



Joint information fusion and multi-scale network model for pedestrian detection

Hexiang Zhang¹ · Ziyu Hu² · Ruoxin Hao¹

Accepted: 12 October 2020 / Published online: 10 November 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

The existing pedestrian detection suffers the low accuracy when the environment changes dramatically. In order to solve the problem, a pedestrian detection model combining information fusion and multi-scale detection is proposed. The model is composed of a retinex algorithm and an improved YOLOv3 algorithm. Retinex algorithm is selected as the preprocessing algorithm to improve the brightness and contrast of pedestrians. The model improves the YOLOv3 algorithm by adding multiple scale detections. The K-means is used to determine the number of optimal anchors and the aspect ratio. By testing on the standard data set, the mean average precision (mAP) of the joint detection model increases from the original 80.69–91.07%, and the recall increases from 65.22 to 87.48%. The comparative experiments show that the improved model performs good robustness and generalization ability on the problem of low pedestrian detection accuracy in complex environments.

Keywords YOLOv3 · Pedestrian detection · Information fusion · Multi-scale network

1 Introduction

Pedestrian detection [1] is a core technology in some applications, such as application in robot vision [2], intelligent monitoring systems, and driverless assistance systems [3], etc. Pedestrian detection technology can determine whether there is a pedestrian in the input image and the approximate location of the pedestrian. The accuracy and real-time performance of pedestrian detection is essential for the evaluation of the superiority of the detection system.

Pedestrian detection belongs to the existing category of object detection, and the development of pedestrian detection can be divided into the following three periods.

(1) The first period is the traditional object detection algorithm, which mainly uses artificial design features for

global feature pedestrian detection. Features are generally extracted by artificial design methods, such as Harr, LBP, HOG [4,5] features, Shapelet and Edgeline features, etc. Support vector machine classifier is mainly used in the classification stage. According to the traditional detection algorithm, Navneet Dalal and Bill Triggs proposed the HOG + SVM [6,7] network structure to improve detection accuracy and robustness, which makes traditional object detection possible for pedestrian detection. Since it is difficult for artificially designed features to express the features of pedestrians in complex environments, the model detection accuracy is low. The algorithm itself has program redundancy and cannot be detected in real time.

(2) The second period is the R-CNN algorithm idea which is based on candidate windows and deep learning classifiers. Representative algorithms are Fast R-CNN [8], Faster R-CNN [9,10], and Mask R-CNN [11]. The two-stage method could achieve the best detection accuracy. However, its disadvantages are also obvious. The generation and detection of candidate windows is a long process, so it is difficult to achieve the purpose of real-time detection.

(3) The third period is the YOLO [12] algorithm based on the regression method, which inputs the entire image into a neural network, and directly provides the predicted

✉ Ziyu Hu
hzy@ysu.edu.cn
Hexiang Zhang
zhxysu@163.com

¹ Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China

² Engineering Research Center of the Ministry of Education for Intelligent Control System and Intelligent Equipment, Yanshan University, Qinhuangdao, China

position and category of the object through network calculation. It reduces part of the detection accuracy but greatly improves the detection speed, making real-time detection possible. Representative algorithms include SSD (Single Shot MultiBox Detector) [13,14] algorithm, three versions of YOLO algorithm, and so on. Among them, the YOLOv3 [15,16] algorithm performs a stronger comprehensive computing capability, with a mAP of 57.9% in the COCO [17] dataset.

Pedestrian images are easily affected by factors such as angle, posture, brightness, and contrast. The difficulty of pedestrian detection increases with the number of parameter variables, which makes pedestrian detection a difficult task in the field of computer vision. In this paper, a method is proposed to reduce the influence of background environment on detection through image preprocessing. Through information fusion, the network structure and parameters are adjusted. In the end, the overall structure can improve accuracy at the expense of a small amount of time. Section 2 introduces the existing Retinex algorithm and YOLOv3 algorithm. The principle and function of the algorithm are briefly described. Section 3 mainly introduces image preprocessing algorithms and information fusion. Section 4 details how to improve the YOLOv3 algorithm, and how to merge the two structures through information fusion. Section 5 verifies the function and effect of the model through a series of comparative experiments.

2 Background

2.1 Retinex algorithm

Retinex algorithm [18] is a commonly used image enhancement method, proposed by Edwin.H.Land in 1962. The retinex mode is based on the following assumptions:

- (1) The real world is colorless, and the colors we perceive are the result of the interaction between light and matter, such as natural phenomena, light reflection and refraction;
- (2) Each color gamut is composed of three primary colors of red, green, and blue at a given wavelength;
- (3) The color of each unit area is determined by the three primary colors. The basic theory of Retinex is that the color of an object is determined by the ability to reflect long (red), medium (green), and short (blue) light, rather than the absolute value of the reflected light intensity. The color of the object is not affected by the unevenness of the light and it shows consistency, that is, Retinex is based on the consistency of the color sense (color consistency). Different from traditional linear and nonlinear methods

which can only enhance one image feature, Retinex can adaptively enhance different types of images.

The Retinex algorithm is divided into 3 categories according to the development processes:

1. The path-based Retinex algorithm is the earliest classic version. For each pixel in the image, the algorithm searches for a 1 dimensional random path in its neighborhood. Then, it calculates the product of the ratio of adjacent pixels on each path to obtain the relative brightness value, and the average value on all paths is the new pixel value.
2. Iterative Retinex algorithm, an iterative form of Retinex algorithm, proposed by Frankle and McCann, is also called multi-resolution Retinex algorithm. The algorithm uses matrix calculations instead of path calculations. In path calculation, pixels are sequentially added to the path, and pixel values are compared in a serial manner. The distance function is determined according to the length of the path. In matrix calculation, the ratio and product operations can be processed in parallel, and the distance function is determined according to the number of iterations.
3. The theoretical basis of the neighborhood-based Retinex algorithm is that the intensity of the illuminance component generally changes slowly, which shows the reflectance of different surface materials varies widely. The algorithm uses low-pass filtering to estimate the illuminance component, also known as the center surround Retinex algorithm. For each pixel in the image, calculate the nonlinear ratio of the pixel value to the weighted value of pixels in the neighborhood, which is the new pixel value of the pixel. The weighted value of pixels in the neighborhood is obtained by spatial convolution of the smoothing function, and the weight is given by the coefficient of the smoothing function.

Jobson evaluated the performance of algorithms using different types of smoothing functions, determined the Gaussian function, proposed a single-scale Retinex (SSR) [19] algorithm, and MSR [20,21] method with color recovery MSRCR (Multi-Scale Retinex with Color Restoration) [16]. This paper uses McCann99 Retinex algorithm to improve the brightness and contrast of the picture.

2.2 YOLOv3 detection algorithm

With the development of the object detection algorithm, Redmon proposed the YOLO algorithm in 2016. Based on the idea of regression, the algorithm can detect without generating any candidate frames, which avoids the time spent in the process of generating candidate frames. This improvement

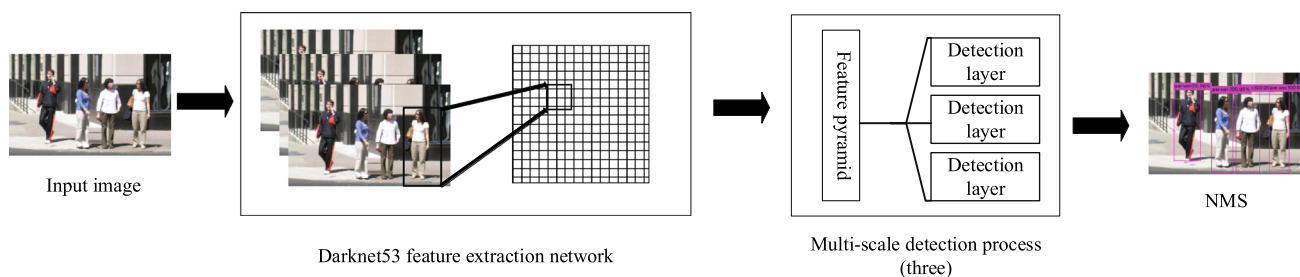


Fig. 1 YOLOv3 algorithm detection flowchart

makes real-time detection possible. Since then, the YOLO series of algorithms have entered into the field of vision. On the basis of the YOLOv2 algorithm, YOLOv3 [22,23] algorithm retains the detection speed and improves the detection precision. the structure of YOLOv3 is shown in Fig. 1.

In YOLOv3, the newly designed darknet-53 network is used as the main network for feature extraction. There are 53 convolutional layers in the entire network. Compared with darknet-13, the network structure is deeper. This structure uses a series of well-performing convolutional layers of size 3×3 and 1×1 , and uses LeakyRelu as a correction function to improve the accuracy of object recognition while maintaining high detection speed. A large number of residual connections are used in the network to increase the depth of the Network. Combing with the FPN (feature pyramid networks) network structure, the two feature maps on the network are sampled, aggregated with the feature maps of the corresponding size in the early stage of the network, and the prediction result is obtained through the convolution network. During the object detection process, YOLOv3 performs a convolution operation on the size of 13×13 feature map, and then sends it to the detection layer to obtain a detection result. Then, the feature map of size 26×26 is fused with the feature map which sampled on the feature map of size 13×13 in the previous network, and the new feature map is convolved and sent to the detection layer to obtain a detection result; furthermore, the features map of size 52×52 is fused with feature map which sampled on the feature map of size 26×26 in the previous network, and the new feature map is convolved and sent to the detection layer to obtain a detection result. Non-maximum suppression is performed on the three detection results to obtain the final detection result.

3 Introduce McCann99 Retinex algorithm

In this paper, the McCann99 Retinex algorithm is used as the model preprocessing algorithm. By modifying the related parameters, the sharpness and contrast of the image are improved, which is beneficial to the information fusion of the model. The multi-resolution descriptions is built by the image pyramids, and iterates layer by layer from top to bottom to

improve the efficiency. Because the pyramid model needs to downsample the input image, the low-resolution image on the upper layer of the pyramid is 1/2 of the high-resolution image on the lower layer. Therefore, the McCann99 algorithm determines the size of the input image. The simplified process of the algorithm is described in the following steps:

1. The original image is transformed to the logarithmic domain S ; the length and width of the original image have strict restrictions and can be expressed as: $rows \cdot 2^n \times cols \cdot 2^n$ and $cols \geq rows, cols \geq 1, rows \leq 5$.
2. The number of image pyramid layers is initialized and calculated. The initialized constant image matrix R is used as the initial value for the iterative operation;
3. From the top layer to the last layer, 8 neighborhood comparison operations are performed. The calculation formula is:

$$NP(x, y) = \sqrt{OP(x, y) IP^*(x, y)} \tag{1}$$

$$NP(x, y) = \sqrt{OP(x, y) \left[OP(x_s, y_s) \frac{R(x, y)}{R(x_s, y_s)} \right]^*} \tag{2}$$

where OP is the product of the previous iteration; NP is the product of the current iteration; IP is the intermediate product result; R is the input image of the layer; the symbol $*$ represents the reset operation. OP 's initial size is $rows \times cols$.

4. The number of pyramid layers is determined by the size of the original image. The size is $rows \cdot 2^n \times cols \cdot 2^n$ and the number of layers is $(n + 1)$. After the n th layer operation finished, the operation result of the n th layer is interpolated and becomes twice the original size.
5. When the bottom layer is calculated, it is the final enhanced image. The algorithm structure is shown in Fig. 2.

Based on the theory of multi-source information fusion, an intelligent model-level fusion method is used to fuse the McCann99 Retinex algorithm model with the YOLOv3 model. This method performs well in terms of information processing capacity, anti-interference ability, and fusion performance. The information fusion model reduces the amount

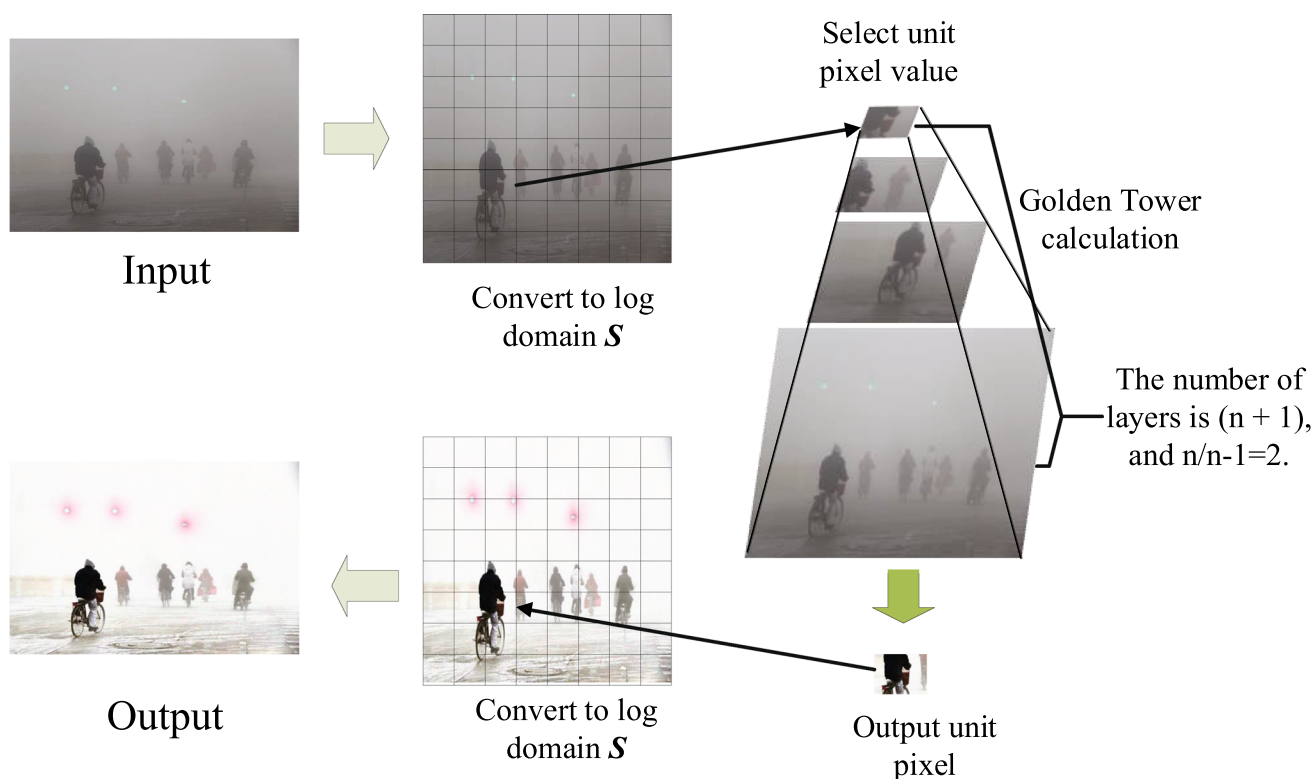


Fig. 2 Pyramid Structure Algorithm

of calculation. For the convenience of description, this paper named the model after information fusion as YOLOv3-99.

4 Improved YOLOv3 algorithm

4.1 Introduction of multi-scale detection

In different scene information, the sizes of the same object may cause detection errors. In order to make the model more sensitive to small objects, the concept of multi-scale fusion is introduced. In the YOLOv3 algorithm, a feature pyramid network (FPN) [24] network is used to sample and fuse features of different layers. The high-resolution features of the low-level features and the high-semantic information of the high-level features are used to detect target objects on 3 different scale feature maps. These three scales feature maps are 13, 26, and 52. Since pedestrian detection objects are small and the background environment is complex, this paper increases the original feature extraction from 3 different scales to 4 different scales. The improved detection model further improves the accuracy of pedestrian detection, strengthens the use of shallow information, and prevents partial information loss after multiple layers of convolution. The detection module assigns accurate anchor boxes to feature maps of different scales to improve the accuracy of pedestrian

detection. The recognition effect of small object is always regarded as the standard to evaluate the algorithm. The overall pedestrian detection network model is shown in Fig. 3.

4.2 Dimensional clustering of object boxes

The YOLOv3 algorithm introduces the anchor boxes mechanism from Faster R-CNN network. The number of anchors and the aspect ratio will affect the accuracy of object detection. In the process of model training, the parameters of candidate box will be adjusted constantly with the increase of the number of iterations. The parameters are finally close to the real box parameters. The YOLOv3 algorithm obtains the number of optimal anchors and aspect ratio by performing K-means [25] dimensional clustering on the COCO dataset. Compared with faster R-CNN artificially setting prior frames, the K-means clustering method is easier to learn by deep convolutional neural networks (CNN). The purpose of this method is to enhance the recognition of dense crowds and small targets, so deep convolutional neural networks are needed to learn pedestrian features from a large amount of samples. The COCO dataset contains about 80 types of object data, including many types of data such as people, cars, grassland, etc. There are flat long boxes that meet pedestrian characteristics, and relatively few thin and tall boxes. The candidate frame types of the experimental

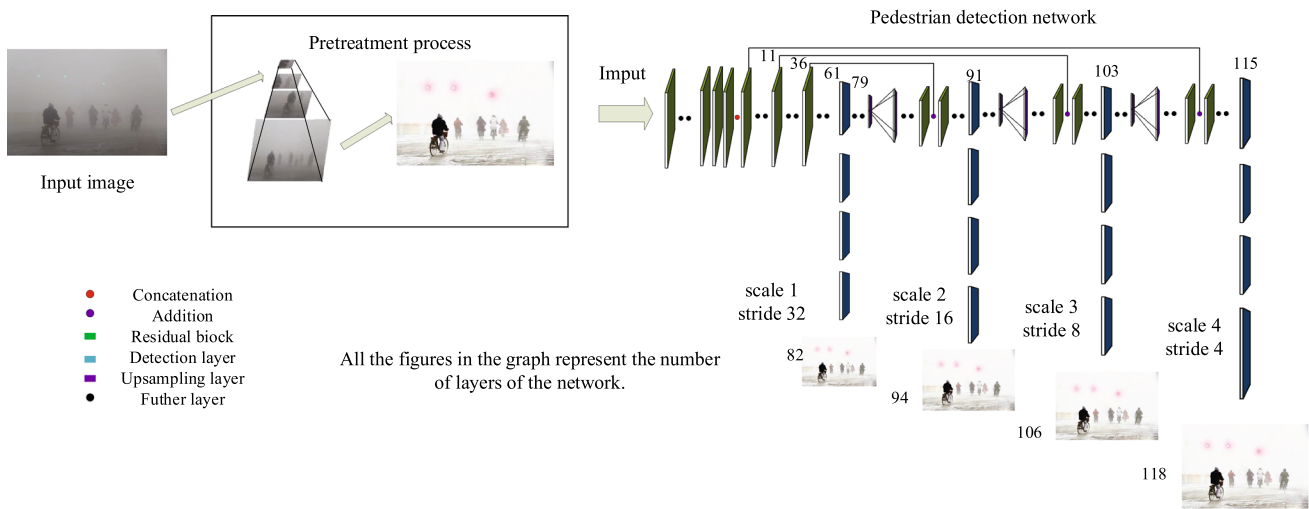


Fig. 3 Structure of pedestrian detection model based on multi-scale detection and information fusion

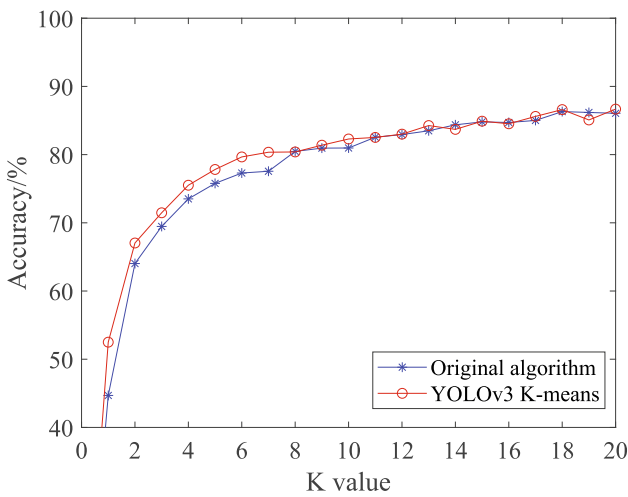


Fig. 4 Change curve of objective function

data set collected in this paper are opposite to the candidate frame types of the COCO data set, that is, most of them are thin and tall data.

By using K-means clustering analysis on the experimental data set, it is obtained that the optimal anchor number, width and height dimensions suitable for the high-density population data set. Traditional K-means clustering uses indirect clustering by measuring the similarity between samples.

Usually, Euclidean distance [26] or Manhattan distance [27,28] is used as the measurement formula, but it will produce a big box advantage which makes larger bounding boxes produce more errors than smaller bounding boxes. In response to the above problems, the YOLOv3 algorithm chooses the Intersection Over Union (IOU) to reflect the error between the candidate box and the real box. The distance for-

mula is:

$$IOU = \frac{|B \cap C|}{|B \cup C|} \tag{3}$$

$$d(box, centroid) = 1 - IOU(box, centroid) \tag{4}$$

where B is the acronym for box, and C is the acronym for centroid. Box is the sample clustering result; centroid is the center of all clusters; $IOU(box, centroid)$ is the intersection ratio of all cluster centers to all cluster boxes. Selecting an appropriate IOU score can achieve a good balance between model complexity and recall. The two methods perform cluster analysis on the object data set of the pedestrian. The change curve of the objective function is shown in Fig. 4. As the number of clusters K value (cluster value) is different, the objective function curve changes accordingly; when the value of K is greater than 7, the objective function curve tends to be smooth, so the choice of K value is 7, which means that anchor is equal to 7. In addition, with different K values, the initial specifications of the network training candidate frame also change. The details are shown in Table 1. Fig. 5 shows the clustering effect of a pedestrian under different K values. The K value gradually changes from 2 to 9. When the K value is equal to 7, the pedestrian clustering effect is stable.

4.3 Determination of candidate boxes

At the time of detection, the picture is input into the convolutional layer to extract pedestrian features of training and testing. Furthermore, there are a total of $S \times S$ small cells in the image. Each small cell predicts B bounding boxes. Each bounding box corresponds to 5 prediction parameters, that is, the 4 coordinates and confidence in the bounding boxes, there are $S \times S \times B$ candidate frames in total. There are not

Table 1 Cluster membership of the experimental dataset

K value	Specification 1	Specification 2	Specification 3	Specification 4	Specification 5	Specification 6	Specification 7
$K = 3$	[0.21,0.77]	[0.07,0.17]	[0.12,0.55]				
$K = 4$	[0.27,0.80]	[0.06,0.17]	[0.17,0.71]	[0.11,0.51]			
$K = 5$	[0.11,0.48]	[0.21,0.77]	[0.16,0.69]	[0.06,0.17]	[0.33,0.81]		
$K = 6$	[0.12,0.67]	[0.20,0.79]	[0.17,0.59]	[0.32,0.80]	[0.10,0.38]	[0.05,0.15]	
$K = 7$	[0.24,0.77]	[0.33,0.79]	[0.20,0.78]	[0.15,0.68]	[0.11,0.49]	[0.06,0.17]	[0.46,0.87]

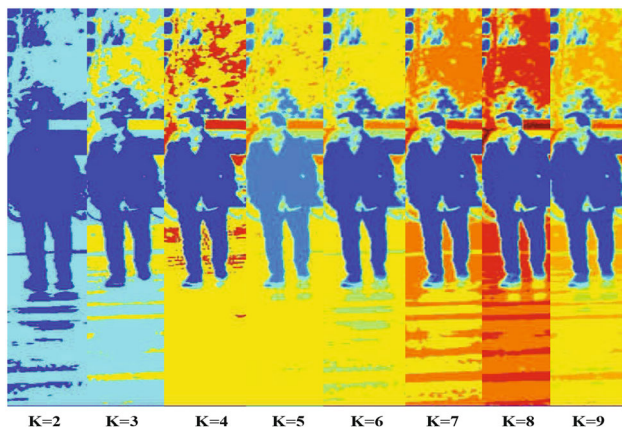


Fig. 5 Image clustering effect map

necessarily objects in the candidate frames, the confidence level without object frame is set to 0, and the confidence level with object frame is:

$$Object = P (Object) \times IOU_{Truth}^{Presd} \tag{5}$$

The object existence probability expression in the box is $P = P_r (Person|Object)$. By multiplying the probability of the person in the test candidate box with the confidence level, a common confidence formula in each candidate box can be obtained:

$$conf = P \times P (Object) \times IOU_{Presd}^{Truth} \tag{6}$$

The above formula is to determine whether any object falls into the candidate box. If not, *conf* is 0, otherwise:

$$conf = IOU_{Presd}^{Truth} \tag{7}$$

IOU_{Presd}^{Truth} represents the ratio of the communication area of the predicted and actual boxes to the area of the union, and also reflects the possibility that the object falls into the candidate box:

$$IOU_{Presd}^{Truth} = \frac{area (box (Truth) \cap box (Pred))}{area (box (Truth) \cup box (Pred))} \tag{8}$$

Table 2 Software and hardware configuration of the experimental platform

Software and hardware	Configuration
Operating platform	Ubuntu 18.04
CPU/GHz	Intel(R)Xeon(R)E5-2667,2.90
GPU	CUDA 10.0, CUDNN 7.6
RAM	16GB
Deep learning framework	Darknet, caffe

Finally, non-maximum suppression is used to filter out the unrelated pedestrian frames and keep valid frames. The content of the output box is $[C, conf, x, y, w, h]$, which is the category, confidence and offset of the object box of the center position, width and height.

5 Experimental verification and result analysis

5.1 Experimental platform and data set

The specific hardware and software configuration of the experimental platform is shown in Table 2. All experimental training is performed on the platform. The initial parameter settings are as follows: learning_rate = 0.001, batch = 64, steps = (40,000, 45,000), max_batches = 50,200, subdivisions = 8.

This paper collects 2,000 pedestrian images as a pedestrian data set. Most of them come from standard data sets (VOC2012, COCO2014), other parts are collected during work. Pedestrians have rich appearances, complex backgrounds, and significant individual differences, accompanied by different degrees of overlap, which meet the requirements of experimental data. The labeling tool is used for manually label pedestrians on the training set. McCann99 algorithm is used to preprocess special environment images, and the pre-processed images have significantly improved contrast and sharpness. Among the 2,000 images, there are 1,400 images in the training set and 400 images in the verification set. Another 200 images are selected as the test set. An example image of a pedestrian dataset is shown in Fig. 6.



Fig. 6 Example picture of pedestrian data set

5.2 Pedestrian detection accuracy experiment

For the convenience of description, this paper renames the improved YOLOv3 network to YOLOv3-Z network. In order to verify the accuracy and credibility of the experiment, the algorithm YOLOv3-Z is compared with Faster-RCNN, YOLOv3, YOLOv2, SSD on the same pedestrian data set. It is well-known that the Faster-RCNN algorithm is the method with the highest mAP value of current object detection methods. YOLOv3 is an improved version of YOLOv2, which is representative of YOLO series algorithms. SSD, a single-shot detector for multiple categories, is faster and more accurate than the previous advanced single-shot detector (YOLO). Therefore, this paper chooses them to compare with the algorithm YOLOv3-Z. They are trained 120000 times, respectively, by using the experimental data set. It is observed that the value of accuracy in network training. The training results are shown in Table 3.

where the accuracy rate P is defined as

$$P = \frac{TP}{TP + FP} \tag{9}$$

The model which performed well is chosen for testing, and Recall and mAP are used as the detection indicators. The

Table 4 Test results of different models on test datasets

Method	Backbone	mAP/%	Recall/%	IOU	Time
YOLOv3-Z	Darknet	87.616	73.68	84.72	0.019
YOLOv2	Darknet	74.783	48.45	70.32	0.026
YOLOv3	Darknet	80.686	65.22	79.62	0.015
Faster R-CNN	caffe	89.153	70.65	86.22	0.039
SSD	caffe	84.122	75.76	82.78	0.021

The metrics in bold indicates the best values

test results are shown in Table 4. The calculation formula of Recall is:

$$R = \frac{TP}{TP + NP} \tag{10}$$

where R stands for recall. TP stands for the prediction result of the network is a pedestrian. The real result is also the number of pedestrian samples. FP indicates that the prediction result is non-pedestrian, and the real result is the number of samples of pedestrians. The larger the Recall value, it proves that the overall network’s ability to detect pedestrians is better, and the detection effect is better. The maximum value of Recall is 1. The calculation formula of mAP is:

$$mAP = \frac{\sum AP}{NC} \tag{11}$$

In Eq. (11), $\sum AP$ represents the sum of the average accuracy of the row human pictures, and NC represents the sum of the row human pictures.

It can be seen from the table that the YOLOv3-Z algorithm is higher than the original YOLO series algorithms in mAP, recall and IOU. Compared with Faster R-CNN and SSD, the comprehensive performance of YOLOv3-Z algorithm is similar. Although YOLOv3-Z is lower than Faster R-CNN

Table 3 Training results of YOLOv3-Z, YOLOv3 and Faster R-CNN

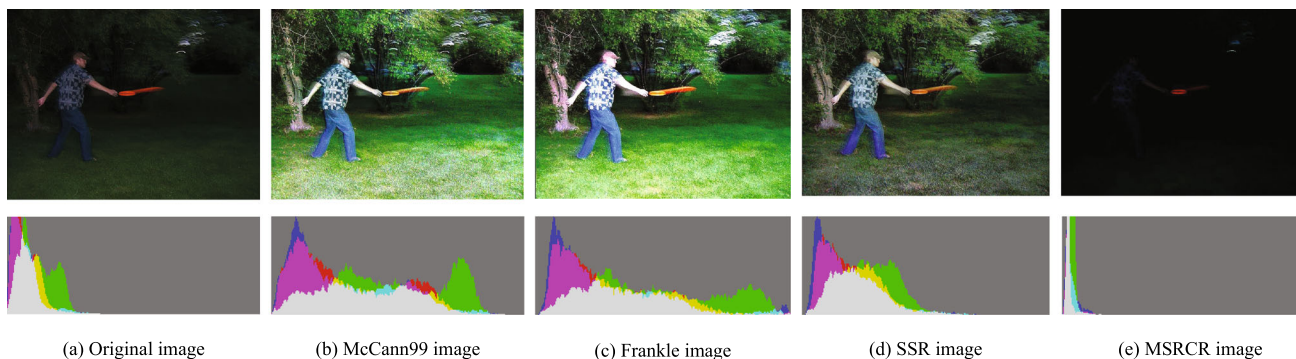
Network iterations	Accuracy P			
	YOLOv3-Z	Faster-RCNN	SSD	YOLOv3
10000	0.68	0.69	0.68	0.67
20000	0.71	0.76	0.72	0.71
30000	0.76	0.80	0.75	0.74
40000	0.77	0.81	0.77	0.75
50000	0.79	0.82	0.77	0.76
60000	0.80	0.83	0.81	0.78
70000	0.81	0.85	0.81	0.78
80000	0.82	0.84	0.81	0.80
90000	0.87	0.89	0.84	0.81
100000	0.84	0.88	0.86	0.79
110000	0.84	0.89	0.85	0.79
120000	0.84	0.87	0.86	0.78

The metrics in bold indicates the best values

Table 5 Quantitative evaluation index of images in Fig. 7

Enhancement algorithm	Mean value	Standard deviation	Information entropy	Color entropy
Image	60.905	48.615	5.016	15.284
McCann99	112.345	74.624	7.472	19.318
SSR	108.436	40.823	7.641	19.817
MSRCR	40.735	21.984	4.343	10.376
Frankle-McCann	105.139	68.872	6.735	17.463

The numbers in bold are the best results

**Fig. 7** Comparison of various image processing methods

in mAP value and *IOU* value, and slightly lower than SSD in recall value, the YOLOv3-Z takes less time than them.

5.3 Comparison experiment of pretreatment algorithm

Experiment 2 explores the effect of the preprocessing algorithm on the image. By using different Retinex algorithms to preprocess the images, the image related indicators are obtained. The classic Retinex algorithms selected are SSR, MSRCR, and Frankle-McCann algorithms. The Mean Value, Standard Deviation, Information Entropy, and Color Entropy are used as indicators to evaluate the quality of the algorithm. The mean value represents the change in the overall brightness of the image. The higher the average value, the greater the image brightness. The standard deviation represents the contrast of the image. The information entropy represents the amount of information in an image. Color entropy is used to measure the degree of color enhancement of an image. The experimental test set is used to test the Retinex algorithms. The comparative data results are shown in Table 5.

It can be concluded from Table 5 that McCann99 is stronger than other algorithms in improving image brightness and contrast. It also has a good performance at the level of rich image information. These experimental results indicate that the enhanced image can highlight more information. It can be observed from the Fig. 7 that the Model algorithm is increasing the value of the high-brightness area. By observing the trend of the curve, the curve is U-shaped.

The greater the height difference, the higher the image contrast. The number on the right side of the color histogram represents the brightness of the image. Like the (b) and (c), the more the number, the higher the brightness. We can conclude that the McCann99 algorithm can greatly improve the brightness, contrast and information of the image.

5.4 Joint detection model test experiment

Experiment 3 is to test the fusion effect and detection speed of the model. We selected SSD-99, Faster R-CNN-99 and YOLOv3-99 for comparison in the same data set. Detection speed and single detection time are used as evaluation indicators. See Table 6 for specific data. For reference, the unmodified Faster R-CNN, SSD, YOLOv3 values are also listed in Table 6.

It can be observed from the table that YOLOv3-99 takes 0.05 seconds to improve the accuracy rate by about 10%. Compared with the Faster R-CNN algorithm, the algorithm proposed in this paper has low accuracy, but has a fast detection speed. It can be clearly observed from Fig. 8 that the accuracy of the joint detection model is higher than that of the ordinary detection model. Experiment 3 proves that the joint detection model improves the detection accuracy by spending a little more time.

Table 6 Related numerical comparison table

Model	Data set	mAP/%	Recall/%	$AP_{50}/\%$	Time
YOLOv3-99	VOC2007+2012	91.1	87.5	90.2	0.020
YOLOv3	VOC2007+2012	80.7	65.2	78.6	0.015
SSD-99	VOC2007+2012	87.4	80.1	86.0	0.027
SSD	VOC2007+2012	84.1	75.8	80.3	0.021
Faster R-CNN-99	VOC2007+2012	92.1	86.2	90.9	0.046
Faster R-CNN	VOC2007+2012	89.2	70.7	86.3	0.039

The metrics in bold indicates the key comparison values

**Fig. 8** Comparison of various image methods

6 Conclusion

The pedestrian detection model based on joint information fusion and multi-scale network is proposed, which combines the YOLOv3 algorithm theory with Retinex algorithm theory. By improving the network structure and adding multi-scale detection, the dimensional clustering of object boxes is reset. Based on the former modification, the accuracy of pedestrian detection is improved. The McCann99 Retinex algorithm is used to preprocess the experimental data set, which enhances the contrast between pedestrians and the overall picture. After preprocessing, the YOLOv3 network model can obtain more image information, which greatly increases the detection range and accuracy. It can be concluded that the mAP of the improved YOLOv3 increases from 80.67% to 91.07%, and the recall increases from 65.22% to 87.48%. This effect only takes about 0.05 seconds. However, this method still need to be improved when dealing with the changing scenario. For example, the number of images in the data set is very small, and the McCann99 Retinex algorithm strictly requires the image input size. Therefore, how to further optimize the algorithm structure and expand the detection range while ensuring the detection accuracy will be the focus of future research.

Acknowledgements This work was supported by Project supported by National Natural Science Foundation of China (No. 62003296), the Natural Science Foundation of Hebei (No. F2020203031), Science and Technology Project of Hebei Education Department (No. QN2020225)

Compliance with ethical standards

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Suprit, D.B., Abhijeet, V.N.: Crowd anomaly detection and localization using histogram of magnitude and momentum. *Vis. Comput.* **36**(3), 609–620 (2020)
- Lin Feng, Y.L., Zan, L., Meng, Z., Feilong, W., Shenglan, L.: Discriminative bit selection hashing in rgb-d based object recognition for robot vision. *Assembly Automation* (2019)
- Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C., Winner, H.: Three decades of driver assistance systems: review and future perspectives. *IEEE Intell. Transp. Syst. Magaz.* **6**(4), 6–22 (2014)
- Pang, Y., Yuan, Y., Li, X., Pan, J.: Efficient hog human detection. *Signal Process.* **91**(4), 773–781 (2011)
- Chengbin, Z., Huadong, M., Anlong, M.: Fast human detection using mi-svm and a cascade of hog-lbp features. In: 2010 IEEE International Conference on Image Processing, pp. 3845–3848. IEEE (2010)
- Navneet Dalal, B.T.: Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893. IEEE (2005)
- Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. *Neurocomputing* **55**(1–2), 169–186 (2003)
- Zhong-Qiu, Z., Haiman, B., Donghui, H., Wenjuan, C., Hervé, G.: Pedestrian detection based on fast r-cnn and batch normalization. In: International Conference on Intelligent Computing, pp. 735–746. Springer (2017)
- Shaoqing, R., Kaiming, H., Ross, G., Jian, S.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
- Yeong-Hyeon, B., Keun-Chang, K.: A performance comparison of pedestrian detection using faster rcnn and acf. In: 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 858–863. IEEE (2017)

11. Kaiming, H., Georgia, G., Piotr Dollár, R.G.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
12. Joseph, R., Santosh, D., Ross, G., Ali, F.: 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)
13. Wei, L., Dragomir, A., Dumitru, E., Christian, S., Scott, R., Cheng-Yang, F., Alexander, C.B.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer (2016)
14. Alexander, W., Mohammad, J.S., Francis, L., Brendan, C.: Tiny ssd: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection. In: 2018 15th Conference on Computer and Robot Vision (CRV), pp. 95–101. IEEE (2018)
15. Chong, L., Rong, W., Jinze, L., Linyu, F.: Face detection based on yolov3. In: Recent Trends in Intelligent Computing, Communication and Devices, pp. 277–284. Springer (2020)
16. Ajit, J., Prerana, M., Vinay, K., Brejesh, L.: Aerial multi-object tracking by detection using deep association networks. In: 2020 National Conference on Communications (NCC), pp. 1–6. IEEE (2020)
17. Tsung-Yi, L., Michael, M., Serge, B., James, H., Pietro, P., Deva, R., Piotr, D., Lawrence, Z.: Microsoft coco: Common objects in context. In European conference on computer vision, pp. 740–755. Springer (2014)
18. Zia-ur, R., Daniel, J.J., Glenn, A.W.: Retinex processing for automatic image enhancement. *J. Electron. Imag.* **13**(1), 100–111 (2004)
19. Zietkiewicz, E., Rafalski, A., Labuda, D.: Genome fingerprinting by simple sequence repeat (ssr)-anchored polymerase chain reaction amplification. *Genomics* **20**(2), 176–183 (1994)
20. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(11), 1330–1334 (2000)
21. Qin, Y., Luo, F., Li, M.: A medical image enhancement method based on improved multi-scale retinex algorithm. *J. Med. Imag. Health Inf.* **10**(1), 152–157 (2020)
22. Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z.: Apple detection during different growth stages in orchards using the improved yolo-v3 model. *Comput. Electron. Agric.* **157**, 417–426 (2019)
23. Mujtaba, A., Jie, Y., Jiang, H., Pourya, S., Xiangjian, H.: Multi-frame feature-fusion-based model for violence detection. *Vis. Comput.* pp. 1–17 (2020)
24. Tsung-Yi, L., Piotr Dollár, R.G., Kaiming, H., Bharath, H., Serge, B.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
25. John, A.H., Manchek, A.W.: Algorithm as 136: A k-means clustering algorithm. *J. Royal Stat. Soc. Ser. c* **28**(1), 100–108 (1979)
26. Zitnick, C.L., Piotr, D.: Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision, pp. 391–405. Springer (2014)
27. Danielsson, P.-E.: Euclidean distance mapping. *Computer Graph. Image Process.* **14**(3), 227–248 (1980)
28. Vishnu, G.N., Guruprasad, K.R.: Multi-robot coverage using voronoi partitioning based on geodesic distance. In: Control Instrumentation Systems, pp. 59–66. Springer (2020)



Hexiang Zhang was born in Tangshan, China. He is studying for a bachelor's degree in control engineering and technology at the School of Electrical Engineering, Yanshan University, Qinhuangdao City, Hebei Province, China. His research directions include multi-target detection, objects tracking and image processing. So far, he has two invention patents.



Ziyu Hu received his B.E. degree in engineering from North China Institute of Aerospace Engineering, Langfang, China, in 2012, and Ph.D. in control science and engineering from Yanshan University, Qinhuangdao, China, in 2018. He is currently a lecturer with the Department of Automation, Yanshan University, Qinhuangdao, China. His current research interests include computer vision, multiobjective optimization evolutionary algorithm, and simulation of complex systems.



Ruoxin Hao was born in Xingtai, China. She is studying for a bachelor's degree in Measurement and Control Technology and Instrumentation Program, Yanshan University, Qinhuangdao City, Hebei Province, China. Her current research interests include computer vision simulation, pedestrian recognition and image processing.