Motion Target Detection and Recognition Based on YOLOv4 Algorithm

Hailan Yu and Weili Chen
School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou, Guangdong, China
Email: 360724201@qq.com

Abstract. With the application of artificial intelligence more and more widely, the target detection of artificial intelligence “eyes” is becoming more and more important, which can give the machine the ability to detect and recognize image content in the environment. In this paper, YOLOv4 deep learning motion target detection algorithm is used to achieve the positioning and recognition of motion targets, to identify and mark the location and type of objects contained in the image, to achieve the detection of pictures, videos and camera real-time image detection, so that the machine has the most basic visual ability. Using YOLOv4 algorithm, the speed requirement of real-time detection can be realized with certain precision, and the model mAP value reached 69.48%.

Keywords. YOLOv4; motion target detection; positioning and recognition; vision.

1. Introduction
With the development of computer technology and the wide application of computer vision principle, it is more and more popular to use computer image processing technology to track the target in real time, and it is of great value to carry out real-time tracking and positioning of moving target in intelligent traffic system, intelligent monitoring system, military target detection and surgical instrument positioning in medical navigation surgery.

Motion target detection system is a system that can give computers the ability to distinguish image information in real time. Motion target detection system aims to obtain information in the field of view through the camera, use the target detection algorithm to detect, identify, locate specific targets in the scene (usually refers to people), and finally output the detection results, so that the machine has similar image vision capabilities to people [1].

The traditional target detection algorithm [2] is only applicable to the detection of objectives with more obvious characteristics and simple background, while in the real world application scenarios, a variety of factors are complex and variable. A variety of factors are complex and changeable, the target shape, detection angle, background environment, etc. will affect the accuracy of identification classification, only through the unified set of abstract features is more difficult to achieve the detection of the target.

In 2014, Girshick et al. proposed the R-CNN algorithm, which successfully applied the deep learning to the target detection field for the first time, but it has many problems with cumbersome steps, slow detection speed and low training efficiency. Kaiming He group proposed the SPPNet detection framework, which increased 24~102 times compared to R-CNN. Ren [3] et al. learned from the design idea of SPPNet to improve R-CNN and proposed the Faster R-CNN algorithm, whose
detection speed and accuracy are better improved, but the speed is too slow to meet the speed requirements of real-time detection [4].

In 2016, Redmon et al. proposed the YOLO algorithm [5], increasing the detection speed to 45Fs, meeting the speed requirements of real-time detection video. In 2017 and 2018, YOLOv2 and YOLOv3 [6] were introduced successively. These two algorithms were optimized by the series founder Joseph Redmon et al. based on the previous algorithm in the same series. They have a good optimization on the detection speed and accuracy. Although it is not Faster R-CNN accurate and cannot detect scenes with dense distribution or small size, it has no small advantages in real-time performance. In 2020, Joseph Redmon, the father of YOLO, announced his retirement from computer vision, and the same year, Bochkovskiy [7] et al. made an improvement to YOLOv3 and released YOLOv4. YOLOv4 is based on the original YOLO target detection framework, using the most advanced algorithm in the field of convolutional neural network in recent years to optimize the model, in the backbone network, activation function, loss function, data processing, network training and other aspects have been optimized accordingly.

YOLOv4, a powerful and efficient target detection network, makes several improvements to the original algorithm to take its effect to the next level and allows it to be trained with a single GPU, enabling everyone to use the GTX 1080Ti or 2080Ti GPU to train a fast and accurate target detector with a significantly reduced hardware cost. YOLOv4 is more flexible and has better overall performance than YOLOv5 [8].

In this paper, a motion target detection system is designed to achieve the target positioning and classification tasks in the general scene. Identification methods include picture recognition, video recognition, and real-time camera recognition.

2. YOLO Algorithm

YOLO (You Only Look Once) is a single-stage target detection algorithm based on regression methods [9], and the basic principle is to enter the entire image, and then the output layer directly returns directly to the classification and location of the Bounding box, with a faster detection speed. The traditional two-phase object detection algorithm [10], such as the R-CNN algorithm, requires multiple steps to complete the object detection, and each module needs to be trained separately. Due to the complex model, there are the disadvantages of many training parameters, long training time and difficult network training optimization. YOLO adopts a completely different processing idea from R-CNN, innovatively proposes the end-to-end detection scheme, solves the detection task as a boundary box and the classification regression task, removes the branch of the candidate area, directly classifies the boundary frame regression of the candidate area on the image, and avoids multiple detection of the same target. Therefore, as the first single-phase object detection algorithm, the YOLO algorithm is significantly faster compared to the existing two-phase object detection algorithm, but it is still poor in terms of accuracy.

The YOLO algorithm is to partition a graph with S*S’s grid and then input it into a convolutional neural network for feature extraction, and then the full connection layer classifies and selects the identified targets. Each of the grids is responsible for detecting the target falling within the grid and predicting B bounding boxes, confidence, and target types, then removing the excess boundary boxes with non-maximum suppression and obtaining the final results. Each grid is to predict the category of the target in which the central point falls [11].

The YOLOv4 feature extraction network adopted in this paper is the CSPDarknet53 + PAnet-SPP structure. The backbone feature extraction network CSPDarknet53 [12], strengthens the feature extraction networks SPPnet and PANet [13], last YOLO head are used to transform the extracted features into prediction results. The CSPdarknet53 consists of a convolutional block and five Resblock_body s containing a single convolution, standardization, and Mish activation functions.

Resblock_body of YOLOv4 utilizes the CSPnet [14] structure, which first convolutions a 3*3 steps of 2 on the incoming feature layer, then produces two branches, Part1 and Par2, Part1 is a large residual edge, connected directly to the end, and Part2 uses residual structure to stack the feature
layers n times. For the backbone Part2, a 1*1 convolution first integrates the number of channels, then enters the loop of the n times 1*1 convolution and 3*3 convolution residual structure, and then performs a 1*1 convolution of the obtained residual structure. Finally the Part2 and the residual edge of the Part1 are stacked and the number of channels is consolidated with a convolution of 1*1.

SPPnet performs three convolutions with the effective feature layer of 13*13 *1024 obtained from the backbone feature extracted network CSP darknet53, maximum pooling of 13*13,9*9,5*5, and then stacking the pooling results with three convolutional cores of 1*1,3*3,1*1.

PAnet convolutions and upsamples the results obtained by SPPnet, convolutions the effective feature layer of 26*26*512, then stacks both, 5 convolution operations, then with the 52*52*256 from CSPDarknet53, then 5 convolution, then a series of downsampling and convolution operations to fuse features to extract more efficient features, and finally outputs three results.

3. Experimental Results and Analysis

3.1. Preprocessing

This paper adopts TensorFlow deep learning framework and YOLOv4 algorithm, and trains the prediction model using VOC 2007 + 2012 Datasets to realize the detection functions of pictures, video, cameras, etc., and complete the recognition and positioning of motion targets.

Using YOLOv4 algorithm, a total of 16551 pictures were trained in the training and verification sets of VOC2007-2012 data set, and three functions were realized: picture detection, video detection and real-time detection. And Model detection performance is enhanced by Mosaic data enhancement, Label Smoothing label smoothing, CIoU, learning rate cosine anneal attenuation, K-means clustering for priority box, and so on.

In this paper, the training effect of the model is improved by Mosaic data enhancement, Label Smoothing label smoothing, CIoU, learning rate cosine anneal attenuation, K-means clustering for priority box, and so on.

This paper uses Mosaic data augmentation, cutting four images into several parts and then combining them into one image for training. Data augmentation can well ease the angular brightness and other additional factors caused by feature extraction errors. In model training, Label Smoothing label smoothing is used to prevent overfitting; the K-means clustering algorithm is used to calculate the width of the a priori box consistent with the dataset to improve the performance of the model.

3.1.1. CIoU. The function of CIoU is to determine the degree of coincidence between the real box and the prediction box, and the higher the value of the CIoU, the more accurate the prediction results, as shown in equation (1):

\[
CIoU = IoU - \frac{\rho^2(b, b')}{c^2} - \alpha v
\]

IoU is the intersection between the real box and the prediction box, the middle term on the right is the distance between the real box and the prediction box divided by the diagonal distance between the real box and the prediction box closure. The closer the real box is to the center of the prediction box, the smaller the value. When the two centers fully coincide, this value is equal to 0, and the higher the degree of coincide, the higher the CIoU. The \(\alpha v\) is related to the width and height of the two boxes. The closer the ratio of the true box to the prediction box width, the lower the value of the \(\alpha v\).

3.1.2. Learning-Rate Cosine Annealing Attenuation. When the gradient descent algorithm is used to optimize the objective function, the Loss value is getting closer to the global minimum, the learning rate becomes smaller to make the model as small as possible from this point. And the cosine annealing uses the characteristics of the cosine function. The cosine value decreases slowly with the increase of x, then it decreases rapidly, and then slowly decreases. This decline mode can cooperate with the
learning rate to produce a good effect in a very effective computational way. The learning rate cosine annealing attenuation effect diagram is shown in figure 1.

3.1.3. Mish Activation Function. Mish is a regular, non-monotonic neural activation function. As shown in figure 2, the Mish activation function diagram appears.

![Figure 1](image1.png) ![Figure 2](image2.png)

**Figure 1.** The learning rate cosine annealing attenuation effect diagram.  
**Figure 2.** The Mish activation function diagram.

The Mish activation function has a smooth, non-monotonic, upper unbounded, lower bound, low cost, good generalization ability and effective optimization ability, which can improve the quality of the results.

3.2. Model Training
The image and video are detected, and the program reads the selected video frame by frame and detects the target in each frame, and then displays it via the OpenCV. The display effect is shown in figure 3 for the video detection diagram.

The results generated by image detection are shown in figures 4 and 5.

3.3. Data Analysis
The respective precision of the 20 categories, recall rate, and AP values are shown in tables 1 and 2: table of different categories correspond to parameter data (a) and (b).

The AP value is plotted as a bar chart, and the calculated mAP value of the model is 69.48 percent. Similar to the fast R-CNN algorithm’s mAP values of 70% and 68.4% on the VOC2007-2012 data set, it can meet the basic target detection needs.
Figure 3. Video detection diagram.

Figure 4. Picture detection results diagram (a).
Figure 5. Picture detection results diagram (b).

Table 1. Table of different categories correspond to parameter data (a).

| Category name | Accuracy (%) | Recall rate (%) | AP (%) |
|---------------|--------------|-----------------|--------|
| Aeroplane     | 91.89        | 61.26           | 70.68  |
| Bicycle       | 88.71        | 53.4            | 64.81  |
| Bird          | 92.63        | 55.35           | 80.15  |
| Boat          | 85.19        | 43.81           | 54.73  |
| Bottle        | 85.04        | 55.67           | 69.94  |
| Bus           | 70           | 53.85           | 54.14  |
| Car           | 93.3         | 60.75           | 78.75  |
| Cat           | 92.42        | 75.78           | 83.03  |
| Chair         | 68.53        | 48.18           | 57.96  |
| Cow           | 92.5         | 48.05           | 74.37  |
| Dining table  | 48.57        | 50.5            | 46.88  |
| Dog           | 86.62        | 66.02           | 75.79  |

Table 2. Table of different categories correspond to parameter data (b).

| Category name   | Accuracy (%) | Recall rate (%) | AP (%) |
|-----------------|--------------|-----------------|--------|
| Horse           | 87.65        | 56.35           | 71.26  |
| Motorbike       | 97.33        | 73.74           | 89.27  |
| Person          | 94.32        | 64              | 86.42  |
| Pottedplant     | 78.33        | 31.97           | 57.25  |
| Sheep           | 86.84        | 42.31           | 64.38  |
| Sofa            | 49.17        | 68.6            | 68.06  |
| Train           | 93.75        | 58.25           | 73.93  |
| TV monitor      | 93.02        | 41.24           | 67.72  |
4. Summary
This paper, a motion target detection system based on the deep learning target detection algorithm, enables the computer to obtain the most basic information from the external image, and lays a foundation for further AI applications. The target detection algorithm based on deep learning can perfectly replace the traditional target detection algorithm under most application scenarios, the detection accuracy and detection speed are greatly improved compared with the traditional algorithm, and the current detection accuracy and speed are the top YOLOv4 algorithm in the target detection algorithm.

Compared with other algorithms, the detection speed and detection accuracy generally have obvious advantage in the case of similar one index. YOLOv4 achieves the speed demand of real-time detection in real time with certain accuracy, and the model mAP value reaches 69.48%.

In the future work, we can try to improve the detection model in the direction of image processing, data processing, etc., so that the recognition effect can be more perfect.

Acknowledgments
This work was supported by Guangdong University Science Research Project (No.2019KTSCX256), the Study of Technology (No.2020KQNCX105), and the Intelligent Science and Technology Team (No.2019XJJXTD01).

References
[1] Cuan Y and Li Y K 2020 Design and implementation of monitoring MoTarget recognition and detection system Smart Computer and Applications 10 (04) 165-169.
[2] Wu X, Song X R, Gao S and Chen X B 2021 A Review of Target Detection Algorithms Based on Deep Learning Sensor and Microsystems 40 (02) 7-18.
[3] Ren S Q, He K M, Girshick R, et al. 2017 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks IEEE Transactions on Pattern Analysis & Machine Intelligence 39 (6) 1137-1149.
[4] Xu D G, Wang L and Li F 2021 A Review of Typical Target Detection Algorithms for Deep Learning Computer Engineering and Application 57 (08) 10-25.
[5] Redmon J, Divvala S, Girshick R, et al. 2016 You only look once: Unified real-time object detection IEEE.
[6] Redmon J and Farhadi A 2018 Yolov3: An Incremental Improvement arXiv e-prints.
[7] Bochkovskiy A, Wang C Y and Liao H 2020 Yolov4: Optimal speed and accuracy of object detection.
[8] Tan X F and Wang Z J 2020 Table Tennis Recognition based on the YOLOv4 improved algorithm Technological Innovation and Application 27 (3) 74-76.
[9] Yang G K 2021 Study on single detection detection of YOLOv4 Electronic World 3(03) 77-81.
[10] Wen J W 2018 Research on object detection algorithms based on deep learning Guangdong University of Technology (05).
[11] Nguyen J Y 2019 Design and implementation of the object detection algorithm based on YOLO Beijing University of Posts and Telecommunications (3).
[12] Guan J L and Zhi X 2020 The mask wear detection method based on the YOLOv4 convolutional neural network Modern IT 4 (11) 9-12.
[13] Yang Y B, Guo Z Y and Cai L M 2021 Study on Vehicle Detection Methods of YOLOv4 Electronic World 3 (05) 80-87.
[14] Xu Z R, Liu and Tan Y T 2020 Vehicle detection and traffic statistics research based on YOLOv4 Modern Information Technology 4 (15) 98-103.