Research on Pedestrian Fall Detection Based on YOLO

Qiu Zhenyu, Fan Xinwen, Chen Kaixuan

School of information, Xiamen University
Student ID: 23020211153961, 23020211153890, 23020211153919

Abstract

For the past few years, with the rapid development of computer science and technology, target detection has become one of the most practical researches in the field of computer vision. At the same time, the trend of aging population leads to frequent falls. Every security industry in society has an increasing demand for falls detection. In order to improve the detection accuracy and running speed of the model, a set of benchmark models is set up.Mosaic data enhancement and label smoothing are introduced into image preprocessing; Redisigning the Resblock module in the backbone network (CSPDarknet53);Improving the SPPNet(Spatial Pyramid Pooling Network) and the PANet(Path Aggregation Network) to make model perform better in fall detection. Kalman filter and Hungarian algorithm are introduced to track the prediction frame in Video Fall detection domain. The experimental results show that the method proposed in this paper is simpler, faster than the benchmark model network, and the predicted results are more in line with people's visual effects. The evaluation indexes in this paper, namely Mean Average Precision (MAP) and F1-Score, have achieved the better results than baseline.

Introduction

Pedestrian fall detection is a problem worthy of attention. The elderly are prone to fall events due to the decline of physical mechanism and function, the degradation of balance system, blurred vision and other reasons, especially in the unattended environment. If they can't be detected in time, it will become more and more serious. The consequences of physical injury and internal trauma caused by fall events will seriously affect the physical and mental health of the elderly who fall, and it will also increase family and social concerns. Therefore, it is a meaningful study to monitor the fall behavior of the elderly and make emergency measures in time. In addition, in the monitoring scene, timely detecting pedestrian falls and reminding relevant staff to deal with them can effectively reduce the consequences of pedestrian accidental falls, improve the service quality in relevant places, such as shopping malls and subways, better protect pedestrian travel safety and minimize the harm caused by emergencies.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

With the continuous development of deep learning and image processing technology in recent years, there are new solutions to the fall problem. Hardware driving and object detection algorithms complement each other, providing new decisions to solve the fall problem. 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, The life dynamics of people who are easy to fall can be captured in real time through HD camera, and can be stored in cloud video for statistical prediction, and this method is easy to install. The optimization of fall detection method from the algorithm level does not need a lot of investment cost on the basis of existing hardware. With the popularization of computer hardware, target detection technology also rises and is widely used in society, such as intelligent video surveillance, industrial detection, face recognition, unmanned driving and so on. However, due to the problems of occlusion, illumination change and scene clutter, pedestrian fall detection technology based on deep learning is still a challenge.

Based on the above problems, this paper proposes a fall detection model based on Yolo. This model is designed to extract the center adjustment parameters, width and height adjustment parameters, category score and foreground confidence of category objects in the original image, and mark the generation box in the original image. Specifically, it includes collecting and screening appropriate fall data sets, image preprocessing (occlusion, mosaic data enhancement, etc.), model design and optimization, model training and evaluation effect. The main contents are as follows:

- In terms of collection and screening of training sets, the fall detection data set of China Hualu cup is selected, which with a total of 4576 images. It includes a variety of elements such as streets, stadiums and interior, and has a certain standardization.
- The initial image is preprocessed with mosaic data enhancement and gray edge revision. The XML annotation will be automatically generated by setting the category mask through the algorithm, and the annotation conversion code will be written to convert it into a TXT format

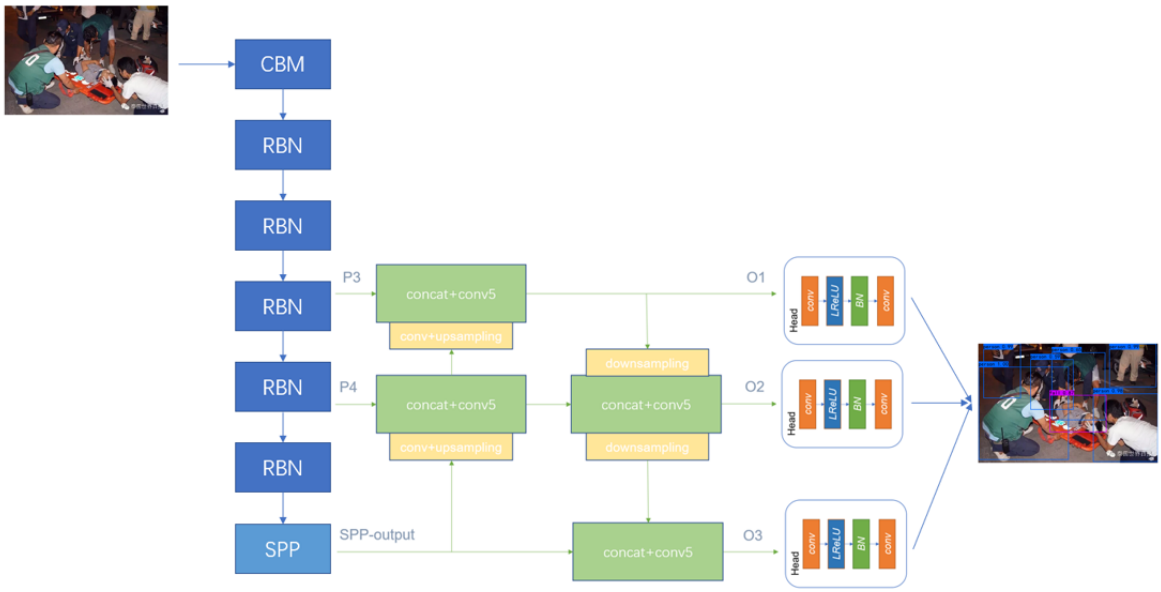

Figure 1: overall model architecture

label file.

- Redesign PAN (Liu et al. 2018), SPPNet (He et al. 2015), backbone network and other network layers, and simplify and improve the old model.
- Train the model, adjust the parameters, and post process the result box to find the best performance. The evaluation includes the index mean precision (MAP) and F1 value, which focuses on whether the generated result box is more in line with the human eye.

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 2014). Mostapha et al. considered the hardware embedded into the sole to reduce intrusiveness constraints(Zitouni et al. 2019), which improved the acceleration and timeliness. 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. 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). 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.

With the rapid development of computer hardware, large-scale data training becomes possible. 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.

Method

Overall architecture

The input image (video frame) enters the backbone feature extraction network. Some features pass through SPPNet and some features flow into PANet. Finally, feature integration, loss calculation and target detection are carried out through the head module. The overall model architecture is shown in Figure 1.

Backbone network module

After the image is input into the model, it firstly enters the backbone feature extraction network, as shown in Figure 2. The residual network module is defined as RBN, the basic convolution block is CBM (Conv-BN-Mish), and the basic residual block is called RB. With the stacking of RBN modules, the width and height of the feature map obtained by the network are constantly compressed, and the number of channels is constantly expanding. The multi-dimensional features extracted by three deep RBN modules are mainly obtained. The whole backbone feature extraction network is obtained by stacking RBN modules. It is a convolution block built by residual network. RBN module extracts shallow feature information through two high-span residual edges under different normalization, and stacks it on the channel with the feature map output by RB module. In RB module, N is the number of convolution layers, N belong to 1,2,8,8,4. In the CBM module, the Mish activation function is used to replace the traditional ReLU activation function in order to improve the accuracy and normalization of the network layer.

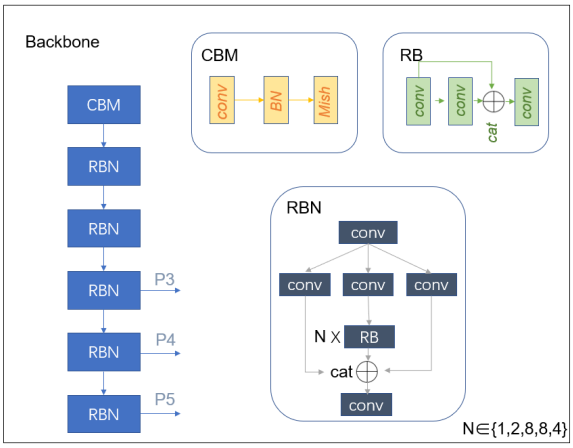

Figure 2: backbone module structure

Neck Module

The neck module is used to connect the backbone for feature extraction and the head module for object detection. SPPNet (spatial pyramid pooling net) and PAN (path aggregation network) are selected as the Neck module of the network.

Spatial Pyramid Pooling Net (SPPNet) In this experiment, the pooling layer of spatial pyramid pooling(SPP) is selected as the first part of the neck module,this structure was modified to retain the output space size, and five sliding check inputs with different scales were set the feature to be max pooled. SPPNet is shown in Figure 3, in which P5 comes from part of the output of the backbone network. After the characteristic channels of the multi pooling layer are combined, the characteristic map with the same size as the original map is output.

Path aggregation net(PAN) In this experiment, the path aggregation model is shown in Figure 4. The lower sampling

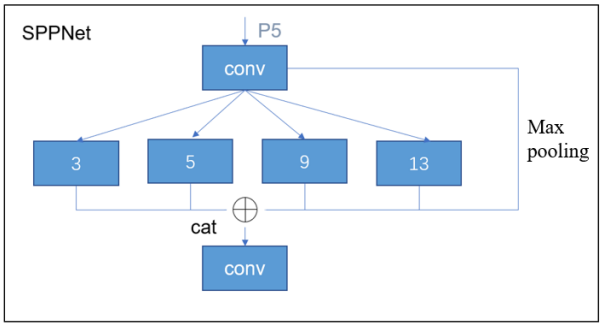

Figure 3: structure diagram of spatial pyramid module

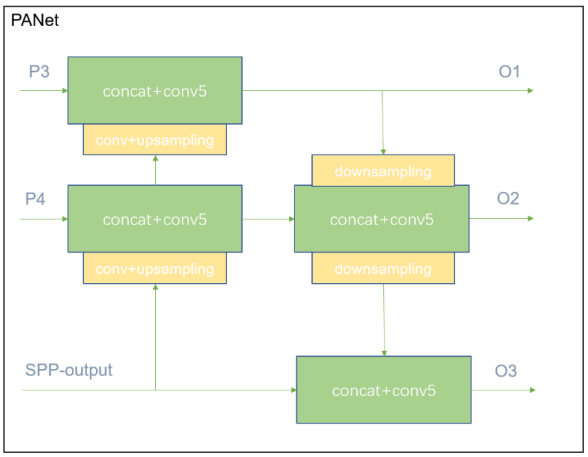

Figure 4: structure diagram of panet module

of the feature map is realized through convolution, the transposed convolution layer is used to replace the upper sampling layer, more learnable parameters are introduced to improve the accuracy of the model, the conv5 structure is modified into a residual block, and 1x1 convolution is applied to this module by controlling the dimension of the middle layer, The function is to splice the features of each pixel on different channels to retain the size of the original feature map. P3 and P4 in Figure 4 are the outputs of the backbone network. After passing through the backbone module, part of the input image flows into SPPNet, then enters PANet for bidirectional feature extraction, and finally outputs three effective feature layers of different scales: O1, O2 and O3.

Head Module

The head module is the output module of the detection network. As shown in Figure 5, the output of the receiving path aggregation module is used as the input of the head module, the features are integrated through the convolution layer with the convolution core size of 3X3, and the integrated features are channel converted through the convolution layer with the convolution core size of 1x1.

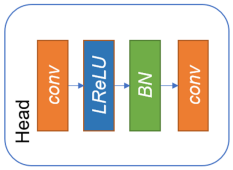

Figure 5: head module structure

Loss introduction

CIOU Loss Mask is defined as the pixel matrix whose intersection over Union ratio between the target frame and the bounding box is greater than the threshold at different scales. Since the input image is mainly the background, that is, most of the negative samples, if all the negative samples participate in the calculation, the loss of the negative samples will be greatly amplified, resulting in the training results biased towards the negative samples. In this experiment, Mask is used to control the sample training,defines the pixels not marked by Mask and ignored by NMS as NMask. CIOU formula is shown in equation (1)

$$CIOU = Mask \times (IOU - \frac{\rho^2(b, b^{gt})}{c^2}) \tag{1}$$

In the above equation, Represents the Euclidean distance between the center points of the prediction Box and the target Box and represents the diagonal distance of the minimum closure area that can include both the prediction frame and the target Box.a an v shows in equations (2) and (3):

$$\alpha = \frac{v}{1 - IOU + v} \tag{2}$$

$$v = \frac{4}{n^2}(arctan\frac{w^{gt}}{h^{gt}} - argtan\frac{w}{h})^2 \tag{3}$$

The width and the height ratio of target bounding box as shows , w/h shows the width height ratio of the prediction bounding box,and we get the CIOU Loss Function as the equation (4):

$$CIOULoss = 1 - Mask \times (IOU - \frac{\rho^2(b, b^{gt})}{e^2} - av) \tag{4}$$

Confidence Loss The confidence loss is shown in equation (5), where Conf represents the confidence value output by the head module.

$$\begin{aligned} ConfLoss = Mask \times BCELoss(Mask, Conf) \\ +NMask \times BCELoss(Mask, Conf) \end{aligned} \tag{5}$$

BCELoss shows in the equation (6):

$$BCELoss(x, y) = -\frac{1}{n}\sum y \times lnx + (1 - y) \times ln(1 - x) \tag{6}$$

Label Loss We use the BCELoss to calculate the Loss of the classification,as shown in the equation (7):

$$ClassLoss = Mask \times BCELoss(Label, Pred) \tag{7}$$

Total Loss Through the weighted combination of the above methods,we show the total loss in equation (8):

$$\begin{aligned} TotalLoss = W_{CIOU} \times CIOULoss \\ +W_{Conf} \times ConfLoss + W_{class} \times ClassLoss \end{aligned} \tag{8}$$

Experiments

Datasets and Experimental Settings

We use 4576 fall images provided by China Hualu Cup algorithm competition as the data set, in which 3660 images are randomly selected as the training set, 458 images as the verification set and 458 images as the test set. The main structure of YOLO is adopted as the baseline, and the number of iterations is set to 100, including 50 training times for frozen trunk feature extraction network and 50 training times after thawing, so as to speed up the training process. The input image size is normalized to 416*416 to reduce the demand for GPU memory, and the backbone feature extraction network adopts CSPDarkNet53, Neck module adopts PANet and SPPNet, and the output module adopts YOLO head structure.

Tuning strategy

Priori boxes generation by K-means clustering. The traditional clustering algorithm uses the Euclidean distance to measure the difference between target clusters, but the error increases with the size. Therefore, in this experiment, the width and height set corresponding to the training set is extracted and the Jaccard similarity coefficient is used for clustering evaluation. That is, the IOU value between the bounding boxes is used as the clustering standard, and finally nine priori boxes' widths and heights under three different scales are obtained, which is used as the benchmark to fit the target box through the training model.

Network structure adjustment. We introduce the CSPResNet module of the backbone network CSPDarkNet53 into an asymmetric high-span residual edge of the BN layer for channel fusion to improve the robustness of the network. Neck module introduces a smaller feature kernel in SPPNet structure to extract fine-grained features. The residual connection is made for the multiple continuous convolutions in the PANet structure, so that the deep network can extract the shallow image feature information. In the Upsample module, the upper sampling layer is removed, the transposed convolution module with learnable parameters is used to replace the bilinear interpolation, and the Mish activation function in the Neck model is replaced by the Leaky ReLU function to improve the accuracy and speed of the model.

Data set expansion. The label files in the data set of China Hualu Cup account for a relatively small proportion, so we remove the image data that does not contain the target category, and converte the foreign LFD fall video data set into the image frame category label in VOC format. Finally, automatically label the non fall category label through the algorithm, and merge them. Now there are 7627 images in the data set.

Figure 6: Fall detection results of images and video frames.

Experiment Results and Analyses

The performance comparison among the above strategies and the baseline on our data set is summarized in Table 1. For baseline, clustering is used to screen the initial width and height, which improves the effect of the model to a certain extent. By adjusting the trunk feature extraction network, the SPPNet structure of Neck module, the continuous convolution layer and upper sampling layer of PANet path aggregation module, the evaluation metrics of the fine tuned YOLO model performs better than the original model structure through fewer training parameters.

	MAP/%	F1(average)
Baseline	49.39	0.490
Improved clustering	50.36	0.500
Improved backbone	51.20	0.513
Improved SPPNet	52.27	0.530
Improved PANet	53.55	0.570
Data set expansion	77.20	0.743

Table 1: Performance(MAP and average F1) of baseline and improved approaches.

In Table 2 We can see that our improved model removes 40MB memory occupation compared with baseline, making it faster and achieving better results in evaluation metrics. With the expansion of LFD fall data set, the model can extract richer hidden layer features from diverse fall styles, making it more robust and its generalization effect better.

	FPS	MAP/%	Model Size/M
Baseline	18	49.39	250
Ous Model	30	77.20	210

Table 2: Accuracy, speed and memory comparison of baseline and our model.

Fall recognition for video

Traditional SORT algorithm calculates the IOU matrix for the boundary boxes detected in each frame and the boxes predicted by the target tracking algorithm, and associates the inter frame ID through the Hungarian algorithm. In this experiment we use DeepSORT algorithm, which adds the target features extracted by convolution operation to inter frame matching, reducing the redundant operation of ID switch while dealing with occlusion(figure 7).

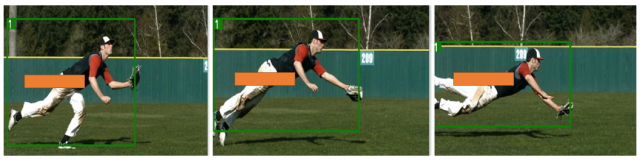

Figure 7: Results of fall detection under partial occlusion for video.

We randomly choose some fall images and videos from the Internet, and their detection effect are shown in Figure 6, which are effective.

Conclusion

In this paper, the Yolo network structure is improved to detect pedestrian falls. In the data preprocessing module, mosaic data enhancement and label smoothing are performed on the original data set, and the RBN basic module in the backbone feature extraction network is modified, another high-span different normalized residual edge is introduced to fuse the characteristics of the shallow layer in the deep layer in different ways. In the post-processing, the boundary boxes with high coincidence degree are discarded by calculating the IOU matrix of the two types of prediction frames. Target detection is carried out on the fall data sets of LFD and China Hualu cup, and the evaluation of the effect of different module adjustments on the model is analyzed. Through the above adjustments, the improved model has a better effect on fall detection, and the evaluation index is improved for the test set.

References

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*, 8(6): 1129–1138.

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.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9): 1904–1916.

Karantonis, D. M.; Narayanan, M. R.; Mathie, M.; Lovell, N. H.; and Celler, B. G. 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.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759–8768.

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.

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. IEEE.

Zitouni, M.; Pan, Q.; Brulin, D.; Campo, E.; et al. 2019. Design of a smart sole with advanced fall detection algorithm. *Journal of Sensor Technology*, 9(04): 71.