Journal of New Media
DOI:10.32604/jnm.2021.017763
Article

Elderly Fall Detection Based on Improved SSD Algorithm

Jiancheng Zou[1], Na Zhu[1,*], Bailin Ge[1] and Don Hong[2]

[1]North China University of Technology, Beijing, 100144, China
[2]Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132, USA
[*]Corresponding Author: Na Zhu. Email: zhuna12340204@163.com
Received: 10 February 2021; Accepted: 02 March 2021

Abstract: We propose an improved a single-shot detector (SSD) algorithm to detect falls of the elderly. The VGG16 network part of the SSD network is replaced with the MobilenetV2 network. At the same time, we change the infrastructure of MobilenetV2 network, the three layers that were not down-sampled at the end were removed, which can make the model structure simpler and faster to detect. The complete Intersection-over-Union (CIoU) loss function is introduced to get a good regression of the target borders that have different sizes and different proportions. We use Feature Pyramid Network (FPN) for up-sampling, it can fuse low-level feature maps with high resolution and high-level feature maps with rich semantic information. For sampling results, we use the Secure Shell (SSH) module to extract different receptive fields, which improves the detection accuracy. Our model ensures that the accuracy of the elderly fall detection remains unchanged, but it greatly improves the detection speed that only takes 10 milliseconds to detect a picture.

Keywords: SSD algorithm; MobileNetV2 network; fall detection

1 Introduction

As the elderly's body functions continue to degenerate, the selfcare ability gets weaker and often becomes easier to fall accidentally in daily life. It could get seriously injured, cause disability, and even endanger their lives. Therefore, the study of fall detection problems for the elderly has become particularly important.

The current methods of fall detection for the elderly can be summarized into three types: The first one is based on wearable devices [1]. Almeida [2] and others designed a crutch with sensors, the angular velocity of the crutches is measured to determine whether the crutches drop on the ground, so as to infer whether the elderly has fallen. Shahiduzzaman et al. [3] proposed a cloud network edge architecture and installed it in a helmet to obtain a smart helmet, the purpose of this is to enhance the effect of fall detection. Peng [4] and others designed a belt that uses an ARM microcontroller as the main control chip and uses the ADXL345 [5] three-axis acceleration sensor to detect falls and alarm. Jiang [6] and others designed a smart watch monitoring device for the elderly, which can carry out GPS positioning for the elderly and notify the family members in time after fall detection. Method of this type performs fall detection by wearing a detection device on the elderly. The wearable device has drawbacks such as it must be worn continuously, needs to charge the battery, maybe forgotten by the user, and prevents being used due to loss of consciousness after falling among others.

The second method of fall detection is based on a scene equipment. Alwan et al. [7] monitored the vibration patterns on the floor to detect person's falls. Haider et al. [8] used low-cost UWB radars, housed the radar in a 14 cm × 7 cm × 1 cm box, and placed it in a 3 m × 5 m room at a height of 1.7 m to detect falls. Ogawa et al. [9] applied infrared array sensors to detect falls that the temperature distribution can be

obtained from the top, and fall detection can be performed according to temperature changes. Method of this type doesn't require wearing the device, but it is problematic to determine whether the information collected by the sensor belongs to the user, and the high cost for installation of such a system can be an affordable issue for average families.

The third method of fall detection is based on computer vision, Hsu et al. [10] used variable triangles to divide the shape of a human body, extracted two posture features of the bone and the center, and then determined a suitable algorithm for human posture recognition. In view of the head movement obviously when the human body falls, and the head is generally within the visible range, Rougier et al. [11] proposed a monocular camera to obtain the 3D motion trajectory of the human head and use it as the basis for human fall detection. First, the human head is detected and tracked, then the 3D motion trajectory of the head is extracted and the human body fall detection is performed according to the obtained speed characteristics. Yu et al. [12] applied multiple cameras to detect the segmented foreground and construct its 3D features. Compared with the traditional OCSVM (a kind of SVM), not only the classifier training is carried out with the corresponding target samples of the human body falls, but also some non-fall samples with more accurate decision boundaries are used to enhance training. Chen et al. [13] extracted the feature points of human bones by using the OpenPose human pose recognition algorithm, judging the possibility of the elderly falls according to the movement of the neck characteristic points and the analysis of the semantic information of the scene autonomously. Zhou et al. [14] proposed an algorithm of human fall detection based on depth images by using the deformable spring multi-component model of the human body, and the detection is performed by analyzing the posture of the human body. This method is currently more popular, and it is also the method that this article focuses on the development of algorithms that are computer vision-based with features of simplicity in operation and it can not only reduce the inconvenience caused by wearable devices, but also reduce the cost.

In this paper, we adopt a computer vision detection method and design a lightweight network model on the basis of the Single Shot (MultiBox) Detector (SSD) network to detect the fall of the elderly. We used improved MobileNetV2 algorithm for feature extraction and change the basic network architecture. At the same time, the smoothL1 Loss function is replaced with the complete Intersection-over-Union (CIoU) loss function, in order to perform a good regression of the target borders which have different sizes and different proportions. This paper uses Feature Pyramid Network (FPN) for up-sampling, it combines high-level features with low-level features, and uses the Secure Shell (SSH) module to extract different receptive fields from the sampling results, which improves the detection accuracy. The algorithm can re-analyze the data and decide to only use the feature maps of the last two stages for prediction that improves effectiveness of the detection significantly.

2 SSD Algorithm

2.1 SSD Network

The SSD network was proposed by Wei Liu in ECCV 2016 [15]. SSD network returns to target category and location directly. It predicts on feature images of different scales and adopts end-to-end training to ensure the accuracy of detection even if the resolution of the image is relatively low. It is based on the Visual Geometry Group16 (VGG16) network [16], and the model structure is shown in Fig. 1. SSD network composed of a front end and a back end. The front end is a VGG16 network, which runs through to the Conv5_3 layer, and the back end is a feature layer network of different scales. The following is the difference between SSD network and VGG16 network:

1) The size of the input picture is different, the size of the input picture of the SSD network is 300 × 300 × 3, and the size of the input picture of the VGG-16 network is 224 × 224 × 3;
2) The FC6 and FC7 layers in the VGG16 fully connected layer are replaced with 3 × 3 convolution layers Conv6 and 1 × 1 convolution layers Conv7, the FC8 layer is removed, 4 sets of convolution layers are added at the same time. The group first uses a 1 × 1 convolution kernel to reduce the channel, and then uses a 3 × 3 convolution kernel to downscale and

increase the channel;

3) The fifth pooling layer (FC5) of the VGG16 network has a size of 2 × 2 and a step size of 2, which is changed to a size of 3 × 3 and a step size of 1.

The SSD network consists of 6 feature layers of different scales, they are Conv4_3, Conv7, Conv8_2, Conv9_2, Conv10_2, Conv11_2, the corresponding sizes are 38 × 38 × 512, 19 × 19 × 1024, 10 × 10 × 512, 5 × 5 × 256, 3 × 3 × 256, 1 × 1 × 256. Because the shallow feature layer is more suitable for detecting small targets and the deep feature layer is more suitable for detecting large targets, the multi-scale detection network needs to perform regression calculation and non-maximum suppression on the feature layers of different depths, and outputs the final result [17–18].

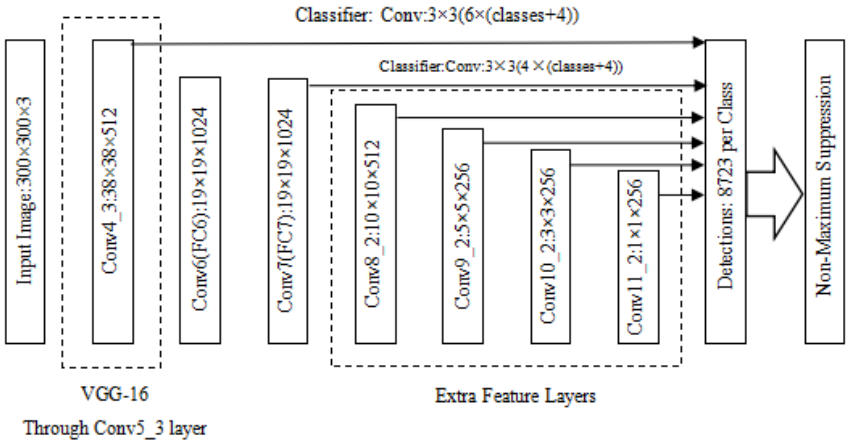

Figure 1: SSD network structure

2.2 Encoding of the Border

The regression goal in SSD is not simple deviation of center point and scaling of width and height. It involves a process of encoding and decoding. Encoding maps the ground truth box to the output space of the SSD which serves as the label information of the SSD. The reason for this is to solve the loss function more conveniently. Decoding is to convert the information of the priori box into the data of the ground truth box. We define the position of the input a priori box as $d=\left[d^x,\ d^y,\ d^w,\ d^h\right]$, the position of the ground truth box is $g=\left[g^x,\ g^y,\ g^w,\ g^h\right]$ and the coding coefficient is $[0.1,\ 0.1,\ 0.2,\ 0.2]$. Then there is the following coding process:

$$\hat{g}^x=\frac{g^x-d^x}{0.1d^w} \tag{1}$$

$$\hat{g}^y=\frac{g^y-d^y}{0.1d^h} \tag{2}$$

$$\hat{g}^w=\frac{\log(\frac{g^w}{d^w})}{0.2} \tag{3}$$

$$\hat{g}^h=\frac{\log(\frac{g^h}{d^h})}{0.2} \tag{4}$$

2.3 Loss Function

The loss function of SSD [19] is shown in Eq. (5).

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \tag{5}$$

where N is the number of positive samples of the prior box, c is the predicted value of category confidence, l is the position of the prior box, g is the position of the ground truth box, α is the weight of the both.

The loss function consists of two parts, one is the category loss and the other is the location loss. The category loss is defined as:

$$L_{conf}(x,c)=-\sum_{i\in Pos}^{N}x_{ij}^{p}\log(\hat{c}_i^{p})-\sum_{i\in Neg}\log(\hat{c}_i^{0}) \tag{6}$$

where $\hat{c}_i^{p}=\frac{\exp(c_i^{p})}{\sum_{p}\exp(c_i^{p})}$, $\hat{c}_i^{p}$ is the probability that the predicted i-th prediction box matches the j-th ground truth box with respect to the category $x_{ij}^{p}=\{0,1\}$, which is equivalent to an indicator. When $x_{ij}^{p}=1$, it means that the i-th predicted box for category p matches the j-th ground truth box. $\hat{c}_i^{0}$ refers to the probability that the i-th prediction box predicted by the network matches the background.

The location loss is defined as:

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}smooth_{L1}(l_i^{m}-\hat{g}_j^{m}) \tag{7}$$

where $smooth_{L1}(x)=\begin{cases}0.5x^2,if|x|<1\\|x|-0.5,otherwise\end{cases}$, l_i^{m} is the position of the prediction box, $\hat{g}_j^{m}$ is the position of the ground truth box.

3 MobileNetV2 Network

The MobilenetV2 [20] network was proposed by Google. This network introduces a depthwise separable convolution and inverted residual block, it replaces the last ReLU6 with a linear activation function in the inverted residual network structure to solve the problem of information loss caused by the low-dimensional ReLU6 operations [21].

3.1 Depthwise Separable Convolution

MobileNetV2 is based on depthwise separable convolution which decomposes the standard convolution into depthwise convolution and pointwise convolution. Comparing with the standard convolution, the depthwise separable convolution can reduce the amount of calculation and improve the speed to a certain extent.

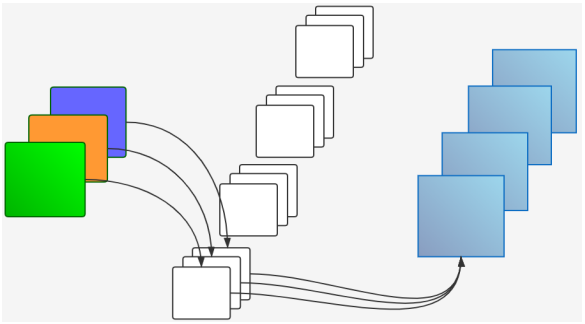

Figure 2: Standard convolution structure

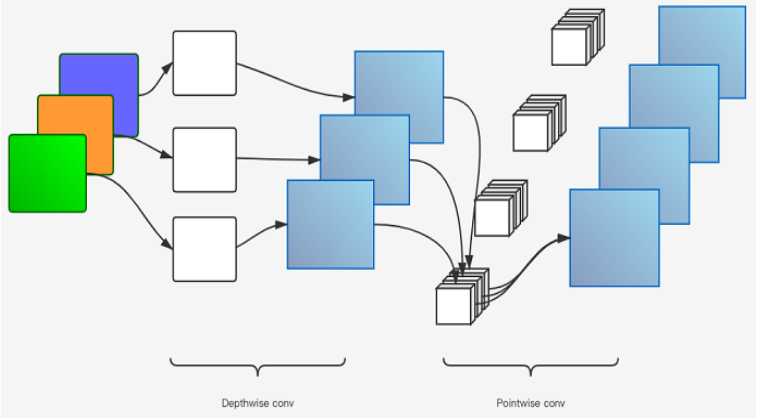

Figure 3: Depthwise separable convolution structure

We suppose that the model input feature image size is $D_F \times D_F$, the number of input channels is M, the number of output channels is N, and convolving with a convolution kernel of size $D_k \times D_k$. Fig. 2 shows the standard convolution structure, the amount of calculation is L_t, which can be calculated as follows:

$$L_t = D_F \times D_F \times M \times D_k \times D_k \times N \tag{8}$$

The depthwise separable convolution structure is shown in Fig. 3. The amount of calculation consists of two parts: depthwise convolution L_d and pointwise convolution L_p, which can be calculated as follows:

$$L_d = D_k \times D_k \times M \times D_F \times D_F \tag{9}$$

$$L_p = M \times N \times D_F \times D_F \tag{10}$$

The two parts of parameter calculations are summarized to obtain the calculations of the depthwise separable convolution structure L_s, which can be calculated as follows:

$$L_s = L_d + L_P = D_k \times D_k \times M \times D_F \times D_F + M \times N \times D_F \times D_F \tag{11}$$

Comparing the calculations of the depthwise separable convolution and the standard convolution, we can obtain:

$$\frac{L_s}{L_t} = \frac{1}{N} + \frac{1}{D_k^{\,2}} \approx \frac{1}{D_k^{\,2}} \tag{12}$$

It can be seen from Eq. (12) that the number of calculation amount of the depthwise separable convolution is greatly reduced compared with the standard convolution. It can be approximated that the computation of the deep separable convolution structure is D_k^2 times less than that of the standard convolution structure. The operation efficiency of the model is improved.

3.2 Inverted Residual Block

In order to play the role of depthwise separable convolution efficiently, MobileNetV2 uses an inverted residual block with a linear bottleneck innovatively [22], it is the opposite of the residual block. The standard residual block first performs 1 × 1 convolution to compress the feature map which reduces the number of channels of the feature map, then performs 3 × 3 convolution, and finally increases the dimension of the feature map through 1 × 1 convolution.

However, the inverted residual block first uses 1 × 1 convolution to increase the dimension of the feature map. Then using 3 × 3 depthwise convolution and finally reducing the dimensionality of the feature map through 1 × 1 convolution. The purpose of this is to increase the features that can be extracted, because MobileNetV2 uses depthwise separable convolution, it has fewer parameters, and the

extracted features will be relatively few. If it is compressed again, which will result in fewer features that can be extracted.

The ReLU6 activation function is used in the inverted residual block, but the ReLU6 needs to be removed in the last layer to use the linear activation function. The features obtained by the depthwise separable convolution correspond to the low-dimensional space and have fewer features. If linear mapping is followed, most features can be retained, while nonlinear mapping, such as ReLU6, will destroy features, cause loss of features, and make the model effect worse.

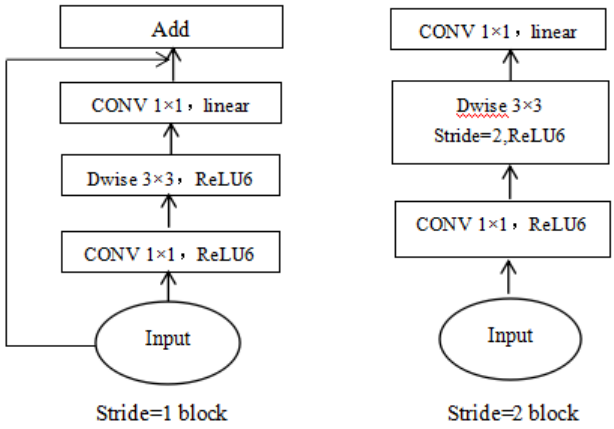

Figure 4: Inverted residual block

4 Model Optimization

Our article replaces the VGG16 network part of the SSD network with the MobileNetV2 network [23], and makes the following improvements on this basis.

4.1 Change Infrastructure of MobileNetV2

Our article changes the basic architecture of MobileNetV2. Considering that the last three layers of the MobileNetV2 network are not down-sampled, there is no larger range of acquisition for the receptive field, but the number of channels is significantly increased, which increases the computational complexity of the model. In order to make the model structure simpler and more effective, the last three layers were removed.

The data is analyzed, as shown in Fig. 5, which is the distribution map of all target sizes. The abscissa is the width w, and the ordinate is the height h. It can be seen that the width and height of the detection target are concentrated between 20 and 160, and there is no particularly large fluctuation. Because it is a monitoring deployment, there is no situation where the target is very small or occupying the entire picture, and the size of the low-level feature map is larger, and more useless frames are generated, so the low-level feature map is not predicted, and select the last two layers of feature maps to predict, which can reduce the generation of more useless frames, increase the proportion of positive samples, and alleviate the imbalance between positive and negative samples.

At the same time, we use FPN [24–25] for up-sampling, it combines low-level features with high-level features. This approach optimizes the problems of low-level features with high resolution but weak semantic information and high-level features with low resolution but strong semantic information. Using the SSH [26] module to extract different receptive fields for the sampling results.

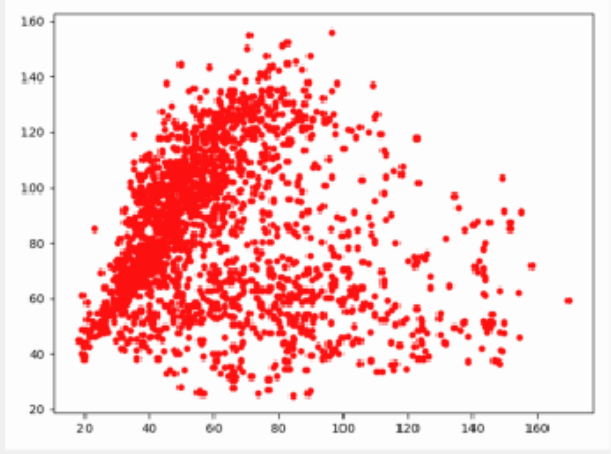

Figure 5: Dimensional drawing of detection target

Our model is shown in Fig. 6.

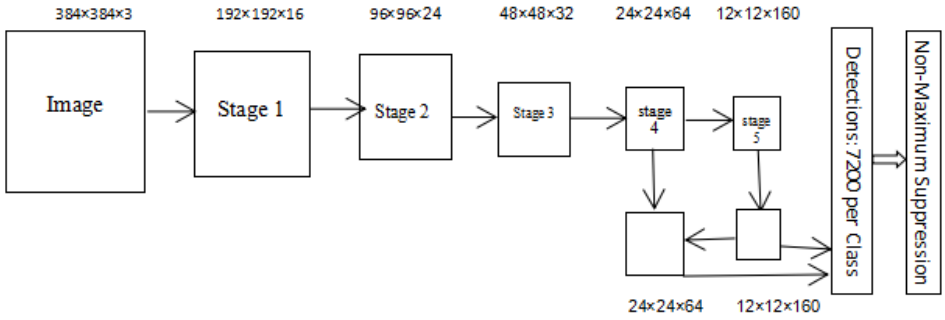

Figure 6: Our model structure

The size of the picture that we input is 384 × 384, then using five convolutions for down-sampling and the last two layers are up-sampled using FPN, finally, the up-sampled feature map is predicted. The detection network needs to perform regression calculation and non-maximum suppression on the feature layers of last two layers. Next our model will output the final result.

The size of the anchor is the first layer (32, 64), the second layer (28, 16). We use different scale values $[1,2,3,1/2,1/3]$ for calculations, 10 anchors are generated at each position of the prediction feature layer, which takes care of various sizes of sitting, lying, standing, and falling effectively. We can calculate that our model has a total of 7200 anchors of different sizes and positions, which is much less than the SSD model that inputs pictures of the same size.

4.2 CIoU Loss Function

Our article replaces smoothL1 loss function with CIoU [27] loss function.

Multiple detection frames may have the same smoothL1 Loss, but the IoU may be very different. Based on this, IoU Loss is proposed. However, IoU Loss cannot optimize the situation where the two boxes don't overlap, nor can it reflect how the two boxes intersect when the IoU value is the same. In order to determine the relative position of the prior box and the ground truth box, GIoU Loss is proposed. But when the ground truth box completely wraps the prior box, GIoU will degenerate into IoU, and its relative position relationship cannot be distinguished, so DIoU Loss is proposed, which can better optimize this kind of problem by adding the normalized distance of the center point. However, DIoU Loss only considers the overlap area and center point distance, but does not consider the aspect ratio. Therefore, CIoU Loss is further proposed on the basis of DIoU. It has scale invariance. Among targets of different

scales, their loss values have the same proportions, which is better for small object detection, and the detection effect of the detector is more balanced. Thus, our paper chooses CIoU Loss as the loss function, which will make the position detection more accurate. See Eq. (13).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon \tag{13}$$

where $IoU = \frac{B \cap B^{gt}}{B \cup B^{gt}}$, B is the prediction box, B^{gt} is the target box, $\rho(\cdot)$ is the Euclidean distance, b and b^{gt} are the center points of B and B^{gt} respectively. c is the diagonal distance between the smallest outer rectangle of B and B^{gt}, α is the parameter used for trade-off, $\alpha = \frac{\upsilon}{(1 - IoU) + \upsilon}$, υ is a parameter used to measure the consistency of the aspect ratio, $\upsilon = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2$.

5 Experiment and Analysis

5.1 Data Set

Our article uses the public data set Multiple cameras fall dataset for fall detection. We process the falling video in the data set and divide the video into frame-by-frame pictures; then dividing the training samples into four postures: Standing, lying, sitting, and falling. We prepare 1851 images for our experiment, including 1461 as the training set and 390 as the test set. Among them, there are 855 samples of standing posture in the training set, 158 samples of sitting posture, 240 samples of lying postures, and 208 samples of falling posture. In the testing set, there are 250 samples of standing posture, 44 samples of sitting posture, 52samples of lying postures, and 44 samples of falling posture, as shown in Tab. 1.

Table 1: Experimental data description

	Sitting	Standing	Falling	Lying
Training set	158	855	208	240
Test set	44	250	44	52

5.2 Experiment Setup

We train with an initial learning rate of 0.1 and use the cosine decay function for learning and setting the size of batch is 32. We input the size of the image is 384 × 384. In order to enhance the image, we perform random horizontal flip, random crop and random color dithering on the image. We input the processed image while keeping the original ratio, and then fill the edge with 0 for a total of 100 epochs.

5.3 Experimental Results

Tab. 2. shows experimental results of fall detection.

Table 2: Experimental results

	Standing	Sitting	Falling	Lying	mAP	mAP per object
AP	0.978	0.996	0.749	0.964	0.922	0.951

The evaluation index uses AP (Average Precision) that means average precision and mAP (Mean Precision) which is the average AP value of multiple verification set individuals. AP is used in the field of target detection commonly. The AP values of standing, sitting, lying, and falling are 0.978, 0.996, 0.964, 0.749, respectively, their mAP value is 0.992, the mAP value of each target is 0.951. When the

confidence threshold is equal to 0.6, we have $F_1Score = 0.86229$.

Our model ensures that the accuracy of the elderly fall detection remains unchanged, but it greatly improves the detection speed that can be up to 10 milliseconds per image. The following are some diagrams of the result.

Figure 7: Experimental results

6 Conclusion

We propose an improved SSD network model to study the problem of elderly fall detection. In order to improve the detection speed and reduce the amount of parameters, the lightweight MobileNetV2 network is introduced to extract features. At the same time, we changed the network architecture and optimized the loss function. By analyzing the data, we choose to use FPN to sample the last two feature layers and extract features. And the SSH module is used to extract the sampling results with different receptive fields. Our model reduces a large number of parameters and greatly improves the detection speed.

Funding Statement: The work of this paper is supported by the National Natural Science Foundation of China under Grant No. 61572038, and the Innovation Project Foundation NCUT.

Conflicts of Interest: We have no conflicts of interest to report regarding the present study.

References

[1] H. Xie, H. C. Jia, M. J. Mao and R. Chen, "A method of comprehensive judging pedestrian fall detection based on KCF conditions," *Internet of Things Technology*, vol. 10, no. 9, pp. 27–30, 2020.

[2] O. Almeida, M. Zhang and J. Liu, "Dynamic fall detection and pace measurement in walking sticks," in *Proc. HCMDSS-MDPnP*, Boston, MA, pp. 204–206, 2007.

[3] K. M. Shahiduzzaman, X. Hei, C. Guo and W. Cheng, "Enhancing fall detection for elderly with smart helmet in a cloud-network-edge architecture," in *Proc. ICCE-TW*, Yilan, Taiwan, pp. 1–2, 2019.

[4] Y. P. Peng, Q. G. He, X. Y. Ke, J. Hu and L. X. Wu, "A fall detection belt based on an acceleration sensor," *Electronic Measurement Technology*, vol. 41, no. 11, pp. 117–120, 2018.

[5] Y. H. Cui and L. Zhan, "Detection of fall of elderly based on three-axis accelerator sensor," *Modern Electronic Technology*, vol. 36, no. 3, pp. 130–132, 2013.

[6] J. B. Jiang, F. C. Wang, Y. X. Zhao and L. J. Yang, "The design of the physical monitoring system for the fall of the elderly based on smart watch," *Fujian Computer*, vol. 34, no. 03, pp. 21–22+13, 2018.

[7] M. Alwan, P. J. Rajendran, S. Kell, D. Mack, S. Dalal *et al.,* "A smart and passive floor-vibration based fall detector for elderly," in *Proc. 2nd Int. Conf. on Information & Communication Technologies*, Damascus, pp. 1003–1007, 2006.

[8] F. Haider and G. Shaker, "Wearable-free wireless fall detection system," in *Proc. 2017 IEEE Int. Symp. on Antennas and Propagation & USNC/URSI National Radio Science Meeting*, San Diego, CA, pp. 2457–2458, 2017.

[9] Y. Ogawa and K. Naito, "Fall detection scheme based on temperature distribution with IR array sensor," in *Proc. ICCE*, Las Vegas, USA, pp. 1–5, 2020.

[10] Y. T. Hsu, J. W. Hsieh, H. F. Kao, and H. Y. Mark Liao, "Human behavior analysis using deformable triangulations," in *Proc. of Multimedia Signal Processing, 2005 IEEE 7th Workshop*, Taiwan, pp. 1–4, 2005.

[11] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Monocular 3d head tracking to detect falls of elderly people," in *Proc. EMBS 06. 28th Annual Int. Conf. of the IEEE*, Canada, pp. 6384–6387, 2006.

[12] M. Yu, A. Rhuma, S. M. Naqvi and J. Chambers, "Fall detection for the elderly in a smart room by using an enhanced one class support vector machine," in *Proc. DSP, 2011 17th Int. Conf.*, UK, pp. 1–6, 2011.

[13] Y. B. Chen, H. W. He, G. Z. Wang and G. T. Wang, "An elderly fall detection system based on machine vision," *Automation and Information Engineering*, vol. 40, no. 5, pp. 37–41, 2019.

[14] M. G. Zhou, "Research on human fall detection algorithm based on computer vision," Ph.D. dissertation, University of Shandong, China, 2013.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016.

[16] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556, 2015.

[17] S. F. Ruan, "A study on detection of pedestrian mask based on improved SSD algorithm," *Science and Technology Economic Guide*, vol. 28, no. 35, pp. 9–13, 2020.

[18] G. H. Yu, H. H. Fan, H. Y. Zhou, T. Wu and H. J. Zhu, "Vehicle target detection method based on improved SSD model," *Journal on Artificial Intelligence*, vol. 2, no. 3, pp. 125–135, 2020.

[19] H. Tang, A. Peng, D. Zhang, T. Liu and J. Ouyang, "SSD real-time illegal parking detection based on contextual information transmission," *Computers, Materials & Continua*, vol. 62, no. 1, pp. 293–307, 2020.

[20] M. Sandler, A. Howard, M. L. Zhu, Andrey Zhmoginov, Liang-Chieh Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, pp. 4510–4520, 2018.

[21] D. Zheng, X. Q. Li and X. Z. Xu, "Vehicle and pedestrian detection network based on lightweight SSD," *Journal of Nanjing Normal University (Natural Science Edition)*, vol. 42, no. 1, pp. 73–81, 2019.

[22] H. Liu, L. S. Zhang, Y. Shen, J. Zhang and B. Wu, "A real-time detection of pedestrian in orchard based on improved SSD method," *Journal of Agricultural Machinery*, vol. 50, no. 4, pp. 29–35+101, 2019.

[23] T. Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," arXiv preprint arXiv: 1612.03144, 2017.

[24] B. Wang and Y. J. Gu, "A study on vehicle detection technology based on FPN image," *Industrial Control Computer*, vol. 33, no. 7, pp. 88–90, 2020.

[25] H. J. Huang, X. H. Duan and X. C. Huang, "Research and improvement of fruit detection based on deep learning," *Computer Engineering and Applications*, vol. 56, no. 3, pp. 127–9133, 2020.

[26] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. ICCV*, pp. 4885–4894, 2017.

[27] Z. Zheng, P. Wang and W. Liu, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI*, New York, USA, pp. 12993–13000, 2020.