ISSN 2616-5775 Vol. 5, Issue 4: 29-34, DOI: 10.25236/AJCIS.2022.050405

YOLOv5-based fall detection method

Sichao Cheng[1], Hekai Zhang[2]

[1]*Tianjin University of Science and Technology, Tianjin, China*
[2]*Heilongjiang University, Harbin, Heilongjiang, China*

Abstract: *With the rising elderly population in China, detecting whether an elderly person has fallen is one of the problems that people need to pay attention to today, however, most of the current detection methods are affected by problems such as expensive, vulnerable to environment and not easy to implement. In order to solve the above problems, this paper proposes a fall detection method with YOLOv5s as the basic network model, which first enhances the original image, and then improves the loss function and NMS non-maximal suppression. The final results show that applying this improved algorithm model can effectively perform fall detection.*

Keywords: *Fall detection; YOLOv5s; NMS non-maximal suppression*

1. Introduction

Currently, falls have become the "main culprit" in causing injuries to the elderly. The World Health Organization reports that more than 300,000 people worldwide die from falls each year, with half of those over 60 years of age. Prospective studies have reported fall rates of 15 to 26 percent for older adults in the community. 4 to 5 percent of Chinese seniors fall two to three times in a year. On average, 44% of falls occurred in the home, with the living room, dining room and bedroom being the most common indoor locations where falls occurred. 22% to 76% of outdoor falls occurred in the street or on the sidewalk. Most falls (59% to 97%) occurred during daytime hours, with a significantly higher rate of daytime falls among older adults in rural areas (88%) than in urban areas (69%). According to statistics, falls are the number one cause of injury deaths among people over the age of 65 in China. How to detect falls and take action in a timely manner has become a worldwide concern.

Studies have shown that timely treatment of falls in the elderly can reduce the risk of death by 80% and the risk of long-term supportive care by 26%. Therefore, it is important to detect falls in surveillance video in a timely manner: on the one hand, it can alert the guardians for help in order to reduce the injuries caused by falls; on the other hand, it can save the public medical resources and reduce the public medical burden of society.

In the last decade, many works have been carried out by researchers in fall detection algorithms. Depending on the devices used and the detection methods, fall detection algorithms are divided into the following three main areas [1-3]: wearable sensor-based methods, scene-based sensor-based methods, and computer vision-based methods. Wearable sensor-based approaches use sensors to automatically detect falls and send help information to health care providers through communication devices such as WIFI, mobile network, and Bluetooth, and thus have many advantages, but many elderly people will often forget to wear the sensors. Scene sensor-based approaches use scene sensors installed in the monitoring area to collect pressure, vibration and sound information to determine whether a fall has occurred. However, such sensors are sensitive to noise information, prone to false alarms, and have high arrangement costs and low accuracy rates. The fall detection method based on computer vision detects whether a fall occurs by passively acquiring human motion information from the monitoring device and processing the acquired video or image. The computer vision-based method does not require the user to wear any equipment, which is a good user experience and high detection accuracy.

In the field of computer vision research, deep learning techniques such as convolutional neural networks (CNN) have been widely used in image classification. CNN are able to extract the information in images directly from a large amount of labeled data layer by layer, extract effective features of images and detect and classify images. The widely used network models for multi-target detection algorithms are Faster RCNN [4], YOLO [5] and so on. Faster RCNN algorithm uses RPN network to extract proposals. The essence of RPN method is to use sliding window mechanism. RPN sliding window on the feature map output from the last convolutional layer, each sliding window generates 9 anchors, due

ISSN 2616-5775 Vol. 5, Issue 4: 29-34, DOI: 10.25236/AJCIS.2022.050405

to the anchor mechanism and edge regression can get candidate regions with different scales and different aspect ratios, by non-maximum YOLO can predict multiple candidate frames and target classes at one time, and YOLO treats the target detection task as a regression problem, which truly achieves end-to-end target detection and can YOLO can predict multiple candidate frames and target classes at once. In this paper, we improve YOLOv5 to make it suitable for fall detection.

2. Introduction to the YOLOv5 network model

YOLOv5 can be structurally divided into 4 parts: input, trunk, neck and head of the trunk part, and its network structure is shown in Figure 1.

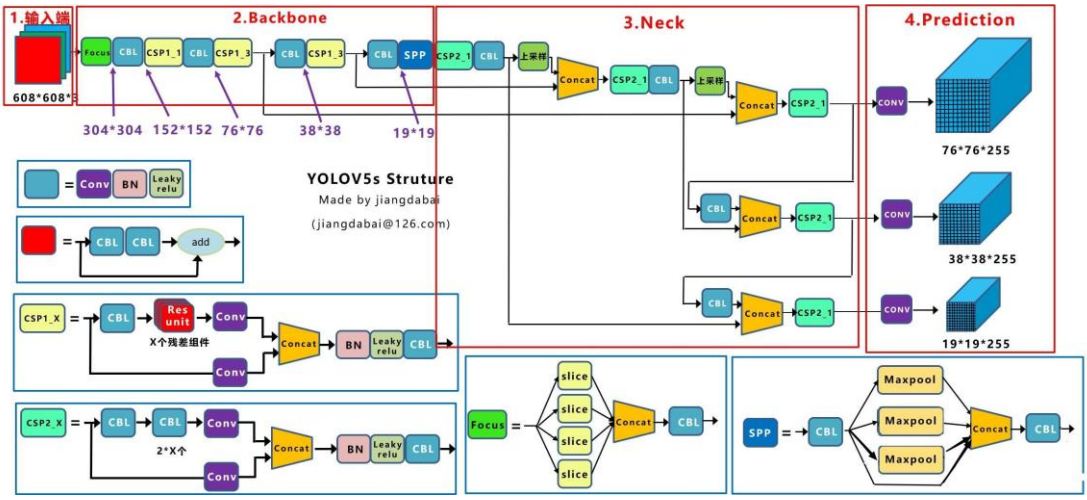

Figure 1: Network structure of yolov5

The input side mainly contains pre-processing of the data, including Mosaic data augmentation [6], and adaptive image padding [7]. Using the Mosaic data enhancement method, 4 images are randomly called with random size and distribution and stacked, which enriches the data and adds many small targets to improve the recognition of small objects and also makes the network more robust. Calculating 4 images at the same time is equivalent to increasing the size of the Mini-batch, which also reduces a lot of GPU memory consumption. In order to accommodate different datasets, YOLOv5 integrates adaptive anchor frame calculation on the Input side to automatically set the initial anchor frame size when changing datasets. The main chain utilizes bottleneck cross-stage local structure BottleneckCSP, which aims to reduce computation and improve inference speed, and spatial pyramid pooling SPP, which implements feature extraction at different scales for the same feature map and helps to improve detection accuracy. Neck network layer contains feature pyramid (FPN), path aggregation structure (PAN). FPN conveys semantic information top-down in the network, while PAN conveys localization information bottom-up, fusing information from different network layers in Backbone to further improve detection capability. head output, as the final detection part, mainly predicts targets of different sizes on feature maps of different sizes.

3. Detection algorithm and optimization

3.1. Model selection

Since most of the monitoring sites are embedded devices with low computing power, they cannot deploy detection models of larger scale. In order to reduce computing costs and enhance practicality, after comparing the performance of the four versions of YOLOv5, the smaller and faster model YOLOv5s model is chosen here in view of the speed requirements for fall detection.

3.2. Loss function selection

When the detection box B and ground truth G of IoU [8] is 0, Loss is 0 and the network cannot be trained. the purpose of GIoU Loss is to solve the problem in IoU Loss when$B$ and$G$ does not intersect, the Loss is 0.

ISSN 2616-5775 Vol. 5, Issue 4: 29-34, DOI: 10.25236/AJCIS.2022.050405

GIoU is defined as in Equation (1).

$$GIOU = IOU - \frac{Ac - U}{Ac} \tag{1}$$

Where, Ac represents the area of the smallest circumscribed rectangle of B and G, and U represents the area of the union of B and G. And GIoU Loss $= 1 -$ GIoU .GIoU Loss Although it solves the IoU Loss The problem that Loss is 0 in GIoU, but there are still some shortcomings. First, because GIoU is mainly converging Ac-U value, it is found that such convergence will cause the network to preferably expand the area of the bounding box to cover the ground truth instead of moving the position of the bounding box to cover the ground truth. The formula of DIoU Loss is as in Equation (2).

$$DIOU = IOU - \frac{\rho^2(b, b^{gt})}{c^2} \tag{2}$$

Among them, the ρ() is B is the Euclidean distance between the G the Euclidean distance between the center point, c is the B is the Euclidean distance between G the length of the diagonal of the minimum external moment. In contrast to GIoU, DIoU restricts not the minimum external moment with B and G The IoU element is also added to make the coverage area of the bounding box and the ground truth closer. In order to enhance the detection speed as much as possible, the DIOU_Loss is used for the bounding box loss function in this paper.

3.3. NMS non-extreme value suppression improvement [9]

In the prediction stage, redundant detection frames are usually removed using NMS. The criterion for judging is the intersection ratio IoU between a certain detection frame and the detection frame with the highest prediction score, and the predicted detection frame will be removed when the IoU is greater than a set threshold. In general scenes, this method is effective, but in pictures with dense targets, the detection frames of different targets are very close to each other with a large overlapping area due to the occlusion between the targets, so they will be incorrectly removed by NMS, resulting in target detection failure. In the surveillance video, the vehicle targets are concentrated in the middle of the road in the image, which is a more dense and easy to produce blocking scene, this paper uses DIoU as the judging criterion of NMS to improve this problem. DIoU considers the distance between the center points of two bounding boxes on the basis of IoU, and the definition of DIoU-NMS is shown in Equation (3):

$$s_i = \begin{cases} s_i, DIOU(M, B_i) < \varepsilon \\ 0, DIOU(M, B_i) \geq \varepsilon \end{cases} \tag{3}$$

M indicates a prediction box with the highest prediction score, and B_i indicates the prediction box that determines whether it needs to be removed, thesi denotes the classification score, and ε The DIoU-NMS considers IoU while judging the distance between two bounding boxes and the center point, and does not remove the prediction box when the distance is far, but considers that another target is detected, which helps to solve the problem of missing detection when the targets obscure each other. In this paper, we use DIoU-NMS to replace the original NMS.

4. Experiments and results

4.1. Data set

A dataset containing 2430 fall images is used in this experiment, and the dataset annotation information includes target type and location information. To ensure as much data for training as possible as well as the universality of the test set, the training and test sets are therefore divided in the ratio of 8:2.

4.2. Experimental environment and model training

The computing platform used for this training: the system configuration is NVIDIA GeForce GTX 1650 GPU, Intel(R) Core (TM) i7-9750H CPU, 16G RAM, OS internal version 19041.1415, CUDA 11.4, and python 3.8 language.

ISSN 2616-5775 Vol. 5, Issue 4: 29-34, DOI: 10.25236/AJCIS.2022.050405

The training parameters are set as follows: the input image size is 640*640, the learning rate is set to 0.02, the learning rate period is set to 0.25, the training batch is 16 groups, and the total number of training sessions is set to 200 in order to have the highest possible accuracy.

4.3. Analysis of experimental results

The following metrics are used to judge training [10]: precision (precision), recall (recall), and mean average precision (mAP, mean average precision). Precision is a measure of accuracy and indicates the proportion of examples classified as positive that are actually positive, as in Equation (4); recall is a measure of coverage and measures the number of positive examples classified as positive, as in Equation (5).

$$Precision = TP/(TP+FP)*100\% \tag{4}$$

$$Recall = TP/(TP+FN)*100\% \tag{5}$$

Where: TP denotes the number of predicted positive cases that are also actual positive cases, the FP denotes the number of predicted positive cases but actual negative cases, and FN denotes the number of predicted negative cases but actual positive cases.

The maximum precision, recall, mAP_0.5, and mAP_0.5:0.95 of the trained model can reach 0.89571, 0.84539, 0.90915and 0.7261, as shown in Figures 2 and 3.

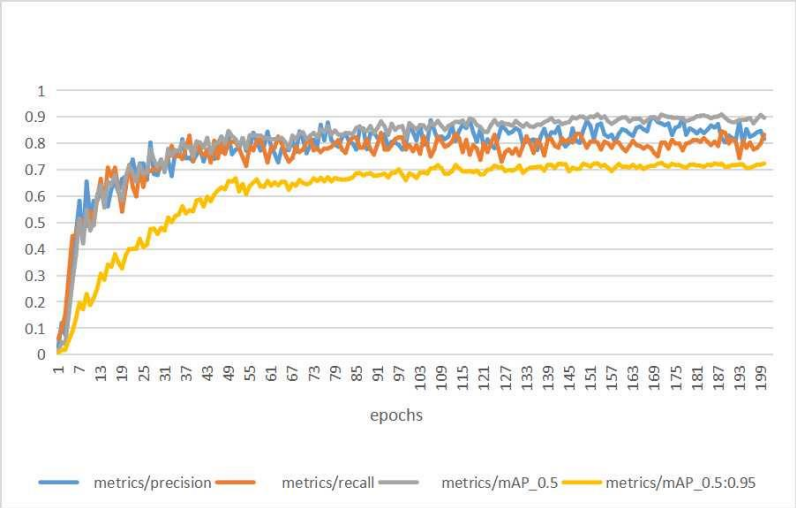

Figure 2: Training index chart

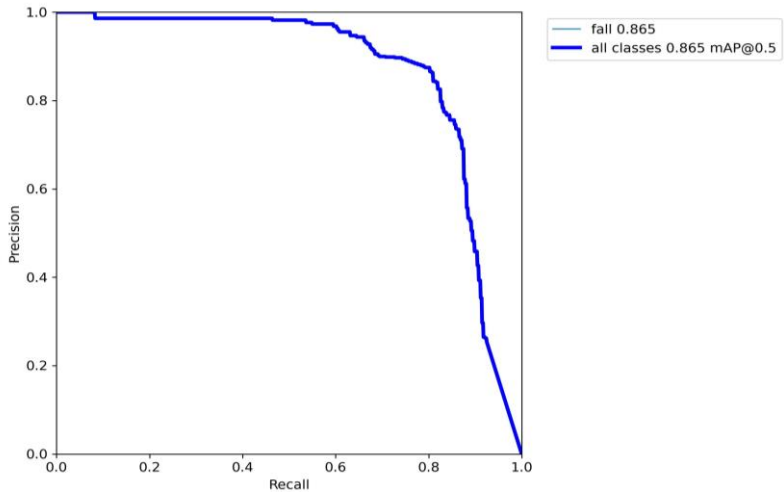

Figure 3: P-R curve

From Figure 2, we can understand that due to the size of the model, the training saturation is reached

ISSN 2616-5775 Vol. 5, Issue 4: 29-34, DOI: 10.25236/AJCIS.2022.050405

in about 50 epochs, and the training effect after that can be said to be negligible. In addition to MAP_0.5:0.95, the other items also reach the effect of 0.9, which is completely sufficient for fall detection.

Our detection also performs well in multi-target images due to the use to NMS non-maximal suppression improvement, as shown in Figure 4.

Figure 4: Multi-target detection effect diagram

Of course, our targets are more often placed on elderly people living alone or on some sparsely populated streets, where the speed and accuracy of single-target detection are perhaps more important, as shown in Figure 5.

Figure 5: Single target detection effect

The test took only 0.093 seconds for one picture, which can perfectly achieve the effect of real-time monitoring.

5. Conclusion

In this paper, we use yolov5s as the basic network model and use DIOU_loss and NMS non-maximum suppression improvement to improve the problem of multi-target occlusion and overlap such that detection fails. The mAP_0.50 of this method can reach 90.92%. For the detection of video streams, one frame detection takes only 0.093 seconds, which is a great advantage for the detection real-time, the method can be used for fall detection in public places. However, the improved detection algorithm in this paper has a lot of room for optimization, and in future research will continue to improve the model, enhance its sharing capability, improve its real-time performance, and also combine multiple factors such as SSD and Atrous filter [11], and replace the underlying network to improve the detection performance of the algorithm. In practice, we can lower the threshold and increase the detection of time intervals to decide whether to alert or not, or add facial expression recognition to achieve a double judgment criterion to achieve the hard requirement of accuracy.

References

[1] Lisa Ku, Suzhen Wang, Yiqian Chen, Chenlong Gao, Chunyu Hu, Xinlong Jiang, Zenyu Chen, Xingyu Xiao. A review of fall detection algorithms based on wearable devices [J]. Journal of Zhejiang University (Engineering Edition), 2018, 52(09): 1717-1728.

[2] Tu, Biqi. Research on fall detection algorithm for the elderly based on multi-sensor fusion [D]. Zhejiang University of Technology, 2017.

[3] Zhu Yan,Zhang Yaping,Li Shusheng,Li Weimin,Liu Yalu. Fall detection algorithm based on deep vision sensor and convolutional neural network [J]. Optical Technology, 2021, 47(01): 56-61. DOI: 10.13741/j.cnki.11-1879/o4.2021.01.011.

[4] Chen, Yijia. Faster RCNN-based target detection system [D]. Harbin Institute of Technology, 2019.

ISSN 2616-5775 Vol. 5, Issue 4: 29-34, DOI: 10.25236/AJCIS.2022.050405

[5] Ruan Qi Yang. Design and implementation of YOLO-based target detection algorithm [D]. Beijing University of Posts and Telecommunications, 2019.
[6] Yang Can. Research on traffic target detection method based on deep learning [D]. East China Jiaotong University, 2021. DOI: 10.27147/d.cnki.ghdju.2021.000076.
[7] Liu, Jingyi. Research on adaptive planar area geometry partitioning algorithm [D]. China University of Petroleum (East China), 2017.
[8] Chen Zhaofan,Zhao Chunyang,Li Bo. A border regression loss function to improve IoU loss [J]. Computer Application Research, 2020, 37(S2): 293-296.
[9] Li Yongshang, Ma Ronggui, Zhang Meiyue. Improving YOLOv5s+DeepSORT for monitoring video traffic statistics [J/OL]. Computer Engineering and Applications: 1-11 [2022-02-25].
[10] Wu Hongwei. Research on license plate detection and recognition system based on deep learning [D]. Dalian University of Technology, 2021. DOI: 10.26991/d.cnki.gdllu. 2021.003057.
[11] Wen Jie-Wen, Zhan Yin-Wei, Li Chu-Hong, Lu Jian-Biao. An Atrous filter design to enhance the detection capability of small targets in SSD [J]. Computer Application Research, 2019, 36(03): 861-865+872. DOI: 10.19734/j.issn.1001-3695.2017.09.0967.