimental environment of this paper is based on Windows 10 64-bit system, 16GB RAM, GPU version is NVIDIA GeForce RTX 3090 24GB Laptop, GPU acceleration library is Cuda11.3, Cudnn10.0, and the software used includes Anaconda, Pycharm, etc., to build deep learning Pytorch framework to implement the training of YOLOv5 target detection model. The input image resolution is 640×640, and the Mosaic data enhancement method is used in the training in order to enhance the model's anti-interference ability. The batch size is 32, the number of training rounds is 100 rounds, the momentum coefficient is 0.937, the weight decay coefficient is 0.0005, and the initial learning rate is 0.01.

Evaluation Indicators

To validate the performance of YOLOv7's algorithm, this paper uses generic target detection evaluation metrics, Accuracy(A), Precision (P), Recall (R), F1 score and mean Average Precision (mAP) to evaluate the model. Two metrics, P and R, are usually used to measure the goodness of the model, and mAP can measure the performance of the whole model.

$$A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \tag{1}$$

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{2}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

$$mAP = \frac{1}{c}\sum_{j=1}^{c} AP_j \tag{5}$$

Training Results

The following pictures can more intuitively reflect the performance of the algorithm on various evaluation indicators. Fig.3 shows the Precision curve of the training set results. Fig.4 shows the Recall curve of the training set results. Fig.5 shows the P-R curve of the training set results. Fig.6 shows the F1 score of the training set results.

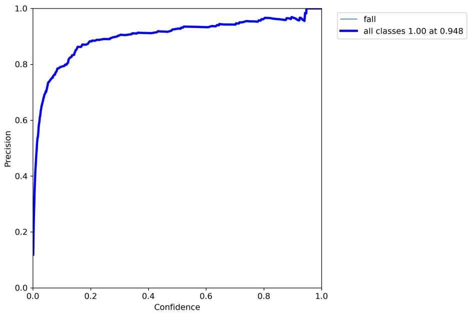

Figure 3: Precision curve on the training set

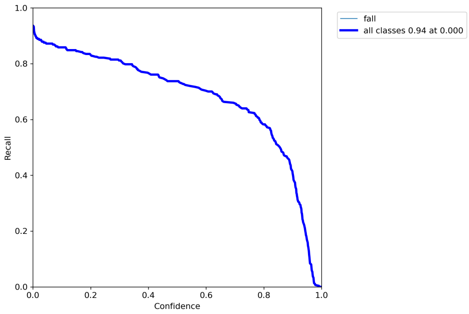

Figure 4: Recall curve on the training set

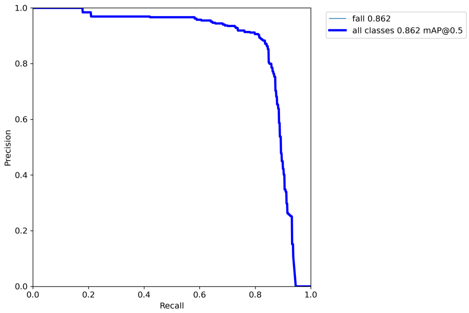

Figure 5: P-R curve on the training set

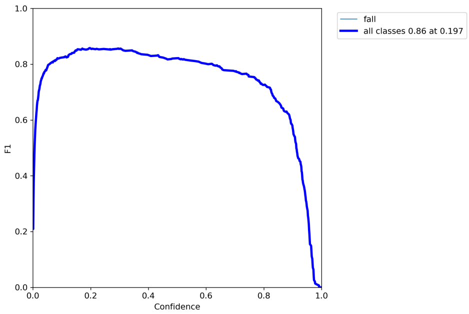

Figure 6: F1 score on the training set

Test Results

In order to show the improvement effect of the network more clearly, images of fallen pedestrians are randomly selected from the test set for testing. The test results of the YOLOv5 and YOLOv7 algorithm parts are shown in Fig.7 and 8. We can clearly see that YOLOv7 detected the fall in the lower left picture, but YOLOv5s did not.

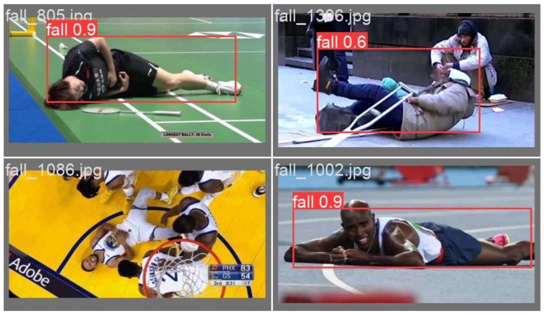

Figure 7: The results of the model YOLOv5s on the test set

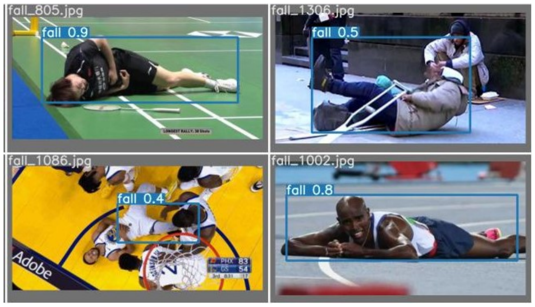

Figure 8: The results of the model YOLOv7 on the test set

In order to reflect the performance improvement of the YOLOv7 algorithm more intuitively, compare the Accuracy (100%), Precision (100%), Recall (100%), and mAP (100%) of the YOLOv7 algorithm with YOLOv5s and other target detection algorithms, and compare the experimental results As shown in Table 1. We can clearly see that YOLOv7 performs better than YOLOv5s in various evaluation indicators.

Models Indications	YOLOv5s	YOLOv7
Precision(100%)	87.3	90.0
Recall(100%)	74.1	81.8
mAP(100%)	82.8	88.2

Table 1: Test set results for YOLOv5s and YOLOv7 models.

Effect Analysis of Coordinate Attention

At the end of the experiment, we improved the YOLOv7 algorithm model and added an attention mechanism called Coordinate Attention(CA). The experimental results are shown in Table 2.

Models Indications	YOLOv7	YOLOv7+CA
Precision(100%)	90.0	87.5
Recall(100%)	81.8	79.8
mAP(100%)	88.2	84.5

Table 2: Test results of models YOLOv5s, YOLOv7 and YOLOv7 with CA mechanism.

Experimental Results and Analysis

In this paper, we propose a YOLOv7-based pedestrian fall detection model with an attention mechanism that helps the network focus on fall poses more accurately. Experiments have proved that the improved algorithm has better detection accuracy and detection effect than the original YOLOv5s network, and is suitable for falling pedestrian detection projects.

Conclusion

In this paper, For the fall detection problem in the field of object detection, we improve the existing yolov5 network, and optimize it in two aspects: model reparameterization and dynamic label allocation. We analyze the propagation path of the gradient to optimize the structure reparameterization for different layers in the network, and propose the model structure reparameterization for different programs. We propose a new label allocation method, which is based on the prediction of the lead head and generates hierarchical labels from coarse to fine for the learning of the lead head and the auxiliary head respectively.Based on the above two improvements, the experiment on our dataset shows our network will eventually increase the mAP from 82.8% to 88.2% compared with yolov5 without increasing the inference time. In addition, we have made some exploration to improve the performance of YOLOv7 network. We have noticed that CA module is flexible and lightweight, and it is easy to plug into the existing classical network. We use the CA module combined with YOLOv7 network to carry out experiments. However, the added modules did not improve the network performance as expected. On the contrary, the speed and accuracy of network inference decline. For the task of fall detection, it may be necessary to design a more appropriate network structure to improve the speed and accuracy.

References

Espinosa, R.; Ponce, H.; Gutiérrez, S.; Martínez-Villaseñor, L.; Brieva, J.; and Moya-Albor, E. 2019. A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-Fall detection dataset. *Computers in Biology and Medicine*, 115: 103520.

Fang, J.-S.; Hao, Q.; Brady, D. J.; Guenther, B. D.; and Hsu, K. Y. 2006. Real-time human identification using a pyroelectric infrared detector array and hidden Markov models. *Optics express*, 14(15): 6643–6658.

Feng, W.; Liu, R.; and Zhu, M. 2014. Fall detection for elderly person care in a vision-based home surveillance environment using a monocular camera. *Signal, Image and Video Processing*.

Foroughi, H.; Aski, B. S.; and Pourreza, H. 2008. Intelligent video surveillance for monitoring fall detection of elderly in home environments. In *2008 11th international conference on computer and information technology*, 219–224. IEEE.

Hussain, F.; Hussain, F.; Ehatisham-ul Haq, M.; and Azam, M. A. 2019. Activity-Aware Fall Detection and Recognition Based on Wearable Sensors. *IEEE Sensors Journal*, 19(12): 4528–4536.

Karantonis, D.; Narayanan, M.; Mathie, M.; Lovell, N.; and Celler, B. 2006. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE Transactions on Information Technology in Biomedicine*, 10(1): 156–167.

Lee, J. K.; Robinovitch, S. N.; and Park, E. J. 2014. Inertial sensing-based pre-impact detection of falls involving near-fall scenarios. *IEEE transactions on neural systems and rehabilitation engineering*, 23(2): 258–266.

Lee, J. K.; Robinovitch, S. N.; and Park, E. J. 2015. Inertial Sensing-Based Pre-Impact Detection of Falls Involving Near-Fall Scenarios. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(2): 258–266.

Mazurek, P.; Wagner, J.; and Morawski, R. Z. 2018. Use of kinematic and mel-cepstrum-related features for fall detection based on data from infrared depth sensors. *Biomedical Signal Processing and Control*, 40: 102–110.

Min, W.; Cui, H.; Rao, H.; Li, Z.; and Yao, L. 2018. Detection of Human Falls on Furniture Using Scene Analysis Based on Deep Learning and Activity Characteristics. *IEEE Access*, 6: 9324–9335.

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*, 6384–6387. IEEE.

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.

Santos, G.; Endo, P.; Monteiro, K.; Rocha, E.; Silva, I.; and Lynn, T. 2019. Accelerometer-Based Human Fall Detection Using Convolutional Neural Networks. *Sensors*, 19(7): 1644.

Thome, N.; and Miguet, S. 2006. A HHMM-based approach for robust fall detection. In *2006 9th International Conference on Control, Automation, Robotics and Vision*, 1–8. IEEE.

Zhuang, X.; Huang, J.; Potamianos, G.; and Hasegawa-Johnson, M. 2009. Acoustic fall detection using Gaussian mixture models and GMM supervectors. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 69–72.

Zitouni, M.; Pan, Q.; Brulin, D.; and Campo, E. 2019. Design of a Smart Sole with Advanced Fall Detection Algorithm. *Journal of Sensor Technology*.