Deep Learning-based Fall Detection

Jason Wei Hoe Chiang and Li Zhang[†]
Department of Computer and Information Sciences, Faculty of Engineering and Environment, Northumbria University,
Newcastle upon Tyne, NE1 8ST, U.K.
[†]E-mail: li.zhang@northumbria.ac.uk

In the modern information era, fall accidents are one of the leading causes of injury, disability and death to elderly individuals. This research focuses on object detection and recognition using deep neural networks, which is applied to the theme of fall detection. We propose a deep learning algorithm with the capability to detect fall accidents based on the state-of-the-art object detector, YOLOv3. Our system is tested on a challenging video database with diverse fall accidents under different scenarios and achieves an overall accuracy rate of 63.33%. The proposed deep network shows great potential to be deployed in real-world scenarios for health monitoring.

Keywords: Fall Detection, Deep Learning, Convolutional Neural Network

1. Introduction

One of the most ubiquitous issues faced by the elderly today is the risk of falling down which could cause devastating injuries or even death. According to an article published by the World Health Organization [1], research suggests that between 28% - 35% of the population over 65 years old suffer at least one fall accident per annum. Subsequently, this figure upsurges to 42% for people who are over 70 years old [1]. The World Health Organization also reported that over 50% of elderly hospitalizations and approximately 40% of non-natural mortalies for the elderly are contributed by fall accidents. Additionally, fall accidents may result in a post-fall syndrome such as dependence, immobilization, and depression, which leads to further constraints in daily activities.

As mentioned, fall accidents may result in a catastrophic post-fall syndrome. Thus, it is crucial that the elderly require proper treatment immediately after post-fall accidents. Based on Rodrigues, Huber and Lamura [2], it is suggested that approximately 33% of the elderly (those who are 65 years old or older) in Europe live on their own. Moreover, research shows that

[*] This work is supported by Northumbria University.

there will be a significant increase in elderly population over the next twenty years. For this reason, investigations and developments for a smart and reliable fall detection system are extremely crucial to resolve this issue [2]. An efficient fall detection system can restore confidence and enable senior citizens to carry on with their normal active lifestyles. Ultimately, in this research, the proposed system aims to deliver an Artificial Intelligence (AI) prototype system that has the capability to recognize a fall accident when a raw video is fed in.

The proposed system consists of three key phases: (1) performing person detection/recognition; (2) identifying if the detected subject has fallen; (3) producing an output video with the respective bounding box and label. The overall activity flow and visual representative of the prototype are shown in Fig.1. In the first phase, a deep Convolutional Neural Network (CNN) is implemented to localize and classify the person within the video. We employ an object detector, i.e. You Only Look Once version 3 (YOLOv3), for this person detection and recognition task owing to its impressive performance and fast processing speed for object detection. In the second phase, the position/orientation of the detected person within the video is used to detect a fall accident. In the last phase, an output video with the respective bounding box and label (person or fall action) will be produced by using the OpenCV library.

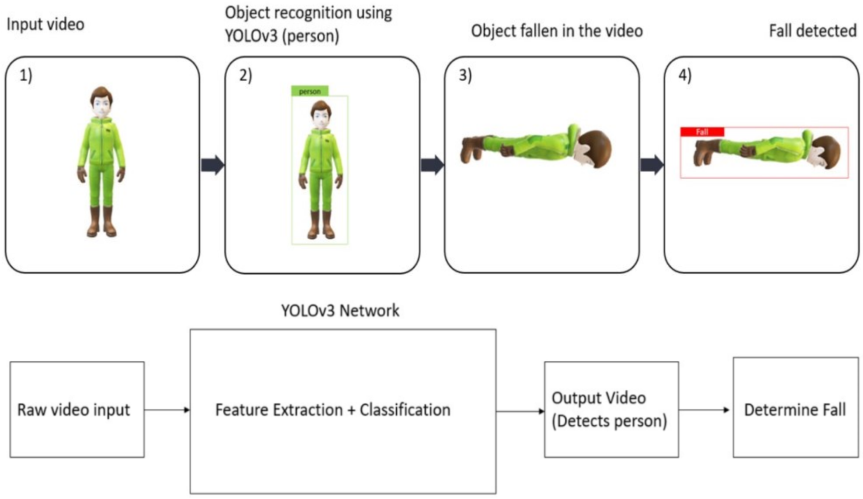

Input video
Object recognition using YOLOv3 (person)
Object fallen in the video
Fall detected
1)
2)
person
3)
4)
Fall
YOLOv3 Network
Raw video input
Feature Extraction + Classification
Output Video (Detects person)
Determine Fall

Fig. 1. The proposed system

1.1. *Research Contributions*

The main contributions are as follows. A deep CNN architecture is developed as a real-time fall accident detector in videos. This fall action detector is an extended development based on a real-time object detector in YOLOv3. In other words, the proposed system showcases the capabilities of utilizing a state-of-the-art deep learning object detector for fall accident detection. Our approach illustrates promising performance when tested on a private video dataset.

The paper is organized as follows. Section 2 discusses state-of-the-art object detectors and related studies on fall accident detection. Section 3 presents the proposed deep learning-based fall detection. System evaluation is discussed in Section 4. Section 5 concludes this research and identifies the future directions.

2. Related Work

In this section, we introduce the state-of-the-art object detector, YOLO, which lays a foundation for this research, as well as related studies on fall detection.

2.1. *Object Detectors*

Proposed by Redmon et al. [3-5], YOLO uses a single CNN model to predict bounding boxes and the class probabilities [3-5]. Firstly, YOLO takes an image and fragments it into an SxS grid. Inside each grid, YOLO generates *m* bounding boxes. Next, YOLO produces a class probability and counterbalance values for each bounding box using the CNN model. Then, those class probability scores of the bounding boxes above a pre-defined threshold are selected and used to locate the corresponding objects within the image. As compared with other object detectors, YOLO can keep up with an astounding speed of 45 frames per second. Furthermore, Redmon et al. [4, 5] have also made modifications such as applying the Darknet classification network to the initial YOLO network over the years to improve its performance. We select YOLOv3 in this research owing to its superior capabilities in detecting objects in real-time with competitive precision. A detailed comparison between YOLO and other object detectors such as Regional CNN (R-CNN), Fast R-CNN and Faster R-CNN, is illustrated in Table 1.

Table 1. Performance comparison between YOLO and other object detectors (where *PASCAL VOC 2007 contains only 9,963 images for training/validation/testing with 20 classes while COCO contains 80,000 training images and 40,000 validation images with 80 classes.)*

Model	Prediction Time (per Image)	Mean Average Precision (mAP)
R-CNN	40-50 seconds	66% (Pascal VOC)
Fast R-CNN	Approximately 2 seconds	66.9% (Pascal VOC)
Faster R-CNN	0.2 seconds	66.9% (Pascal VOC)
YOLOv1	Able to detect images in real-time at 45 frames per second (FPS)	63.4% (45 FPS) (Pascal VOC)

YOLOv2/9000	Able to detect images in real-time at 45 FPS	76.8% (45 FPS) (Pascal VOC)
YOLOv3	Able to detect images in real-time at 45 FPS	57.9% (20 FPS) (COCO)

2.2. *Fall Detection*

Vision-based technical advancements and related studies showed promising performances for fall detection. For instance, Poonsri et al. [6] proposed a method based on the combination of background subtraction and Gaussian model for human detection. Subsequently, aspect, orientation and area ratios were computed for feature extraction to classify the human postures to inform fall detection. However, the issue within background subtraction is that it may not be able to accurately detect multiple human subjects which will contribute to an inaccurate classification outcome. Moreover, this method may not work if the target subjects were occluded by other objects.

Fielding et al. [7] proposed a vision-based health and emotion well-being monitoring system for fall and hazard detection. Their work conducted image description generation integrating object detection and classification using Faster R-CNN, RNN-based attribute classification, and template-based language generation. Their system was able to not only detect hazards and fall actions, but also describe an input image with natural sounding language. Evaluated using IAPR TC-12 and several other private fall datasets, their model showed superior performance. Their further developments [8, 9] have led to two end-to-end deep learning models for a variety of human attribute (age, gender and ethnicity) prediction, human action recognition and encoder-decoder RNN-based image description generation. Other fall detection techniques also include using electromagnetic and infrared sensors.

3. The Proposed Methodology

This research aims to diminish the limitations mentioned in related studies to deliver a real-time fall detector with precision, based on YOLOv3. Specifically, the proposed system first uses YOLOv3 to perform person detection within the video frame whilst simultaneously using the height-width ratio of the detected human subject to perform a fall detection. Moreover, the system will be able to identify when the subject has fallen based on the colour of the bounding box and label. The system contains four key stages: (1) receiving a raw video as an input, (2) person detection and recognition, (3) fall detection, and (4) producing an output video with the respective bounding box and label.

3.1. *Object Detection and Recognition*

Firstly, we use the OpenCV library to perform video processing. To achieve robust object detection, we implement the YOLOv3 object detector using the PyTorch framework. YOLOv3 is trained using the COCO dataset as mentioned in Table 1, therefore having the capability to detect, localize and classify up to 80 object classes (e.g. person, car, cat, dog, etc.). The network contains 75 convolutional layers with the fully connected layers removed for the purpose of multi-labeling. YOLOv3 is designed to deal with input of various sizes [5]. In addition, the maxpooling layer from the previous iterations of YOLO was also removed in YOLOv3 and replaced with convolutional layers with the stride value of 2, instead. Thus, downsampling is applied to the feature map instead of performing the maxpooling process.

To predict the bounding boxes, the network utilizes dimension clusters as the anchor boxes as those of its previous version YOLOv2/9000 [4]. It predicts four coordinates for each bounding box, namely, t_x, t_y, t_w, t_h. If the cell is offset from the top corner of the image and the bounding box anchors have both a width and a height, then the predictions will correspond to [5]:

$$b_x = \sigma(t_x) + c_x \tag{1}$$

$$b_y = \sigma(t_y) + c_y \tag{2}$$

$$b_w = p_{w^{e^{t_w}}} \tag{3}$$

$$b_h = p_{h^{e^{t_h}}} \tag{4}$$

Moreover, an objectness score for each bounding box is predicted using logistic regression function. The score will be 1 if the previous bounding box (anchors) intersects a ground truth object more than any other previous bounding box. Subsequently, a threshold of 0.5 is introduced in deciding whether to ignore or accept the prediction. If the intersection is above the predefined threshold (0.5), then the prediction will be ignored. Additionally, only one bounding box prior is allocated for each ground truth object. On the contrary, the model will not make any classification and localization dismissed if a bounding box prior is not assigned [5].

Lastly, the class prediction is achieved by using an independent logistic classifier and cross entropy. The use of binary cross entropy and independent logistic classifier has improved upon YOLO's capability when dealing with more complex datasets where labels often overlap with each other, for example, when the labels are both golden retriever and dog [5]. The implementation of YOLOv3 by Redmon et al. [5] is modified in this research to only detect a single class (i.e. person), since that is the only class that is required for fall detection.

3.2. *Fall Detection*

The following implementation of fall detection is built upon the modifications discussed in Section 3.1. The main reason that this methodology was implemented is because of the superior precision of YOLOv3 in detecting human in multiple different orientations/positions, e.g. person fallen down, sitting on a couch, standing upright, etc. Based on our observations in various videos and images of fall accidents, we notice that post-fall positions of the persons in majority of the videos and images were approximately 90° off from a person standing upright as indicted in Fig 2.

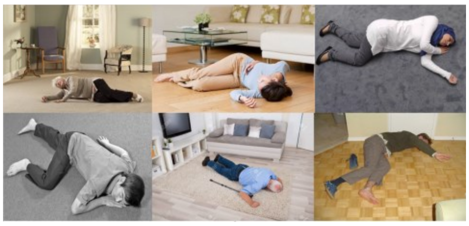
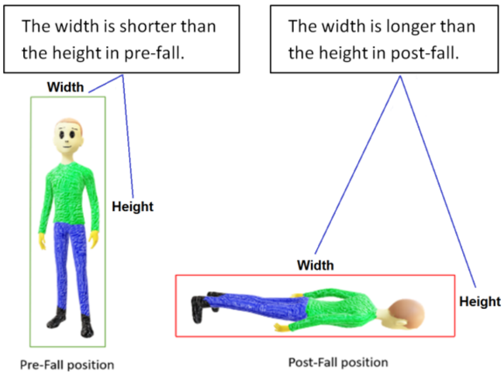


Fig. 2. The position of a person after a fall accident



The width is shorter than the height in pre-fall.
The width is longer than the height in post-fall.
Width
Height
Width
Height
Pre-Fall position
Post-Fall position

Fig. 3. The visual representation of the concept for fall detection

Following the concept above, we first retrieve the ratio (height and width) of the bounding box of a detected person within the video. Subsequently, the ratio of the bounding box can be used to determine if the person has fallen or not based on the dimension of the bounding box.

Since a video is just a continuous series of a single frame, this concept is achieved by implementing the Python Image Library (PIL). Firstly, the whole frame is obtained using the "Image" function by PIL. Subsequently, the OpenCV function "crop" is used to crop the detected object using the coordinates of the object generated by YOLOv3 network. Lastly, we use the "size" function provided by PIL to obtain the ratio of the detected object within the whole frame. The ratio can then be used to determine a fall action.

4. System Evaluation

We evaluate the proposed system using a video data set. The testing is separated into two parts: (1) Person Detection, and (2) Fall Detection.

4.1. *Single Class (Person) Detection & Fall Detection*

Before utilizing the bounding box's dimension to inform a fall accident detection, (1) we first test if the system only detects one class (i.e. person) using a total of ten video randomly selected online. After meticulously testing on different videos, the person detection achieves a 100% accuracy rate. (2) We subsequently conduct fall detection. First, we collect a total of 30 fall accident videos from http://le2i.cnrs.fr/Fall-detection-Dataset?lang=fr, https://www.youtube.com/, as well as capturing one video from a smartphone.

The test videos consist of those with different lighting conditions, fall positions, video angles, video qualities and lengths (e.g. 6-30 seconds). The results were measured by the amount of correct classifications and the maximum FPS achieved from the output video which will be compared with the original video. Our system is tested on a setup with AMD Ryzen Threadripper 1950X 16-Core CPU processor, NVIDIA GTX 1070TI 8GB and 64GB RAM.

Evaluated using the above dataset, our system is able to achieve a fall detection accuracy rate of 63.33% when the raw videos are used as inputs. After testing using all 30 videos, we spot a few minor flaws, e.g. occlusions in the lower bodies, which consequently affect the bounding box dimension calculation. However, despite the minor flaws, the system still performs relatively well under various conditions.

5. Conclusion

We develop a deep learning-based fall detector, based on YOLOv3. The proposed system starts by localizing and classifying human subjects within the frame using YOLOv3. Subsequently, we utilize the bounding box ratio and dimension to classify and determine a fall accident, which is achieved by using multiple libraries such as OpenCV and PIL. The experimental results indicate that the proposed system shows promising performance in dealing with real-time fall detection. To the best of our knowledge, there is still little to no system that utilizes a deep CNN to perform a fall detection. Thus, we believe that this work has introduced a new way to tackle this everlasting issue.

For future work, we aim to equip our current proposed work with transfer learning [10] which could greatly improve the overall practicality and robustness of the system and overcome the minor flaws discussed above. Besides that, we also aim to improve the system by implementing functionalities such as an immobilization timer which can be used to set off the alarm when the fallen subject is not moving for a certain period of time indicating that he/she might be in a fatal condition.

References

1. World Health Organization. WHO Global Report on Falls Prevention in Older Age. (2007). [online] Available at: http://www.who.int/ageing/publications/Falls_prevention7March.pdf
2. R. Rodrigues, M. Huber, M. and G. Lamura. Facts and Figures on Healthy Ageing and Long-term Care. *Europe and North America, Occasional Reports Series 8*. Vienna: European Centre. (2012).
3. J. Redmon, S. Divvala, R. Girshick and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788). (2016).
4. J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271). (2017).
5. J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. (2018).
6. A. Poonsri and W. Chiracharit. Improvement of fall detection using consecutive-frame voting. In *Proceedings of International Workshop on Advanced Image Technology (IWAIT)* (pp. 1-4). (2018).
7. B. Fielding, P. Kinghorn, K. Mistry and L. Zhang. An Enhanced Intelligent Agent with Image Description Generation. In *Proceedings of International Conference on Intelligent Virtual Agents*, 110-119. USA. (2016).
8. P. Kinghorn, L. Zhang and L. Shao. A Hierarchical and Regional Deep Learning Architecture for Image Description Generation. *Pattern Recognition Letters*. (2019).
9. P. Kinghorn, L. Zhang and L. Shao. A region-based image caption generator with refined descriptions. *Neurocomputing*. 272 (2018) 416-424.
10. T.Y. Tan, L. Zhang and C.P. Lim. Adaptive melanoma diagnosis using evolving clustering, ensemble and deep neural networks. *Knowledge-Based Systems*. (2019).