ISSN 2618-1576 Vol. 4, Issue 5: 24-28, DOI: 10.25236/FSR.2022.040507

Multi-Target Detection of Table Tennis Video Based on CNN and Yolov3

Wanyue Li[a], Jie Cui[b]

School of Economics and Management, Shanghai University of Sports, Shanghai, China
[a]liwanyue.sus@outlook.com, [b]1004899295@qq.com

Abstract: *In order to explore the effect of convolutional neural network (CNN) on the detection of athletes and balls in table tennis, and to solve the problems of low accuracy and weak generalization ability of table tennis athletes training data, this paper analyzed videos of table tennis athletes. For the detection of multiple targets in the videos, including athletes and balls, we used Yolov3 as a deep learning framework, and CNN as an automatic detection method when processing images. We trained and test the video data to improve the stability and accuracy of target detection, through modifying its model on the basis of the Yolov3 model. Finally, we detect the movement trajectories of athletes and balls in table tennis videos stably, and the accuracy is above 0.8.*

Keywords: *Convolutional Neural Network; Yolov3 Model; Table Tennis; Multi-Target Detection*

1. Introduction

With the development of the information age, neural network has been developed for a long time. In the past ten years, neural network can be used in many fields such as image classification, recommendation system and knowledge engineering. With the continuous development of neural networks and the integration of related disciplines, the related application is also more extensive and in-depth. The recognition of researchers and the reports of media have made neural networks attract more attention, and image recognition using convolutional neural networks (CNN) has also become a hot research issue in recent years.

Convolutional neural networks are widely used in research in various industries. Shang Zhiliang, Wang Wei and others [1] used convolutional neural network to train garbage images, and the recognition rate of garbage in daily life reached 92.26%, which was 1.09% higher than the original network model. Hou Maoze, Ma Yanqiong and others [2] constructed a water pollution traceability model based on ConvNet convolutional neural network. Their model can successfully identify the three-dimensional fluorescence spectrum of actual wastewater, and the identification accuracy rate is as high as 75%. Lu Fan [3] proposed a model based on convolutional neural network to explore the estimation of human body posture. The results show that this method has advantages and can effectively improve the accuracy of key point positioning of the human body. Zhou Yifeng and Yang Binfeng [4] proposed the convolutional neural network detection model, and used the Bootstrapping algorithm to calculate the athlete's center of gravity position in the video cutting diagram, which achieved a high detection rate and a low false positive rate of the model.

In general, the convolutional neural network has been involved in the environment, water body and athletes in the existing research, but there is less research on the target detection in sports video pictures, especially table tennis athletes and multi-target detection. There is still a lot of experience to dig further.

2. Model Overview

2.1 The Architecture of CNN

Convolutional neural network (CNN) is a kind of neural network, which consists of convolutional layer, pooling layer and fully connected layer. The network structure is shown in figure 1, and it is more suitable for the structure of the image. It has a good effect on image feature extraction and image classification, and the training parameters can be adjusted according to the weight sharing in convolutional neural network. By adjusting the training parameters according to the weight sharing in

ISSN 2618-1576 Vol. 4, Issue 5: 24-28, DOI: 10.25236/FSR.2022.040507

the convolutional neural network, the network structure is simpler and the adaptability is stronger.

When a picture is input, the convolution layer extracts a certain feature of the picture, and its convolution kernel can be regarded as a feature extractor [5]. Multiple convolutions represent multiple extraction of image features. Color images are composed of multiple color channels and display three-dimensional data information. Then the pooling layer selects the features and reduces the number of features, thereby reducing the number of parameters and preventing overfitting. At the same time, pooling can effectively reduce the number of neurons, so that the network remains invariant to some small local morphological changes, such as a wide perception field of vision for fast-moving ping-pong balls. Multiple pooling operations can enhance the invariance due to images being scaled, deformed, and translated [4]. During feature extraction, using a smaller sampling area can increase the data of neurons and extract more data information. The fully connected layer generally uses the softmax function for full connection, and finally obtains the image feature values extracted by the convolutional neural network.

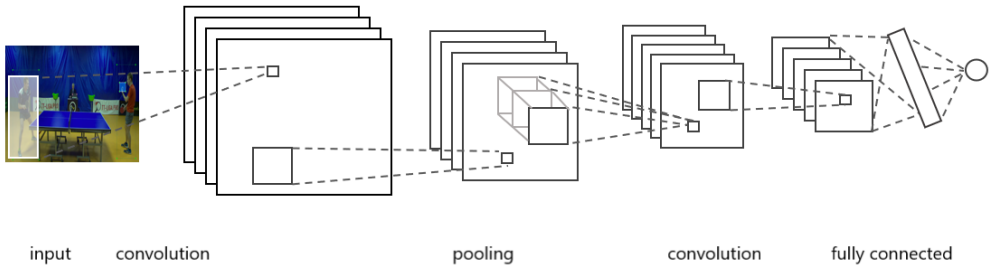

Figure 1: Basic structure of convolutional neural network.

2.2 The Network Architecture of Yolov3

The Yolov3 algorithm is a target detection algorithm suitable for small objects, which improves the accuracy of target detection by dividing the input image into regions for detection. Generally, CNN is used to process images [6], and Yolov3 combined with convolutional layers can also be regarded as a fully convolutional network (FCN). The Yolov3 network structure is mainly divided into three parts, as shown in Figure 2. The Darknet-53 structure is usually called Darknet-53 because there are 53 convolutional layers in the network. The calculation formula is as follows:

$$53 \text{ convolutional layers} = 2 + 1 * 2 + 1 + 2 * 2 + 1 + 8 * 2 + 1 + 8 * 2 + 1 + 4 * 2 + 1 \quad (1)$$

Among them, y1 in the three branches of Yolov3 is suitable for detecting large object targets, such as the target detection of athletes in the input image. Y2 is based on y1 for resampling and feature reconnection, which is suitable for moderate target detection. Y3 is based on y2 for resampling and feature reconnection, which is suitable for detecting small object targets, such as target detection of ping-pong balls in the input image.

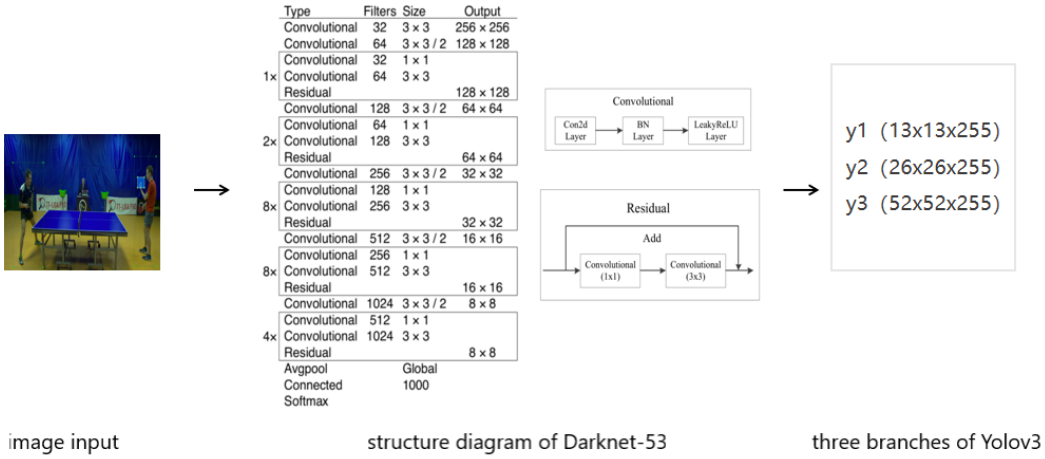

Figure 2: Network structure diagram of Yolov3.

ISSN 2618-1576 Vol. 4, Issue 5: 24-28, DOI: 10.25236/FSR.2022.040507

When the picture is input, the picture will go through multiple layers of depth convolution and achieve the effect of dimensionality reduction. There are full convolution feature extractors in different dimensions, and multiple convolution kernels are interleaved to achieve the purpose of dimension reduction and feature extraction. Then each fully convolutional feature layer is concatenated to obtain the prediction result of each feature layer. The final prediction result is obtained by regressing the result according to the confidence level.

3. Analysis of the Experimental Process

3.1 Environment Configuration

The operating environment is configured as follows:

i7 processor windows 10×64, the development environment is Python 3.6.0 and above, the editing environment is pytorch-gpu 1.5.0, and environments such as opencv-python, matplotlib, and Tqdm are configured.

3.2 Data Set

In this experiment, the training video of table tennis players is used as the training video and test video in the experiment [7]. By intercepting the number of training video frames, 239 video cutting images were generated as the data set of this experiment. The test video is used to verify the effect of this experiment in other videos.

3.3 Ball and Athlete Detection

In this experiment, the detection of table tennis balls and athletes in the cutting diagram is mainly divided into the following four steps.

Step 1: Use the python language to cut the training video into frames, and then manually label the images with labeling. There are two data labels, athlete and ball, as the training set.

Step 2: Use the python language to divide the data set of the above training video into training set, test set, and validation set [8], the ratio is 8:1:1. Then we convert the tagged XML file into a TXT file using Python, in order to use the results in further experiment expediently.

Step 3: Perform preprocessing operations on the established model, and change the classes of the yolo layer under yolov3 and the filters of the fully connected layer (convolutional) according to the categories to be trained in the experiment. The value of classes represents the type of training, and the value of filters depends on the size of the classes.

Step 4: Train and optimize the established model. The model is better, when we modified the epoch parameter and increased the number of training times. As shown in Figure 3, when the number of training times is increased from 100 to 250 times, the average accuracy of the model also increased from about 0.45 to about 0.8.

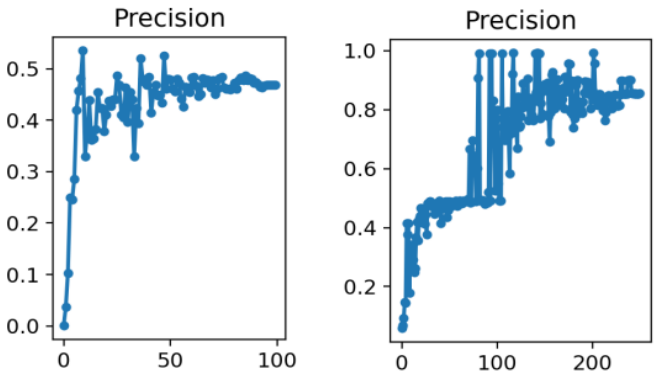

Figure 3: Accuracy comparison of 100 training times (left) and 250 training times (right).

ISSN 2618-1576 Vol. 4, Issue 5: 24-28, DOI: 10.25236/FSR.2022.040507

4. Analysis of Results

The method proposed in this experiment is to first cut the training video into frames, and then manually mark the position of the table tennis athlete and the ball in a specific place in the video cutting map, and then represent the two labels by a rectangular box, and then extract these images, and then iterate and train the data, and then calculate the average accuracy.

The experimental results before modifying the model are shown in Figure 4 on the left. The accuracy of the balls in the test video is only 0.46. In the process of video playback, the balls cannot be detected in multiple seconds, there are many seconds, because of the more light or dark video background and large movements of athletes. This also shows that before the model modification, the original model training effect is not good, and the stability of the detection target is not high.

After modifying the model, the training results show that the average accuracy of athletes and balls is around 0.8. The experimental results of the test video are shown in Figure 4 on the right, and it can be seen that the precision of the athletes' balls also reached above 0.8. In the process of the entire test video playing, there were few cases that the table tennis balls could not be detected. This shows that the modified model detection target is more stable and more accurate, and the model is suitable for other test videos that match to the Angle.

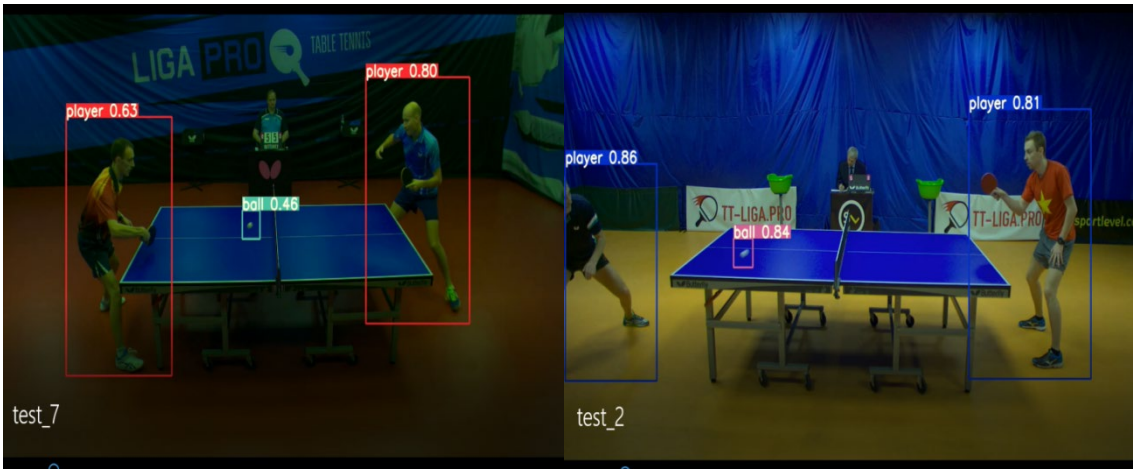

Figure 4: Accuracy comparison of before (left) and after (right) model modification.

5. Conclusion

In this experiment, we introduced convolutional neural network into the target detection algorithm, and proposed a multi-target detection algorithm based on the Yolov3 deep learning framework and regional convolutional neural networks. By modifying the Yolov3 model, we improved the stability of multi-target detection, especially for fast-moving targets like small table tennis balls. Despite this, there is still room for improvement in this experiment, such as fewer training times, and the detection accuracy needs to be improved. We also need to systematically summarize experience, demonstrate and continuously revise the model, in order to make more contributions to the sports cause of our country.

References

[1] Shang Zhiliang, Wang Wei, Yang Mingzhen, Jia Mingzhen, Ma Cunliang. Garbage image processing and improvement based on convolutional neural network [J]. Internet of Things Technologies, 2022,12(08):93-96+99.

[2] Hou Maoze, Ma Yanqiong, Tian Senlin, Ouyang Hao, Zhao Heng, Li Yingjie, Tie Cheng, Zhao Qilin. Research on water pollution traceability based on convolutional neural network identification of three-dimensional fluorescence spectrum[J/OL].Environmental Monitoring in China:1-8[2022-08-26].

[3] Lu Fan. Research on human pose estimation based on deep convolutional neural networks[D]. Southwest Jiaotong University, 2021.

[4] Zhou Yifeng, Yang Binfeng.A palyer detection method using convolution nerual network in sports videos[J].Journal of Xiangtan University(Natural Science Edition),2017,39(01):95-98.

[5] Ouyang Ruiqi, Yong Yang, Wang Bingxue.Application of convolution neural network in aircraft type

ISSN 2618-1576 Vol. 4, Issue 5: 24-28, DOI: 10.25236/FSR.2022.040507

recognition[J].Ordnance Industry Automation, 2017,36(12):71-75.
[6] Yang Lanlan, Gao Mingyu, Wang Chenning, Feng Dongjie, Lv Xinrui. Research on facial expression recognition based on data enhancement[J].Computer Products and Circulation, 2020(11):128-129.
[7] Video source website. https://lab.osai.ai/datasets/openttgames/.
[8] Cao Xiaoming, Zhang Yonghe, Pan Meng , Zhu Shan, Yan Hailiang. Research on student engagement recognition method from the perspective of artificial intelligence: analysis of deep learning experiment based on a multimodal data fusion[J].Journal of Distance Education, 2019,37(01):32-44.