Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.037266
Article

Application of Depth Learning Algorithm in Automatic Processing and Analysis of Sports Images

Kai Yang*

North China University of Water Resources and Electric Power, Henan, 450000, China
*Corresponding Author: Kai Yang. Email: yangkai@ncwu.edu.cn
Received: 28 October 2022; Accepted: 17 February 2023; Published: 26 May 2023

Abstract: With the rapid development of sports, the number of sports images has increased dramatically. Intelligent and automatic processing and analysis of moving images are significant, which can not only facilitate users to quickly search and access moving images but also facilitate staff to store and manage moving image data and contribute to the intellectual development of the sports industry. In this paper, a method of table tennis identification and positioning based on a convolutional neural network is proposed, which solves the problem that the identification and positioning method based on color features and contour features is not adaptable in various environments. At the same time, the learning methods and techniques of table tennis detection, positioning, and trajectory prediction are studied. A deep learning framework for recognition learning of rotating flying table tennis is put forward. The mechanism and methods of positioning, trajectory prediction, and intelligent automatic processing of moving images are studied, and the self-built data sets are trained and verified.

Keywords: Deep learning algorithm; convolutional neural network; moving image; trajectory; intelligent processing

1 Introduction

Sports images in physical education are mainly identified by artificial feature extraction or deep learning methods [1]. It can be seen from the research mentioned above the status of human action recognition methods based on standard artificial features that, limited by the performance of artificial features, it is difficult to handle more complex human action recognition tasks, and the drawbacks are too significant and consume too much human and material resources [2]. In recent years, Convolutional Neural Networks (CNN) have achieved good results in object recognition, image classification, etc. [3]. At the same time, CNN has also been introduced into the video domain for action recognition [4]. The main methods of using CNN for action recognition are 3D Convolutional Neural Networks (C3D), Long-term Recurrent Convolutional Networks (LRCN), two-stream Convolutional Neural Networks, etc. [5]. The application of computer technology in the field of sports is not uncommon. With the continuous improvement of the competitive level of various

sports, to maintain a high economic level and seek a breakthrough, more people use image processing technology to process the competition and training videos of domestic and foreign athletes to assist in the technical and tactical analysis. The image proves imaging technology is to process the video by some man so that technical and tactical analysts can quickly identify the useful technical and tactical information in the video.

In today's Internet world, much data information is generated daily. Compared with text data, picture data is simpler and more intuitive in information communication, so images have become an indispensable information carrier in people's daily lives. Effective retrieval of image information is more important [6]. Early image retrieval was mainly based on text. Later, the content semantics of images appeared. Content-based image classification has become one of the important ways to retrieve images effectively and has been applied in many fields, such as medical imaging, intelligent transportation, sports, etc. [7]. In the field of sports, its related information data proliferates. In addition to the rapid development of the Internet, sports are a fundamental reason for constituting a part of social culture [8]. Sports visual images add many dynamic colors to civilized human life and play an extremely important role in many fields, such as politics, economy, and education [9]. First of all, for the live situation of sports, sports visual images can truly reflect. Secondly, through sports visual images, the essence of sports can be found, which affects the humanistic spirit of the audience and makes people have a positive and healthy attitude towards life [10,11]. Furthermore, those who are engaged in research or study of sports-related work can conduct research and study through sports images [12,13]. In daily life, it can also provide people with visual enjoyment and enrich people's lives [14–17]. With the development of artificial intelligence technology, many CNNs with different structures have appeared. These networks generally have higher accuracy but also have deeper and more complex network structures. Although CNN can achieve real-time performance on some high-performance GPUs or TPUs However, on some mobile devices with limited computing performance, it is difficult to complete the most basic model training [18]. For example, the VGG1612 network, although its classification accuracy is very high, it has hundreds of millions of parameters and operations, which makes it almost impossible to use it on devices with average performance [19]. If you need to use CNN on devices with low computing performance, especially small embedded devices and mobile terminals, you need to solve the problem of too large models [20,21]. Studies have shown that some existing convolutional neural network models have serious over-parameterization problems, which will lead to the problem of high memory usage and high computing costs for the network, so it can be solved from this perspective, and the model can be simplified, which both solves the memory problem and improves speed [22]. The reduction and compression of CNN have long been a key research direction in the field of deep learning [23,24]. Scholars also use deep learning methods to study fading channel parameters, signal distortion, media access control (MAC) protocol, radio signal types, and various identification and classification tasks related to cellular systems [25–28].

This paper mainly studies the human action recognition of table tennis based on a convolutional neural network. The key to action video recognition based on a convolutional neural network is to build a reasonable and efficient network structure. This paper proposes a method of table tennis recognition and positioning based on a convolutional neural network, which solves the problem that the recognition and positioning method based on color features and contour features is not adaptable in various environments, and deeply studies the action features, accurately extracts them and correctly identifies and classifies them. Table tennis action information has the characteristics of spatial information and time series information. Therefore, according to the characteristics of action information, a dual-stream convolutional neural network model is constructed to improve the accuracy of table tennis action recognition. In this paper, without relying on prior knowledge,

by constructing a large number of nonlinear LSTM units to approximate the high-order nonlinear table tennis movement process, the long-term trajectory prediction of table tennis in various rotating states is realized, and the visual perception of the table tennis robot is satisfied. The accuracy and real-time requirements of the system.

2 Establishment and Processing of Data Sets

The experimental data set in this paper is obtained in two ways. First, there is a corresponding public table tennis video dataset on the Internet, namely the UCF101 public dataset. This dataset contains a large number of sports videos, including 13320 videos of 101 kinds, and the size is about 6.6G. Among them, there are 143 video games of table tennis, all of which are shot with unlimited live scenes, including a large number of table tennis matches with rackets. In a large number of videos, the angles are variable, the lighting information and scene information are variable, the video pixels are low, the video length is short, and each video does not exceed 10 s. However, in the video of table tennis in the UCF101 dataset, most athletes are nonprofessionals. The movement technique is relatively simple, and there is no accurate distinction between the technical characteristics of table tennis. The movement data set needs a large number of data sets to verify. Therefore, while using the public data set, this paper also made 109 video data sets, accurately divided the four types of table tennis technical action features, and supplemented the original public data set. The self-made dataset also has lower video frame pixels and shorter video time, which is convenient for post-processing. At the same time, to correctly divide the technical characteristics of table tennis, the UCF101 and the self-made dataset were processed and captured into a short video containing only one technical action. The ratio of the training set to the testing set is about 4:3, and the positive and negative ratio of the training set in this paper is close to that of the testing set.

The UCF101 dataset used and the self-made table tennis dataset are presented in the form of short videos. The two datasets are classified according to the above data forms and stored in folders, respectively. The video format is not acceptable during online training. For the video data set to be processed, the video data format is processed into RGB format by framing. This article uses Python and OpenCV to clip every frame in the video and names the pictures according to the specified format. The specific steps can be divided into introducing cv2 first, using Video Capture to read the video data from the file, and checking whether it is opened normally: after opening successfully, is opened returns to true and then obtains the entire frame number. At this point, we can start to set the frame number. In this paper, keyframes are processed for each short video, and one frame is extracted every three frames. Then, a variable is defined to control the end of the reading video cycle and carry each frame of the image. Use the while loop to read the video frame. After reading the specified frame, the variable that controls the end of the loop is represented by the current frame in the loop until the end of the frame. Finally, put each frame of the generated picture in the custom folder, and name the pictures, in turn, starting from 1. jpg. The processed picture size of each frame is 224×224. After 252 videos are processed, 11453 RGB images are obtained. Then, data tags are generated for the images, and the data of the four table tennis action technologies are scrambled respectively. The training set and test set are randomly composed at a ratio of 7:3, generating a train .list file and a test. List files representing image storage. The path to an action folder for.

Because the training of deep neural networks for human motion recognition needs to rely on a large number of data samples, process the data set to generate a new data set, and then increase the basis of table tennis for the number of video data sets for technical actions. If the data set is too small and the samples are insufficient, there is a risk of overfitting when training the network. When the

training occurs. After size reads the whole batch, you can enhance the original data through a python script to extend the dataset. Such operations are often used in the training of many large data sets to prevent overfitting and improve the robustness of the model. There are several methods:

(1) Image Flipping and Mirroring

Random flipping and mirroring of the training set can make the model more adaptive to the image direction. If the images are flipped and mirrored with the same label, it is necessary to add random image flipping and mirroring to the training model, which can expand the training data set.

(2) Blurring and random brightness change of image

Image blurring and random variation of brightness are common and simple methods in image processing. Image blurring and random variation of brightness are common and simple methods in image processing. Through different random brightness changes, the training data set can cover as many light-brightness scenes under different lighting conditions as possible. Then the reason for using this method is to make noise to the image, which is convenient for the following processing. Noise will produce random errors on the original image and form a new image. Training with these error data can improve the robustness of the network.

3 Convolutional Neural Network

In the early research of visual cells, Wang et al. first proposed the concept of the receptive field [29]. Later, Raghu et al. further proposed a neurocognitive machine on this basis, which is usually regarded as the first implementation of the convolutional neural network [30]. The convolutional neural network has the characteristics of weight sharing, local perception, and downsampling, which makes it have excellent performance in image and speech recognition. As shown in Fig. 1, the structure of a convolutional neural network mainly includes an input layer, convolution layer, pooling layer, full connection layer, and output layer. A convolutional neural network can effectively extract features, especially for two-dimensional images. It can combine feature extraction and training processes to achieve effective feature extraction during network training. The trained network is invariant to light changes, scale changes, and other distortions, have strong versatility, and can be successfully applied in various fields.

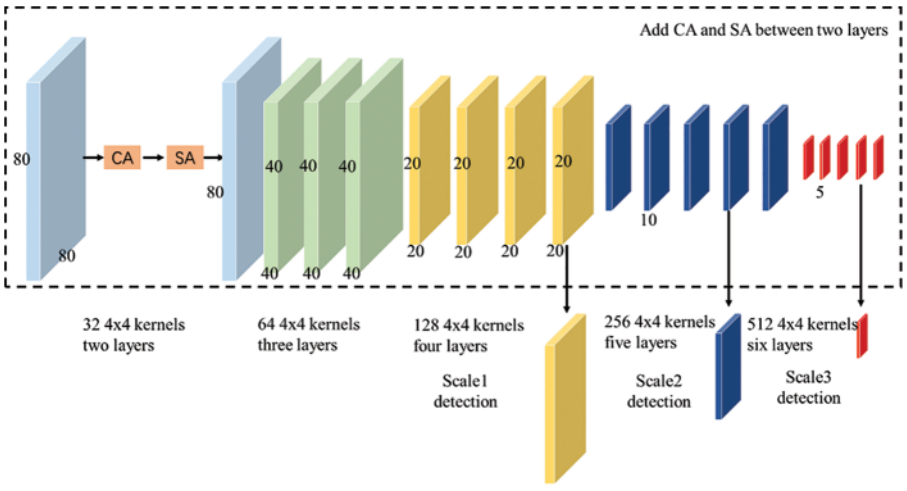

Figure 1: Structure diagram of convolutional neural network

Parameter setting is also very important in network model construction, mainly including iteration number, batch size, gradient descent algorithm, learning rate, classification number, discard rate, etc. Some parameters have been determined by analyzing the results of the comparative experiment in the previous section, while others are selected according to the specific conditions of the experiment. The hyperparameters used in the improved convolutional neural network model in this chapter are: epochs = 100, batch size = 127, optimizer is Adam, learning rate = 0.0006, and class number = 9.

(1) Convolute layer

As an important part of the convolution neural network, the convolution layer is composed of a series of learnable convolution kernels [31]. Convolution is a mathematical operation, whose main purpose is to extract local features from input image data, similar to a filtering operation. A convolutional neural network is still a hierarchical network, which makes use of the spatial correlation between layers to connect each layer only with the neuron nodes of adjacent layers, that is, local connection [32,33]. The convolution kernel glides through the entire image data at a certain step size through a small receptive region and then outputs the operation of the convolution kernel and the original image to form the local feature map of the image. Many different convolution kernels will eventually extract different feature maps, and each feature map represents a certain feature of the original image. Each convolution kernel shares weight and bias, which reduces the number of parameters to learn the network model. As shown in Fig. 2, assuming that the pixel value of a two-dimensional image is composed of O or I, a convolution kernel is used to perform convolution operation with the original image. First, the convolution kernel is superimposed on the corresponding position of the image array. Then, the elements of the convolution kernel are multiplied by the elements of the image array and accumulated to generate a new value. Then, the convolution kernel is repeatedly used to perform the sliding window convolution operation in the image array. This process will eventually generate a new array with dimensions different from the original data. The array usually contains more channels, but its height and width are smaller.

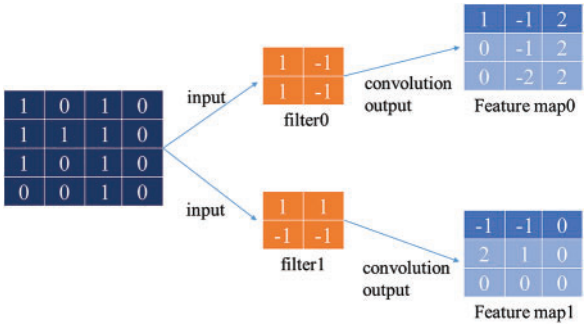

Figure 2: Schematic diagram of the convolution operation

The convolution layer in the convolution neural network is composed of several convolution units. The first layer can usually only extract lower-level features. Only by conducting multi-layer convolution operations can more complex higher-level features be extracted from lower-level features. Finally, the parameters can be learned through the backpropagation algorithm.

(2) Pooling

Pooling is equivalent to dimensionality reduction in the spatial scope so that the network model can extract features in a wider range and improve the model's generalization ability. At the same time, the pooling operation also reduces the size of the input data of the next layer, thereby reducing

the amount of calculation and the number of parameters to prevent overfitting. There are two main methods for pooling the input image: Max Pooling and Average Pooling.

Maximum pooling: the sampling value is the maximum value in the pooled window element, as shown in Fig. 3. 2 × 2 pool the input characteristic graph in the pool window.

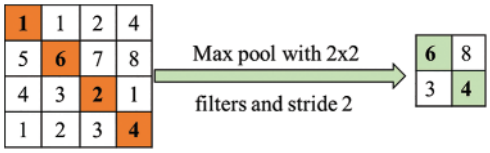

Figure 3: Max pooling diagram

Average pooling: the sampling value is the average value of all elements in the pooling window, as shown in Fig. 4.

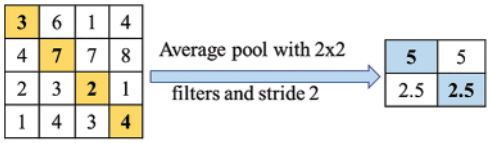

Figure 4: Diagram of average pooling

(3) Full connection layer

The full connection layer is usually set at the end of the network structure. The connection convolution layer and pooling layer synthesize the extracted two-dimensional feature vectors and form the one-dimensional feature vector l. Each neuron in the full connection layer is connected with all the neurons in the previous layer to integrate the local information in the network, so it occupies most of the parameters in the network. In practice, the full join operation can still be implemented by convolution. The main difference between the full connection layer and the convolution layer is that the neurons in the convolution layer and the input data are locally connected, and the neurons share parameters. As shown in Fig. 5, suppose that after n (n ER) convolution layers and pooling layers, five 3's are generated × 3, using 3 of the same size × 3. The convolution kernel performs convolution operations with these five characteristic maps. The output value of a neuron in the full connection layer can be obtained by adding the results of the convolution operation. By analogy, if the whole connective layer contains 100 neurons, then 100 × 5 3 × 3, resulting in too many parameters of the full connection layer. The feature vector of the full connection layer is a high-level abstract expression of the previous network input image, which can highly purify the features. At the same time, the extracted features can also be used to train classifiers such as Softmax. SVM for recognition and classification tasks.

4 Table Tennis Recognition and Positioning

4.1 Table Tennis Recognition Network Construction

To complete the recognition of multiple ping-pong balls in a complex background environment, this paper builds a convolutional neural network that accepts image input and output image categories and chooses to construct the network on the open source, efficient, and stable deep learning framework Caffe (Convolutional Architecture for Fast Feature Embedding). Caffe provides a complete tool kit for building a network, including a modular hierarchical structure, multi-class training algorithms, reference models, etc. It also supports the operation of CPU and GPU with a simplified structure

and fast operation speed. In addition, Caffe also provides C++, Python, Matlab, and other language interfaces for joint debugging. The overall structure of the table tennis recognition network built with the Caffe tool in this paper is shown in Fig. 6.

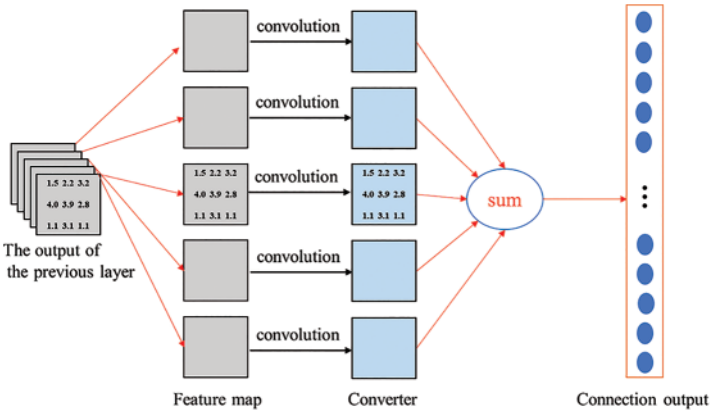

Figure 5: Schematic diagram of neuron output in full connector layer

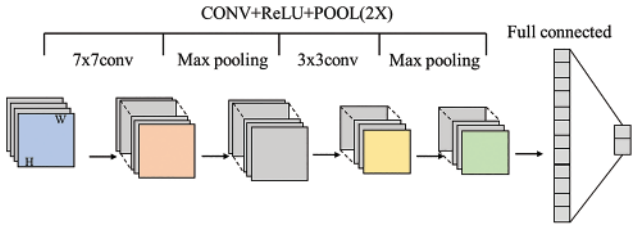

Figure 6: Table tennis recognition network structure

As shown in the figure, the network designed in this paper contains two convolution layers, two maximum pooling layers, two RcLU activation layers, and two full connection layers. The first convolution layer accepts images as input, including 16 with a size of 3×3. The convolution step is 1 pixel. The feature map output from this roll-up layer is input to the second roll-up layer after the activation of the first ReLU layer and the downsampling of the first pooled layer. This layer contains 4 pieces with a size of $3 \times 3 \times 16$; the convolution step is 1 pixel. The feature map output from the second convolution layer also passes through the activation layer and pooling layer and then is input to the full connection layer, which is converted to a size of 2×1 to output the judgment result of whether the image contains a ping-pong ball.

In the network, one convolution layer, one activation layer, and one pooling layer are connected in sequence to form a group of typical feature extraction structures. The convolution layer is responsible for identifying certain features in the graph, the activation layer is responsible for increasing nonlinear operators to increase the richness of features, and the pooling layer is responsible for screening the features with the largest proportion of weight. Because the features of the table tennis ball to be recognized in terms of shape and color are very prominent and maintain high consistency in different images, it can be considered that the feature information that the network needs to extract is relatively concentrated, so the network designed in this paper only contains two groups of feature extraction structures, which will not affect the recognition accuracy while reducing the complexity of the network.

4.2 Network Training of Table Tennis Recognition

Using the table tennis image data set collected in the previous article, this paper trains the table tennis recognition network through the backpropagation algorithm. Network training is also carried out under the Caffe framework. It runs on an x86 host equipped with a Tesla K40c GPU, with 11439 MB of GPU memory.

Since Hinton et al. proposed the backpropagation algorithm (BP) in 1986, the algorithm has been widely used in the back training of neural networks due to its accuracy and ease of use. This paper also chooses this algorithm to train the table tennis recognition network. The derivation process of the BP algorithm for the network training proposed in this paper is as follows.

When the $O_k = [o_{k1}, o_{k2}]$ image is input, the network will output a two-dimensional vector. Set the category label of the image as y. Select the Softmax Loss layer applicable to the classification problem to calculate the error. If it is substituted into the formula, the error between the network output and the label is:

$$L(y_k, O_k) = \log\left(\sum_{j=1}^{2} e^{o_{kj}}\right) - o_{y_k} \tag{1}$$

This error is used to reverse-train the parameter of the upper layer of the output layer; that is, the influence of the error on the parameter is required. Since there is no direct relationship between the error term and the parameter term, it needs to be disassembled through the chain rule:

$$\frac{\partial L}{\partial W_N} = \frac{\partial L}{\partial O_k} \cdot \frac{\partial O_k}{\partial W_N} \tag{2}$$

If the output of the layer above the output layer is X_k, and the output layer offset is b, then $O_k = W_N \cdot X_k + b$, as shown in Fig. 7.

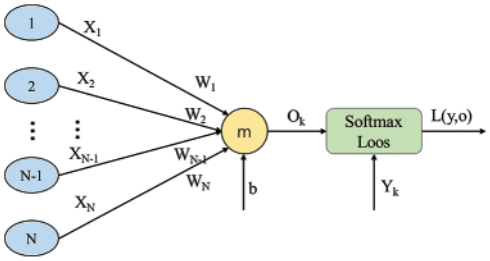

Figure 7: Schematic diagram of error function

The formula after disassembly represents the derivation of the Softmax Loss function O_k, which is calculated as follows:

$$\frac{\partial L}{\partial O_k} = \frac{\partial}{\partial O_k}\left(\log\left(\sum_{j=1}^{2} e^{o_{ij}}\right) - o_{y_k}\right) = \frac{e^{O_k}}{\sum_{j=1}^{2} e^{o_j}} - \delta_{kj} \tag{3}$$

O_k takes the derivative of W_N:

$$\frac{\partial O_k}{\partial W_N} = \frac{\partial}{\partial W_N}(W_N \cdot X_N + b) = X_N \tag{4}$$

Substituting into the above equation:

$$\frac{\partial L}{\partial W_N} = \frac{\partial L}{\partial O_k} \cdot \frac{\partial O_k}{\partial W_N} = \left(\frac{e^{O_k}}{\sum_{j=1}^{2} e^{o_{kj}}} - \delta_{ky} \right) X_N \tag{5}$$

The weight is updated according to the gradient descent method, and its change amount should follow the direction of the negative gradient of the error. Set the weight change rate as n, and the weight change amount can be obtained as:

$$\frac{\partial L}{\partial W_N} = \frac{\partial L}{\partial O_k} \cdot \frac{\partial O_k}{\partial W_N} = \left(\frac{e^{O_k}}{\sum_{j=1}^{2} e^{o_{ij}}} - \delta_{ky} \right) X_N \Delta W = -\eta \frac{\partial L}{\partial W_N} = -\eta \left(\frac{e^{O_k}}{\sum_{j=1}^{2} e^{o_{jj}}} - \delta_{ky} \right) X_N \tag{6}$$

In conclusion, using the chain rule and gradient descent method, the input weight parameters of the output layer can be updated according to the error value between the network output and the label. Similarly, the steps to update the weights of other hidden layers of the network according to the error value are the same.

4.3 Experimental Results and Analysis of Table Tennis Recognition Network

Because the image with a smaller scale is faster during training, this paper first uses the size of 64 × 48. During the training process, the accuracy of the test will be more than 99% if the test is conducted after every 100 iterations.

After training the network with a small map dataset, based on not changing the network structure, directly use the initially trained parameters as the initial value of the network and further fine-tune the network with a large map dataset to speed up the training of the large map dataset. In this data set, 5000 images are also randomly selected as the training data set, and another 1000 images are used as the test data set.

Through the training of small image datasets and large image datasets, the network can better suppress the interference of the environment background and extract the characteristics of table tennis. The characteristic diagram of the output of the two-layer convolution layer in the network is shown in Fig. 8. The two input images both select the pictures of yellow and white bicolor balls under the white edge background of the table. The white part of the characteristic diagram represents the area with a large weight value, and the black part represents the area with a small weight value. It can be seen that the part with the largest weight value in the two images corresponds to the area where the table tennis ball is in the original image.

After preliminary training, the table tennis recognition network proposed in this paper has fully met the requirements of recognition accuracy. To make the network reach the standards of high accuracy and high efficiency at the same time, this paper designs a group of contrast experiments. By comparing the accuracy and running time of the network under different structures, we find the network structure that can complete the identification faster without sacrificing accuracy. A total of 6 network structures were designed in the experiment. Each network has passed 6000 training sessions of 5000 large map datasets without taking any pre-training measures, and then 1000 large maps are used as tests to calculate the average recognition accuracy and average recognition time of the 100 tests. Each training and test run on the same x86 host with Tesla K40c GPU. The experimental results are shown in Fig. 9.

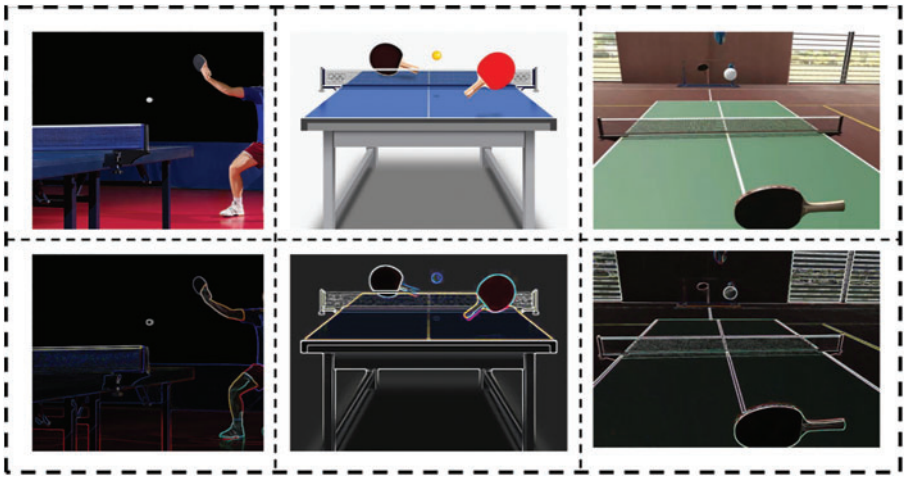

Figure 8: Output characteristics of the convolution layer under different inputs

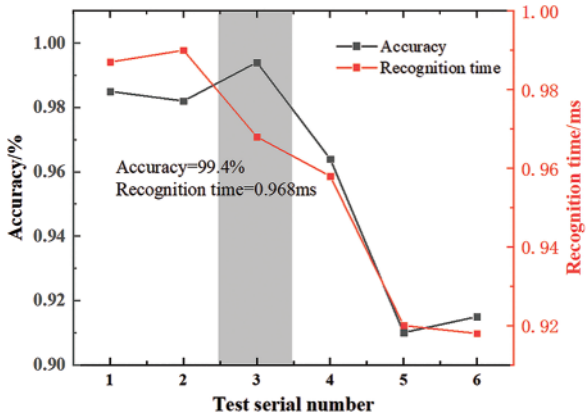

Figure 9: Comparison of accuracy and time under different network structures

It can be seen from the figure that with the reduction of the number of convolution layers (such as the comparison between Group 1 and Group 6) and the reduction of the number of nodes in the full connection layer (such as the comparison between Group 1 and Group 3), the recognition time of the network decreases significantly, but the recognition accuracy also decreases. The increase in convolution kernel size has no significant effect on the recognition time but decreases the recognition accuracy. In six experiments, the third group of networks achieved 1.55 ms identification time and kept the accuracy above 99% by appropriately reducing the number of nodes in the full connection layer and maintaining the structure of the two-layer convolution layer, the nodes of the connection layer from 1 to 6 are 95, 96, 35, 15, 35 and 95, respectively, and the convolution layers are 3×3, 5×5, 3×3, 3×3, 3×3 and 3×3, respectively. After careful consideration, this paper considers that the third group of network structures is the optimal structure of the current network.

To compare with the traditional classification methods based on color features and better analyze the performance of this network. The test data set used in the experiment is also from the large image data set collected in the previous article. Five hundred images are randomly selected from the images with category 1 (including table tennis) to form a positive example set O (excluding table tennis ball).

The set of positive and negative examples includes three kinds of table tennis balls and five kinds of lighting conditions. The precision of the two methods can be calculated from Fig. 9 P, and recall ratio R, the quasi-rate p, and the full rate R are calculated according to the formula method.

By analyzing and comparing experimental data, it can be seen that the color-based comparison method has a faster recognition speed, but because the data set contains table tennis images with multiple colors and multiple lighting conditions, the adaptability of this method are poor, and its can only accurately identify balls in a single color and a fixed environment, so the recognition recall rate is low. The recognition method based on the neural network proposed in this paper can accurately identify the table tennis balls under various colors and illumination in the data set with high recall and precision. However, due to the complexity of the operation mechanism of the neural network, the recognition speed of this method is lower than that of the comparison method. However, considering that the frame rate of the visual system is 120 Hz, the visual perception algorithm of table tennis only needs to complete all operations inside, so the recognition speed of this method is enough to meet the real-time requirements.

To sum up, the table tennis ball recognition method based on the convolutional neural network proposed in this paper can judge whether a 640×480 picture contains a table tennis ball in a relatively short time and can more accurately identify the table tennis ball image under three colors and five lighting conditions, meeting the accuracy and real-time requirements of the table tennis robot visual perception system. In addition, the network proposed in this paper can achieve the above effect only by using a relatively simple structure. If you want to further distinguish more colors, more environment table tennis, or other balls, you can also do further research by deepening the network structure, increasing the number of training times, and using other methods.

4.4 Table Tennis Detection Based on Convolutional Neural Network

As the primary task of the table tennis robot system, the vision system needs to ensure the fast recognition speed and accurate target recognition and location of the table tennis ball. Early research was mainly based on the color and contour features of table tennis balls. This method has great advantages in detection speed, but its disadvantages are also obvious and difficult to overcome. For example, it is easy to be interfered with by light, color, and other external factors. When the ambient light intensity is large or the surrounding color is similar to the color of the table tennis ball, and the color of the table tennis ball changes, the recognition accuracy of this method will be greatly reduced.

In recent years, deep learning has continuously made breakthrough achievements and has also made continuous innovation and development in image recognition, target detection, and other fields. Compared with the traditional target detection algorithm, the target recognition method based on the convolutional neural network has higher robustness and stronger anti-interference ability and is suitable for detection tasks in various environments. However, a large number of parameters bring about the problem of low detection speed, and the detection effect of deepening the network depth is not ideal for small targets. This paper, based on a convolutional neural network, balances the detection speed and accuracy and builds a network suitable for table tennis target detection; In the design of a convolutional network, this paper designs many detection techniques for small targets, allowing deeper networks. It still has a strong detection ability for small targets.

In the learning process of neural networks, we need to learn a lot of labeled data. In the case of limited data, the method of data enhancement can increase the diversity of training samples to a certain extent and prevent the model from meeting the situation of fitting and matching in the training process. For example, by clipping multiple images and then splicing the cropped images, the sensitivity

to object position is reduced in the process of learning these images so that in the process of network recognition, the location information of objects will be less. Impact on the identification results. We can also adjust the brightness, saturation, hue, and contrast of objects to reduce the dependence of the model on color. In recent years, in the development of deep learning, different data enhancement methods have been proposed, and good experimental results have been achieved.

Data enhancement can be divided into two categories: offline enhancement and online enhancement. When the data set is small, offline enhancement can be used to save time in obtaining new data. Online authentication is usually used for large data sets. During network training, each batch of data is acquired and then rotated, translated, flipped, folded, and other changes are made to the data. Many machine learning models have gradually been able to support this enhancement method and can use GPU for optimization calculation.

In recent years, mainstream data enhancement technologies include Mixup, Cutout, Cutmix, Mosaic enhancement, etc. Mixup is mixing two random images in a certain proportion. When classifying, it will also be allocated according to the proportion of each image. The implementation method is shown below.

$$\tilde{x} = \lambda x_i + \left(1 - \lambda x_j\right) \tag{7}$$

$$\tilde{y} = \lambda y_i + \left(1 - \lambda y_j\right) \tag{8}$$

where and $\left(x_j, y_j\right)$ are sample data randomly selected from the training set samples, and $\lambda \in [0, 1]$. Therefore, mixup can expand the distribution of training by combining prior knowledge, that is, linear interpolation of feature vectors. The implementation of mixup is simple and does not increase a lot of calculation, so it is one of the good data enhancement methods according to yours and in the article. The starting point of the Cutout is similar to the method of random erasure.

Considering that this paper only needs to detect the table tennis target, the table tennis ball occupies fewer pixels in a picture, which causes a waste of training costs in network training. Moreover, the feature map after convolution is dozens of times smaller than the original image, and the network's learning ability for small targets such as table tennis is greatly reduced. This paper refers to Mosaic enhancement, and copies the table tennis ball in the picture during network training, The ping-pong ball in the picture is copied three times without affecting other objects. During network initialization, there will be three prior boxes for each target to detect. Then, in each copied image, there will be 12 prior boxes to learn the characteristic information of the ping-pong ball target together, which effectively strengthens the richness of the data set and prevents the network from overfitting.

In the experiment of data enhancement, this paper first used Mosaic enhancement. During the network training, four pictures were spliced into one picture and sent to the network for training. However, the detection effect of table tennis was not improved, and the detection accuracy was also reduced. This is because the size of the table tennis ball was scaled after splicing, and the network was more difficult to obtain the position information of the table tennis ball, which caused the problem of decreased positioning accuracy. In this paper, the idea of Mosaic enhancement is referred to, combined with the target characteristics of table tennis and the needs of the subject. The table tennis ball in each picture is copied three times, which does not change the resolution of the input image, but also increases the diversity of training samples. The network has improved detection accuracy to a certain extent. As shown in Fig. 10.

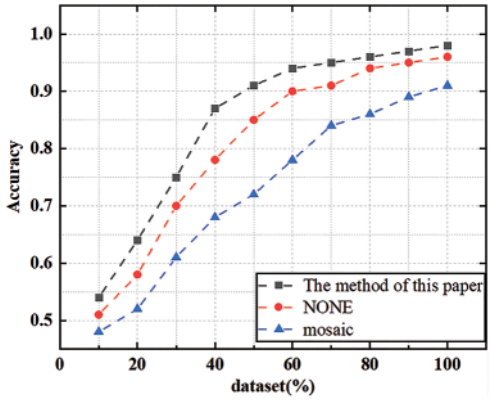

Figure 10: Comparison results of data enhancement accuracy

It can be seen from Table 1 that the current one-stage YOLO series network models have achieved high accuracy in the table tennis target detection task, but the detection speed is insufficient. Their network depth is deeper, and the amount of computation and parameters are also larger, which is not suitable for this research task. The traditional color segmentation algorithm has an advantage in detection speed. It can detect the target quickly, but the detection accuracy is not ideal. Especially when the color of the table tennis ball changes, the detection ability of this method will decline again. The network structure used in this paper meets the detection speed requirements of real-time hits of table tennis and has a very high accuracy [29].

Table 1: Test comparison results

Object detection network	Detection speed	Precision
YOLOv3	20 ms	94.6%
YOLOv4	19 ms	96.1%
The network structure of this paper	19 ms	98.5%
Color segmentation algorithm	2.2 ms	76.2%

After experiments and summaries, the table tennis detection network designed in this paper can achieve high-precision detection of table tennis targets under different complex scenes and interference, but the requirements for high-precision cannot fully meet the requirements of the table tennis robot vision system, and the detection speed needs to be compared. Therefore, this paper designs some comparative experiments to compare the detection accuracy and speed of this network with other existing methods. It is hoped that the network detection speed can also be significantly improved.

5 Conclusion

In this paper, a table tennis ball recognition and location method based on a convolutional neural network are proposed to solve the problem that the recognition and location method based on color features and contour features is not adaptable in various environments. If we want to further distinguish table tennis in more colors and environments, or other ball games, we can further study it by deepening the network structure and increasing the training times. With the self-built table tennis image dataset for training and testing, complete the recognition of table tennis balls in the picture and

image plane positioning, and solve the prediction method state based on motion modeling in various table tennis rotations. In this paper, table tennis detection based on convolutional neural network studies the learning methods and techniques of detection and positioning of table tennis, a typical high-speed rotating flying object, and puts forward a deep learning framework, positioning, and trajectory prediction for the recognition and learning of rotating flying table tennis. The table tennis detection network designed in this paper can realize the high-precision detection task of table tennis targets under different complex scenes and interferences, and at the same time, the network detection speed can be improved.

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] V. Senchenko, V. Lopatina and A. Butsanets, "Calculating the longitudinal and vertical displacements of a moving object by digital image processing methods," *E3S Web of Conferences. EDP Sciences*, vol. 258, no. 5, pp. 02005, 2021. https://doi.org/10.1051/e3sconf/202125802005

[2] E. Ieno and T. C. Pimenta, "Decision-making system for detection of moving vehicles using a field programmable gate array combining conventional techniques of digital image processing with a fuzzy integral," *Journal of Electronic Imaging*, vol. 27, no. 4, pp. 02005, 2018. https://doi.org/10.1117/1.JEI.27.4.043001

[3] N. Carroll, "Towards an ontology of the moving image," in *Aesthetics*, 1st ed., vol. 1. New York: Routledge, pp. 101–112, 2017. https://doi.org/10.4324/9781315303673

[4] M. Uzair, R. S. A. Brinkworth and A. Finn, "Bio-inspired video enhancement for small moving target detection," *IEEE Transactions on Image Processing*, vol. 30, no. 1, pp. 1232–1244, 2020. https://doi.org/10.1109/TIP.2020.3043113

[5] C. Yu, X. Wang and J. Zheng, "Computation of wind wave flow field with moving boundary based on image processing," *Tehnički vjesnik*, vol. 28, no. 4, pp. 1075–1081, 2021.

[6] C. Chen and D. Li, "Research on the detection and tracking algorithm of moving object in image based on computer vision technology," *Wireless Communications and Mobile Computing*, vol. 2021, no. 4, pp. 1232–1244, 2021. https://doi.org/10.1155/2021/1127017

[7] D. Li and X. Chen, "Research on moving target statistic algorithm based on image processing," in *2019 22nd Int. Conf. on Electrical Machines and Systems (ICEMS)*, IEEE, vol. 11, no. 2, Harbin, China, pp. 1–4, 2019.

[8] M. Jian, W. Zhang, H. Yu, C. Cui, X. Nie *et al.,* "Saliency detection based on directional patches extraction and principal local color contrast," *Journal of Visual Communication and Image Representation*, vol. 57, no. 1, pp. 1–11, 2018. https://doi.org/10.1016/j.jvcir.2018.10.008

[9] C. Premachandra, S. Ueda and Y. Suzuki, "Detection and tracking of moving objects at road intersections using a 360-degree camera for driver assistance and automated driving," *IEEE Access*, vol. 8, no. 6, pp. 135652–135660, 2020. https://doi.org/10.1109/ACCESS.2020.3011430

[10] X. Lu, M. Jian, X. Wang, H. Yu, J. Dong *et al.,* "Visual saliency detection via combining center prior and U-Net," *Multimedia Systems*, vol. 28, no. 2, pp. 1689–1698, 2022. https://doi.org/10.1007/s00530-022-00940-8

[11] D. Lv, "Scale parameter recognition of blurred moving image based on edge combination algorithm," *International Journal of Computing Science and Mathematics*, vol. 15, no. 2, pp. 168–182, 2022. https://doi.org/10.1504/IJCSM.2022.124002

[12] M. Jian, J. Wang, H. Yu, G. Wang, X. Meng *et al.,* "Visual saliency detection by integrating spatial position prior of an object with background cues," *Expert Systems with Applications*, vol. 168, no. 4, pp. 114219, 2021. https://doi.org/10.1016/j.eswa.2020.114219

[13] N. Safaei, O. Smadi, A. Masoud and B. Safaei, "An automatic image processing algorithm based on crack pixel density for pavement crack detection and classification," *International Journal of Pavement Research and Technology*, vol. 15, no. 1, pp. 159–172, 2022. https://doi.org/10.1007/s42947-021-00006-4

[14] M. Jian, J. Wang, H. Yu and G. G. Wang, "Integrating object proposal with attention networks for video saliency detection," *Information Sciences*, vol. 576, no. 2, pp. 819–830, 202. https://doi.org/10.1016/j.ins.2021.08.069

[15] X. Li and H. Zheng, "Target detection algorithm for dance moving images based on sensor and motion capture data," *Microprocessors and Microsystems*, vol. 81, no. 1, pp. 103743, 2021. https://doi.org/10.1016/j.micpro.2020.103743

[16] M. Eshkevari, M. J. Rezaee, M. Zarinbal and H. Izadbakhsh, "Automatic dimensional defect detection for glass vials based on machine vision: A heuristic segmentation method," *Journal of Manufacturing Processes*, vol. 68, no. 5, pp. 973–989, 2021. https://doi.org/10.1016/j.jmapro.2021.06.018

[17] Y. Shen, F. Gou and Z. Dai, "Osteosarcoma MRI image-assisted segmentation system base on guided aggregated bilateral network," *Mathematics*, vol. 10, no. 7, pp. 1090, 2022. https://doi.org/10.3390/math10071090

[18] Y. Jiang, C. Li, Y. Zhang, R. Zhao, K. Yan *et al.,* "Data-driven method based on deep learning algorithm for detecting fat, oil, and grease (FOG) of sewer networks in urban commercial areas," *Water Research*, vol. 207, no. 3, pp. 117797, 2021. https://doi.org/10.1016/j.watres.2021.117797

[19] Y. Zhang, C. Li, Y. Jiang, L. Sun, R. Zhao *et al.,* "Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model," *Journal of Cleaner Production*, vol. 354, no. 5, pp. 131724, 2022. https://doi.org/10.1016/j.jclepro.2022.131724

[20] M. T. Cao, K. T. Chang, N. -M. Nguyen, V. -D. Tran, X. -L. Tran *et al.,* "Image processing-based automatic detection of asphalt pavement rutting using a novel metaheuristic optimized machine learning approach," *Soft Computing*, vol. 25, no. 20, pp. 12839–12855, 2021. https://doi.org/10.1007/s00500-021-06086-5

[21] W. Jia, S. Xu, Z. Liang, Y. Zhao, H. Min *et al.,* "Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector," *IET Image Processing*, vol. 15, no. 14, pp. 3623–3637, 2021. https://doi.org/10.1049/ipr2.12295

[22] L. Huang, G. N. McKay and N. J. Durr, "A deep learning bidirectional temporal tracking algorithm for automated blood cell counting from non-invasive capillaroscopy videos," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Cham, Springer, vol. 12908, no. 2, pp. 415–424, 2021.

[23] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 Int. Conf. on Engineering and Technology (ICET)*, vol. 5, Antalya, Turkey, pp. 1–6, 2017.

[24] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, "A Survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022. https://doi.org/10.1109/TNNLS.2021.3084827

[25] H. Chen, Y. Zhang, Y. Cao and J. Xie, "Security issues and defensive approaches in deep learning frameworks," *Tsinghua Science and Technology*, vol. 26, no. 6, pp. 894–905, 2021. https://doi.org/10.26599/TST.2020.9010050

[26] Y. N. Malek, M. Najib, M. Bakhouya and M. Essaaidi, "Multivariate deep learning approach for electric vehicle speed forecasting," *Big Data Mining and Analytics*, vol. 4, no. 1, pp. 56–64, 2021. https://doi.org/10.26599/BDMA.2020.9020027

[27] Y. Zhou, H. Alhazmi, M. H. Alhazmi, A. Almarhabi, M. Alymani *et al.,* "Radio spectrum awareness using deep learning: Identification of fading channels, signal distortions, medium access control protocols, and cellular systems," *Intelligent and Converged Networks*, vol. 2, no. 1, pp. 16–29, 2021. https://doi.org/10.23919/ICN.2021.0004

[28] Q. Cao, W. Zhang and Y. Zhu, "Deep learning-based classification of the polar emotions of "moe"-style cartoon pictures," *Tsinghua Science and Technology*, vol. 26, no. 3, pp. 275–286, 2020. https://doi.org/10.26599/TST.2019.9010035

[29] J. Wang, Y. Chen, R. Chakraborty and S. X. Yu, "Orthogonal convolutional neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 55, no. 24, pp. 11505–11515, 2020. https://doi.org/10.1109/CVPR42600.2020

[30] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Advances in Neural Information Processing Systems*, vol. 34, no. 4, pp. 12116–12128, 2021.

[31] J. Hu, X. Weng, L. Yang, S. Leiv and H. Niu, "Centrifugal modeling test on failure characteristics of soil-rock mixture slope under rainfall," *Engineering Failure Analysis*, vol. 142, no. 4, pp. 106775, 2022. https://doi.org/10.1016/j.engfailanal.2022.106775

[32] K. Ramey, "Expanded visions: A new anthropology of the moving image, by Arnd Schneider," *Alphaville: Journal of Film and Screen Media*, vol. 23, no. 7, pp. 115–119, 2022. https://doi.org/10.33178/alpha

[33] Q. Liu and H. Ding, "Application of table tennis ball trajectory and rotation-oriented prediction algorithm using artificial intelligence," *Frontiers in Neurorobotics*, vol. 18, no. 8, pp. 12–17, 2022. https://doi.org/10.3389/fnbot.2022.820028