Membrane Technology
ISSN (online): 1873-4049

Research on Table Tennis Rotation Prediction Based on Deep Learning and Multi-scale Feature Fusion

Wujunyu Xu
Wuxi institute of technology, Wuxi, Jiangsu ,214121, China

Abstract

This paper designs a table tennis target detection network based on feature fusion network. In the feature extraction network, the cross-layer connection network (CSPNet) is used to strengthen the learning ability of the convolutional neural network and reduce the number of parameters of the network to improve the detection speed of the network; for the existing network, the detection accuracy of small targets such as table tennis is low and the positioning ability is low. For the problem of poor, this paper uses a feature fusion network, adds a bottom-up connection on the basis of the feature pyramid network, and finally uses an adaptive pooling method to fuse the feature information of each feature map, and the upper-level semantic information rich features the layer and the lower layer are connected with the feature layer rich in target location information to perform feature fusion to enhance the network's ability to locate small targets. Because the network in this paper only needs to detect the table tennis target, and the table tennis target in a single picture is small, resulting in a waste of training costs, this paper proposes a new data augmentation method, which combines the data in each picture during training. Multiple copies of Ping Pong, based on the existing data, further increase the richness of the data set. After adjusting and optimizing the network structure, the network of this paper can complete real-time tracking and accurate positioning of table tennis under different background and lighting conditions.

This paper builds a LSTM-based rotating table tennis trajectory prediction network. By stacking LSTM networks, the task of predicting the trajectory of table tennis can be realized, and real-time and certain accuracy can be met. For different types of spinning balls, there are specific trajectory changes in their trajectories to follow; therefore, this article attempts to infer the general type of ping pong ball rotation based on the flight trajectory of the ping pong ball. This article decomposes the table tennis movement into three coordinate movements, and calculates the movement speed of the table tennis in the three coordinate axis directions through the obtained three-dimensional coordinate information of the first five moments of the table tennis, passing the speed and the set threshold of Compare the type of rotation to judge the general type of rotation, and control the end joint of the robot to hit the ball at a specific angle for the ball of different rotation type, which effectively improves the success rate of the table tennis robot. Compared with the traditional physical model, the network in this paper has higher anti-interference ability and accuracy.

Keywords: Rotating table tennis; deep learning; multi-scale feature fusion; LSTM

1 INTRODUCTION

Target detection and trajectory prediction of rotating objects have great practicability and importance in various fields. For example, in the military field, it is necessary to accurately locate the target of the enemy's high-speed movement to make effective prevention in advance; in sports, the trajectory of fast-moving balls can help athletes and coaches conduct effective scientific analysis and judgment, and improve competition. Ability; application results in related fields can also be used in aerospace, industrial production and service robots to complete dynamic interactive tasks.

Table tennis is a sport with strong competitive, ornamental and interesting properties. It is very popular all over the world and has a very large mass base and enthusiasm in China. In international table tennis competitions, China’s table tennis Ball players also often achieve excellent results. The research of table tennis robots can help table tennis players to practice in daily training, saving training costs and improving training efficiency. Many aspects of knowledge learning are involved in the field of table tennis robot research, such as deep learning, robot kinematics, robot control, etc., which have very high research value and significance.

Table tennis has the characteristics of fast movement speed, high-speed rotation, small size and small mass, which brings great research difficulty to the research of table tennis. Since the 1980s, a large number of scholars have begun to study table tennis robot systems. Traditional research methods are mainly based on the color, contour and other characteristics of the ping-pong ball. However, the surrounding environment in the actual scene will cause great interference to the detection algorithm and cause the problem of low detection accuracy, and it needs to be set according to the color of the ping-pong ball. The detection threshold is set, which results in the limitations of the application scenario; most of the table tennis trajectory prediction methods are based on the establishment of physical models, and then the parameters of the model are solved. When trajectory

prediction is performed, the physical model is mainly based on the information at the current moment and the previous several moments of information are used to predict the next position information, and then the predicted information is used to predict the next moment. When long-term predictions are required, errors are likely to accumulate and the prediction accuracy decreases; table tennis with rotation is or high-order nonlinear motion, the physical model is difficult to approximate the real motion and can only adapt to certain specific rotational motions. The research in this paper is based on the seven-degree-of-freedom KUKA robotic arm system, which can more flexibly simulate human arm movements to control the angle of the racket. It is the basis for the robot to receive and even hit the rotating ball.

According to the requirements for real-time and accuracy of the table tennis robot vision system, this article focuses on the shortcomings and problems of traditional methods, combined with the current fast-developing neural network method to carry out research on table tennis target detection and spinning ball trajectory prediction based on deep learning, and based on the seven-degree-of-freedom KUKA robotic arm physical system for verification.

2 LITERATURE REVIEW

2.1 Research status of target detection and trajectory prediction

In the process of table tennis, the robot's vision system needs to complete the following subtasks: ①Detect the table tennis target. ②Locate the position of the table tennis. ③According to the historical position of table tennis, accurately predict its trajectory. It is mainly divided into two parts: target detection and trajectory prediction. In these two fields, many excellent research results and methods have been adopted by scholars. Traditional target detection methods have the characteristics of fast detection speed, but the detection accuracy is often affected by different factors; the development of deep learning networks continues now that it is mature, traditional target detection methods are gradually being replaced. The method based on deep learning has the advantages of extremely high detection accuracy and strong robustness, but the large number of parameter calculations of the neural network brings the problem of low detection speed, and a large number of data sets are required to train the network.

Trajectory prediction mainly uses Kalman filtering, the establishment of table tennis flight models and rebound models, and curve fitting methods. The curve fitting method is to obtain a certain number of frames of position information to solve a specific function $f(x)$ parameters to calculate the position information at the future time. The idea of establishing the flight model and rebound model of the table tennis is to calculate the flight status of the table tennis through force analysis. And then approximate the movement of the table tennis on the x and y axes. It is a linear motion, the z -axis is approximated as a quadratic function curve, which is solved by curve fitting, and then a rebound model of the table tennis ball is established through SVM. And a function $f(x)$ that is as consistent as possible with the actual data y is learned and obtained the entire flight data of table tennis. The Kalman filter method is to predict the position information at $k+1$ based on the current time k and the position information at the previous few moments. And then use the information at $k+1$ time and k time as input to predict the position information at $k+2$ time. Iteratively realizes long-term prediction of data, but the error of this method will continue to accumulate, and the prediction accuracy is not ideal.

2.2 Current status of target detection research

The principles of traditional target detection methods can be roughly divided into optical flow method [1], color segmentation method [2], background difference method [3] and inter-frame difference method [4]. The optical flow method uses the changes in the time domain of the pixels in the image sequence. The correlation between adjacent frames to find the correspondence between the previous frame and the current frame. Thereby calculating the position information of the moving object. The use of optical flow method needs to be based on two assumptions: ①the brightness of the moving object is constant. ②the time of the movement of the object is continuous or the movement speed of the object is not very fast. Once the light in the scene changes greatly or the speed of the object is extremely fast, the detection success rate of the optical flow method will be greatly reduced, and the calculation complexity is relatively high. The inter-frame difference method is to calculate the difference between adjacent frames in a piece of video. The background is filtered in the process of making the difference, while the moving object is retained. By judging the absolute value of the gray difference, it exceeds a certain threshold. It can be judged as a target to realize the detection of the moving target. Its advantage is that the amount of calculation is small, so the detection speed is fast. But for slow-moving object detection, there will be incomplete detection results. The detection effect is not good when the object overlaps in different frame times, so it is generally used in combination with other methods. The background difference method needs to calculate the background image first. When detecting the moving target, it is necessary to make the difference between the image at a certain moment and the background image. If the difference is greater than the set threshold, it can be judged as a moving object. The disadvantage is that it is often

limited by many background factors, and every time the difference is made, the background information at the current moment must be calculated, which brings a very large amount of calculation and its anti-interference ability is also poor.

The target detection network based on deep learning has achieved very good research results in recent years, attracting many scholars at home and abroad to study and study. Convolutional neural network [5] has powerful computing functions. Each convolutional layer extracts and integrates the features of the input image. As the network deepens, the feature information of different objects will be displayed very effectively. The network can be based on these features. Classify objects directly. After training with a large amount of training data, the convolutional neural network can complete the detection and classification tasks of different objects very efficiently. Target detection based on deep learning is currently mainly divided into single-stage network and two-stage network [6]. Early two-stage networks mainly include R-CNN [7], Faster R-CNN [8], etc. The main idea is to divide the detection task into two stages. First, a number of candidate frame regions are generated on the original image to represent There may be areas where the target object exists, and then the objects in these candidate frames are regressed and classified, and the position of the candidate frame is adjusted. The advantage of the two-stage detection network is that the detection accuracy is very high. However, the structure of the two-stage network increases the complexity and the amount of calculation, which leads to a lower detection speed to a certain extent.

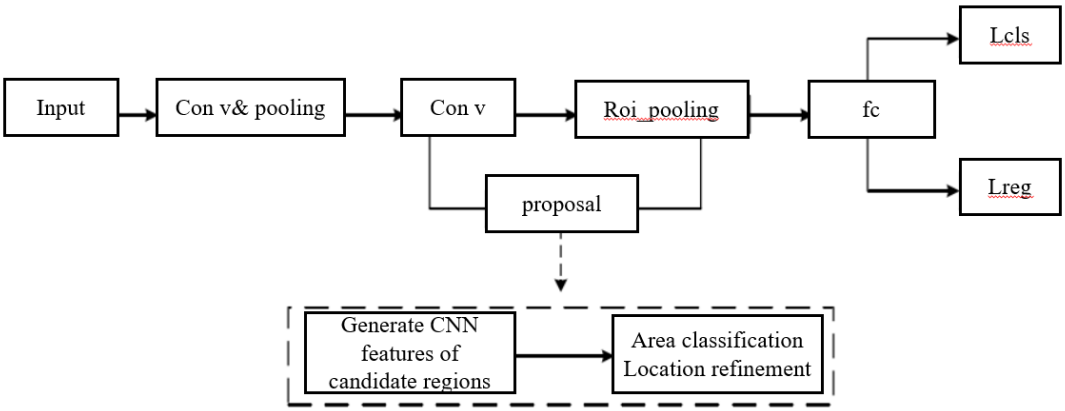


Fig.1. Two-stage target detection network structure diagram

The single-stage network discards the operation of selecting the candidate frame area on the basis of the two-stage, but directly sets a fixed number of candidate frames in each small grid, and then directly performs the regression detection task on the objects in the frame. The operation speeds up the network detection speed, but its accuracy will drop a lot accordingly. Typical networks include SDD [9], YOLO [10] series, etc.

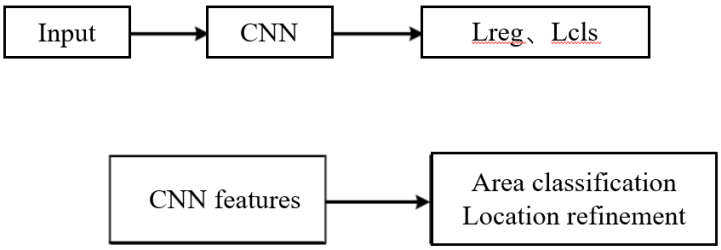


Fig.2. Single-stage target detection network structure diagram

2.2 Research status of trajectory prediction

The traditional trajectory prediction method mainly estimates the motion state of an object based on its historical position, and then comprehensively predicts the future motion trajectory based on the estimated value and the current motion state. It is mainly divided into two categories; the first category requires the establishment of a motion model of the object, and the trajectory prediction of the object according to the method of Kalman rate wave [11]. Kalman filter assumes two variables: position and velocity, and they are all randomly obeying Gaussian distribution. Each variable has a mean to represent the center of the random distribution, and variance to represent uncertainty. The trajectory prediction method based on Kalman filter is a recursive estimation method. In the prediction stage, the current motion state is estimated through the obtained observation value at the last moment, and then the current motion state is used to predict the future object motion traces of. The

accuracy of this method largely depends on the accurate establishment of the model. The prediction accuracy is easily interfered by the accumulation of errors and the model is greatly reduced. In the estimation process, the errors will also superimpose and lead to the decrease of the prediction accuracy. The second type of method also needs to establish a motion model [12], but uses a curve equation to fit the object's motion trajectory when predicting, and an observer is used to modify and optimize the predicted value. The first method is mostly used to predict the trajectory of table tennis. The force analysis of the table tennis is modeled. Considering that the rotation of the table tennis is affected by Magnusri, Andersson uses the force analysis to use the second method. The polynomial is fitted to the trajectory to obtain the movement state of table tennis [13]. Based on this, Nonomura team established a flying motion model of spinning table tennis [14]. Later, some scholars proposed a similar motion model and combined the Kalman filter method to estimate the motion state and predict the trajectory, and obtained high accuracy. This year, machine learning-based trajectory prediction methods have also been continuously proposed. For example, Walker J et al. proposed trajectory prediction based on optical flow method [15], which converts video sequences into discrete motion sequences, which can well recognize table tennis balls. And trajectory prediction, but it can only predict a short period of time in the future, and cannot achieve the effect of long-term prediction.

Table tennis has the characteristics of fast speed, small size, small mass and high-speed rotation. This brings high demands on the real-time and accuracy of the table tennis robot vision system. Traditional table tennis target detection methods mainly use color segmentation, contour search, and multi-sensor-based target detection methods. These methods require a small amount of calculation, so they can reach a high detection speed, but the detection accuracy is very susceptible to light intensity and background. The influence of the surrounding environment such as interference has brought about the problem of low detection accuracy. For example, the color segmentation algorithm mainly sets the detection threshold according to the color of the table tennis. In the trajectory prediction task, the main method is based on the establishment of a physical model; The rotating table tennis model has the characteristics of high-order nonlinearity. However, the traditional model uses an approximate linear model to approximate the actual model, which often has large deviations and this error will accumulate with iterations, resulting in prediction accuracy Very low and usually only suitable for a certain type of spinning ball.

Under the current background of the development of AI technology, the combination of artificial intelligence and robots is the general trend, and it also brings broader development space and research value to robotics research. At the same time, it also brings more research possibilities to table tennis robots. For the problems of the above-mentioned traditional methods, deep learning methods can be well improved, but the detection network of deep learning has a large amount of calculation and requires a large amount of training data, which brings some new problems. Therefore, in view of the shortcomings of traditional methods in the task of table tennis target detection and trajectory prediction, this article tries to combine the deep learning method with stronger generalization ability and anti-interference ability with the table tennis visual system, and at the same time predict the trajectory. Taking the rotation characteristics of table tennis into account, researches on the target detection of table tennis based on deep learning and the prediction of the trajectory of the rotating ball are carried out.

3 TABLE TENNIS DATA COLLECTION AND PROCESSING

In the field of machine learning, the reason why neural networks have very powerful capabilities is largely dependent on the training set data it uses. The data set used by machine learning has the true label and category of the target. These large data sets with labeled data are sent to the neural network for training in batches, so that the neural network can learn very rich information that represents the characteristics of the target. After the data is trained and learned, the network can quickly and accurately identify the target. So as to complete tasks such as classification or regression. At the same time, we also need to prepare an appropriate test set to test the network capabilities of our trained network.

In the task of detecting objects in table tennis, a large amount of training data is also needed to train the network. Each picture must be manually labeled, labeled, and calibrated with category and table tennis position information. In the task of trajectory prediction, trajectory data of table tennis is needed. It takes a lot of time and energy to produce these data, and there is currently no publicly available large-scale table tennis data set. Therefore, in order to complete the research of this article, a large number of table tennis picture data under different environments, different colors and different lighting conditions were manually collected, and the data set required for network training was further constructed.

3.1 Table Tennis Vision System

The visual perception system of table tennis is equivalent to the human eye. The task to be completed is mainly to detect the table tennis in real time, and then predict the trajectory in the future based on the detected trajectory. The vision system used in this article is composed of high-speed binocular cameras. The selected industrial camera model is WP-UT230. The sampling

frequency during operation can reach 150FPS, which can meet the real-time requirements of the table tennis vision system. The physical map is shown in Figure 3.

Fig.3. Industrial binocular camera

The camera parameters are shown in Table 1.

Table 1 Camera parameters

Phase element size	$4.8 \times 4.8\mu m$
Resolution	1920×1200
Frame rate	$150(FPS)$
Exposure time	$16\mu s - 1s$
Bit depth	8-bit/10-bit
Image format	BayerRG8, BayerGB8 Mono8/10Packed
interface	Data transmission: USB3.0 interface / lens interface: standard C port

The picture taken by a single camera is a two-dimensional image, and the ping pong ball needs to obtain a three-dimensional table tennis coordinate for trajectory prediction. At this time, the obtained two-dimensional coordinate needs to be converted to obtain the corresponding three-dimensional coordinate. Two cameras need to be calibrated to obtain camera parameters. The calibration of the camera parameters adopts the classic Zhang-type calibration method, and the camera parameters after calibration are shown in Table 2 and Table 3.

Table 2 Left camera parameters

Camera parameters	Numerical value
focal length	[1771.52282 1781.72145]
Main point	[991.26516 599.12267]
Tilt parameter	0
Distortion parameter	[-0.06660 0.12069 0.00057 0.00757 0.00000]

Table 3 Right camera parameters

Camera parameters	Numerical value
focal length	[1775.74740 1779.20301]
Main point	[968.25786 590.47443]
Tilt parameter	0
Distortion parameter	[-0.1475 0.18440 0.00110 -0.00259 0.00000]

The main purpose of the camera parameter calibration is to obtain the conversion relationship between the world coordinate system, the camera coordinate system, and the image coordinate system, so as to convert the detected two-dimensional image coordinates into the three-dimensional coordinates of the object space. The corresponding relationship between the three is shown in Figure 4.

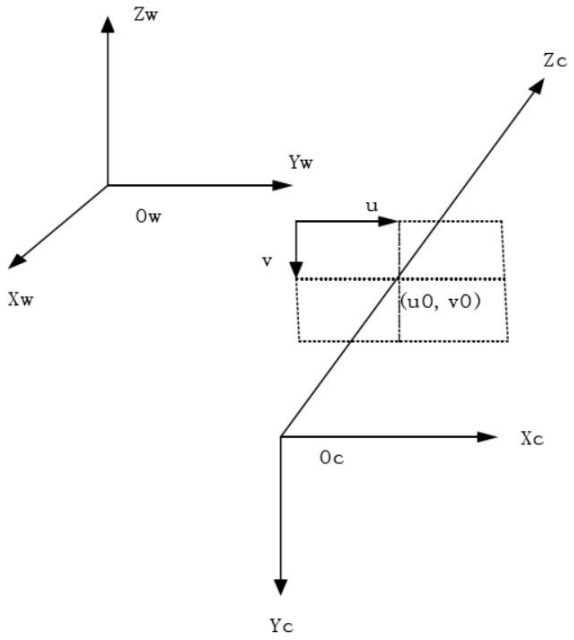


Fig.4. Schematic diagram of the correspondence between the three coordinate systems

The transformation of objects in different coordinate systems is obtained through matrix transformation. Equation (1) is the coordinate transformation relationship between the world coordinate system and the camera coordinate system.

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = [r_1\ r_2\ r_3\ t] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \quad (1)$$

In formula (1), 1 represents the coordinates of the object in the camera coordinate system, and 2 represents the three-dimensional position of the object in the world coordinate system. The conversion formula from the former to the latter is shown in formula (2).

$$z_c \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \quad (2)$$

$[x_p\ y_p]$ is the two-dimensional coordinate position of the object in the photo taken by the camera, and f is the focal length value of the camera.

The coordinates of a two-dimensional image are calculated from the position $[0 \quad 0]$ of the upper left corner. The coordinates of the lower left corner $[1200 \quad 1920]$of the image collected in the vision system of this article are relative to the origin of the camera coordinate system, that is, the pixel coordinates of the principal point. Each two-dimensional the conversion formula of the image in the image coordinates to its corresponding pixel point is shown in formula (3). The d_x, d_y in the formula represents the size on the x -axis and y -axis where the coordinate point is located in the image coordinate system. And the coordinate value of the main point pixel.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} (3)$$

The conversion between the pixel coordinates in the image coordinate system and the left side in the world coordinate system can be obtained by the formula. When setting the position of the chessboard in the world coordinate system, z_w is 0, then the following coordinate transformation formula can be obtained (4).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = s \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} [r_1 \quad r_2 \quad t] \begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} \quad (4)$$

u and v represent the coordinates of the pixel, s is the scale factor, and γ, u_o and v_o are the parameters in the camera. $[r_1 \ r_2 \ t]$ represents the transformation matrix from the camera coordinate system to the world coordinate system.

The homography transformation refers to the relationship between the two-dimensional coordinates of an object in the image and the corresponding three-dimensional position in the world coordinate system, and its transformation matrix is shown in formula (5).

$$H = s \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} [r_1 \ r_2 \ t] \quad (5)$$

It is not difficult to see from the above formula that as long as the internal and external parameters of the camera are obtained, the corresponding homography transformation matrix can be obtained, and the three-dimensional position coordinates of a certain point in the image in the world coordinate system can be calculated. In the process of acquiring images, the distortion problem caused by the camera also needs to be solved. Equation (6) is the radial distortion formula, and (7) is the tangential distortion formula. To eliminate the distortion problem in the image, the final target point coordinates can be obtained by formula (8).

Where $r$is the distance between the coordinates of the target point and the imaging center, and $(k_1, k_2, k_3, k_4, k_5)$ is the distortion parameter in Table 2.1 and Table 2.2, respectively.

$$\begin{cases} x_r = (k_1 r^2 + k_2 r^2 + k_5 r^2) x_p \\ y_r = (k_1 r^2 + k_2 r^2 + k_5 r^2) y_p \end{cases} (6)$$

$$\begin{cases} x_t = 2k_3 x_p y_p + k_4 (r^2 + 2{x_p}^2) \\ y_t = 2k_4 x_p y_p + k_3 (r^2 + 2{y_p}^2) \end{cases} (7)$$

$$\begin{cases} x = x_p + x_r + x_t \\ y = y_p + y_r + y_t \end{cases} (8)$$

3.2 Table tennis data collection and processing

Based on the above table tennis vision system platform, this article collects the image data set needed for the research, and processes the collected data according to different research purposes and methods. In addition, for the data set used for deep learning network learning, if there is no label information, then the network cannot match the learned features with the corresponding categories, and it cannot meet the requirements of recognition and classification capabilities. Therefore, the annotation of the data set has played a vital role in the early stage of the work. At the same time, in order to enhance the efficiency and speed of model learning, this article also needs to perform some image processing operations on the labeled tags to increase the richness of the data, so that the model can learn the characteristic information of table tennis more efficiently.

3.3 Data collection and annotation

For an efficient neural network, in its training process, the greater the richness of the data in the training set, the more information it learns, and the stronger the generalization ability of the network. Networks for different detection purposes require different data set characteristics. For the detection task of table tennis in this article, it has the following characteristics: ①The target is small, and the proportion of pixels in the entire picture is low. ②The table tennis balls have different colors, mainly yellow, white, yellow and white. Therefore, it is necessary to collect table tennis data of different colors and provide them for network training; ③Background interference, mainly because there may be contours or colors similar to table tennis in the background object interference and too dark background light intensity will also interfere with the detection of table tennis balls. After the above analysis, this paper collects table tennis data sets of different colors under different lighting conditions, and annotates the data on this basis. The current table tennis competition mainly uses yellow and white table tennis, and gradually began to use yellow and white table tennis, the environment in the process of table tennis is also changing with the venue. In different scenes, the light intensity, other decorations in the scene, and the colors of the clothes of the contestants or spectators are different. In view of the above factors that will affect the detection of table tennis, this article collects image data in different environmental scenarios as shown in the figure below figure 5. At the same time, many distractors were artificially set up, such as tables, clothes and other background elements. Red and black racquets are also used for table tennis rackets. For these collected data sets, the manual labeling method is used to label, as shown in Figure 6.

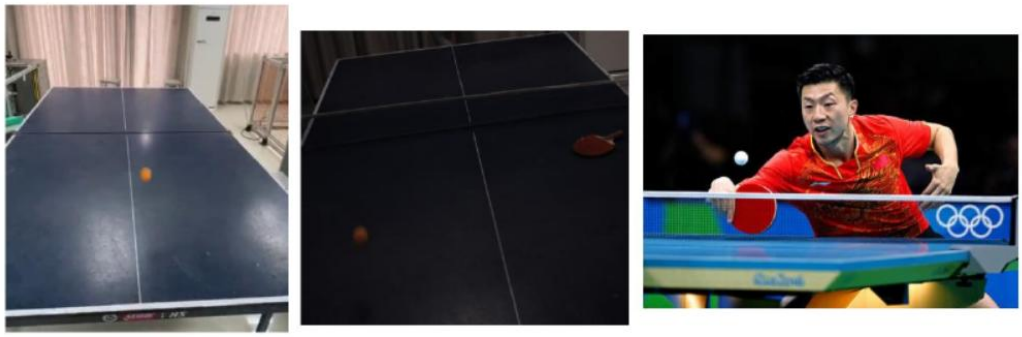


Fig.5. Part of the original data set

Fig.6. Part of the label data set

In order to increase the richness and diversity of the data set, this article also carried out some preprocessing operations on the collected data set. First, perform horizontal and vertical flip operations on part of the data set. Swap the left and right parts of the image and the top and bottom parts. Flip horizontally, as shown in formula (9).

$$[x_r \ y_1 \ 1] = [x_0 \ y_0 \ 1]\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ W & 0 & 1 \end{bmatrix} = [W \ -x_0 \ y_0 \ 1] \quad (9)$$

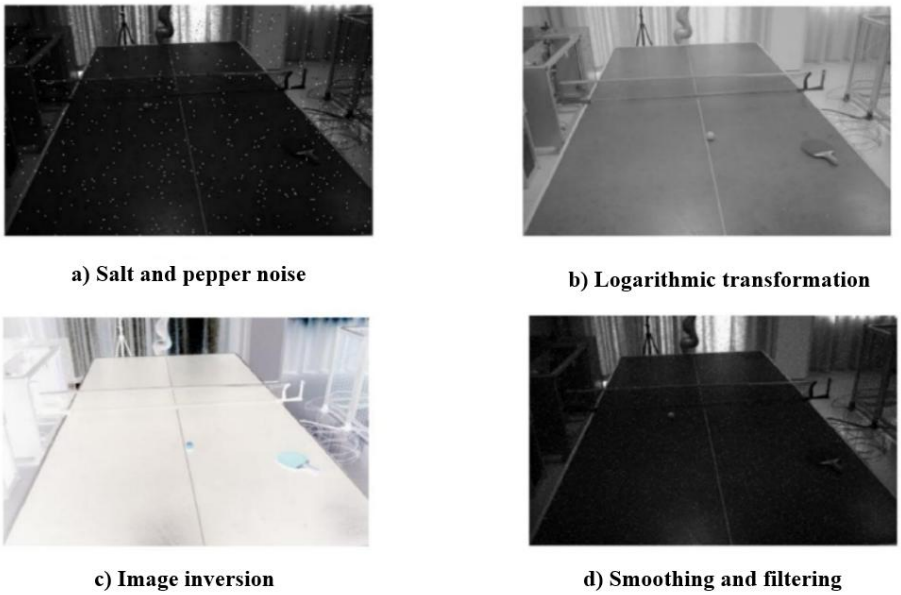
Flip vertically, as shown in formula (10).

$$[x_r \ y_1 \ 1] = [x_0 \ y_0 \ 1]\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & H & 1 \end{bmatrix} = [x_0 \ H \ -y_0 \ 1] \quad (10)$$

In this paper, 60% of the data in the data set are randomly subjected to the above two transformations, and 30% are randomly selected for rotation changes. The rotation angles are clockwise 90° and 270° and counterclockwise 90° and 270°, formula (3-11) θ is the angle of rotation:

$$[x_r \ y_1 \ 1] = [x_0 \ y_0 \ 1]\begin{bmatrix} cos\,\theta & sin\,\theta & 0 \\ -sin\,\theta & cos\,\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} (11)$$

In order to enable the network to learn the characteristic information of table tennis under different lighting intensities in different scenarios, this paper also performs some image processing algorithms on some data sets. Figure 7 (a) Adding salt and pepper noise is to randomly appear some white and black spots in the image. For small targets such as table tennis, adding salt and pepper noise can cause some interference. In the process of network learning, these artificial interferences It can increase the anti-interference ability of the network to a certain extent. Figure 7(b) is a logarithmic transformation of the original image. The overall brightness of the image after logarithmic transformation is greatly increased compared to the original image, and the image becomes smoother, and its color characteristics are also changed, which is different from the brightness of the surrounding environment. The value is small. Figure 7(c) is the reverse transformation of the image, the white and gray details in the transformed image are enhanced. Figure 7(d) uses a 3×3 smoothing filter to process the image. After low-frequency enhanced spatial filtering, the overall brightness of the image decreases and becomes blurred.



a) Salt and pepper noise

b) Logarithmic transformation

c) Image inversion

d) Smoothing and filtering

Fig.7. Data enhancement

In the network training based on deep learning, a large number of training data sets are needed to train the network, but there are currently no public data sets available. This article collects different backgrounds and interferences under the binocular vision system of the table tennis robot. Table tennis data set of balls of different colors. On this basis, image processing operations are performed on the pictures, using image processing such as inversion, noise addition, logarithmic transformation, filtering, etc., which further increase the richness of the training set on the basis of the existing data set.

4 TABLE TENNIS TARGET DETECTION BASED ON FEATURE FUSION NETWORK

As the primary task of the table tennis robot system, the vision system needs to perform accurate target recognition and position positioning of the table tennis ball while ensuring a faster recognition speed. Early research was mainly based on the color and contour characteristics of table tennis balls. This type of method has great advantages in detection speed, but the shortcomings are also very obvious and difficult to overcome, such as being very susceptible to light, color, etc. Interference from external factors. When the light intensity in the environment is very high or the surrounding color is similar to the color of the ping pong ball, and the color of the ping pong ball changes, the recognition accuracy of this method will be greatly reduced.

In recent years, deep learning has continuously achieved breakthrough results, and has also continued to innovate and develop in the fields of image recognition and target detection. Compared with the traditional target detection algorithm, the target

recognition method based on convolutional neural network is more robust, has stronger anti-interference ability, and adapts to detection tasks in a variety of environments. However, a large number of parameters bring about the problem of low detection speed, and The deepening of the network depth is not ideal for the detection of small targets. Based on the convolutional neural network, this paper balances the detection speed and accuracy, and builds a network suitable for the detection of table tennis targets; in the design of the convolutional network, this paper designs many small target detection techniques to make the network deeper It still has strong detection capabilities for small targets.

4.1 Data enhancement

The learning process of the neural network needs to learn a large amount 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 encountering simulations during training. The situation. For example, by cropping multiple images, and then stitching the cropped images, the model reduces the sensitivity to the object position during the process of learning these images, so that in the process of network recognition, such The position information of the object will have less influence on the recognition result. We can also adjust the brightness, saturation, hue and contrast of the object to reduce the model's dependence on color. In recent years, in the development of deep learning, different data enhancement methods have been proposed and have achieved very good experimental results.

Data enhancement can be divided into two categories: offline enhancement and online enhancement. When the data set is small, in order to save the time of acquiring new data, an offline enhancement method can be used. Online enhancement is usually used for larger data sets. It is performed after each batch of data is acquired during network training, and the data is changed such as rotation, translation, flipping, and turning. Many machine learning models have gradually been able to support this enhancement method, and GPUs can be used for optimized calculations.

In recent years, mainstream data enhancement technologies include Mixup, Coutout, Cutmix, Mosaic enhancement and so on. Mixup is to mix two random pictures in a certain ratio, and when sorting, they will also be allocated according to the proportion of each picture. The realization method is shown in formula (12) and (13).

$$\tilde{x} = \lambda x_i + (1 - \lambda x_j) \quad (12)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda y_j) \quad (13)$$

Among them, (x_i, y_i) and (x_j, y_j) are sample data randomly drawn 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 the amount of calculation. It is one of the good data enhancement methods.

Fig.8. Mixup data enhancement

The starting point of Cutout is similar to the method of random erasure. The idea is to randomly select a fixed-size positive area and fill it with 0. At the same time, the data will be normalized to avoid the impact of filling 0 on training. Cutmix's processing method is also relatively simple. It also operates on two pictures. The idea is to randomly generate a crop frame, crop the corresponding position of A, and then use the ROI of the corresponding position of the B picture to place it in the cropped position of a new sample is formed, and finally the loss is calculated by using a weighted summation method.

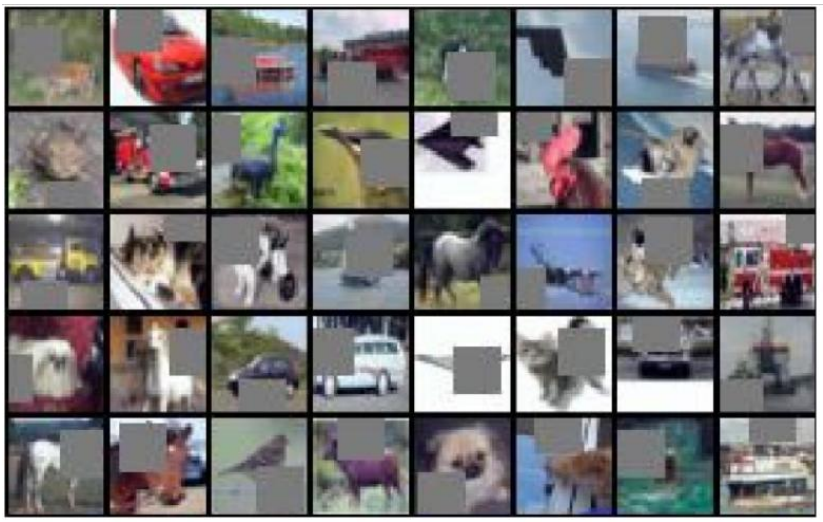


Fig.9. Cutout data enhancement

The Mosaic data enhancement method is proposed in YOLOv4. The idea is to flip, zoom, and change the color gamut of the four input images, and then stitch them into one image and input it into the network for training. Its advantage is that it enriches the background of the detected object, and the stitched pictures are equivalent to processing the data information of four images at a time during network calculation, which effectively increases the ability and efficiency of network learning.

Fig.10. Mosaic data enhancement

Fig.11. This paper uses data enhancement methods

Considering that this article only needs to detect the table tennis target, the table tennis occupies fewer pixels in a picture, which causes a waste of training costs in network training, and after convolution, the feature map is slightly smaller than the original image. Ten times, the network's ability to learn small objects such as table tennis is greatly reduced. This article refers to Mosaic enhancement. During network training, the table tennis in the picture is copied, and the table tennis in the picture is passed through without affecting other objects. After three copies, for each target in the network initialization, there will be three a priori boxes to detect, then in each image after copying, there will be 12 a priori boxes to jointly learn the characteristic information of the table tennis target, which is effective Strengthen the richness of the data set to prevent the network from over-fitting.

4.2 Feature Fusion Network

The FPN network consists of two lines, a bottom-up line and a top-down line, which are connected horizontally. The bottom-up process is the forward propagation process of the network, and the feature map will become smaller and smaller after the convolution kernel is calculated. The top-down process is to up-sample the feature maps, and the horizontal link is to use the 1×1 convolution kernel to fuse the feature maps of the same size generated by the two. In this way, the bottom-level positioning detailed information can be used to make the network learn the target The feature can also learn more accurate location information, especially for the detection of small objects, which brings a very significant improvement.

This paper improves on the basis of FPN, adding a bottom-up connection after FPN network, so that four effective feature layers (N1-N5) can be reached. Convolutional neural networks usually use the last or a certain feature layer of the network for target detection and localization, but this only uses the information of one of the feature layers, and the information of each feature layer has its own for the target detection task Specific functions, so in the end, through the adaptive pooling layer in C, the feature layer in B is pooled into feature layers of the same size, and max is used to fuse them together to obtain the final feature layer for target detection. Such a design can fuse the feature information of each feature map, so that the network's ability to detect the position of small objects is greatly improved.

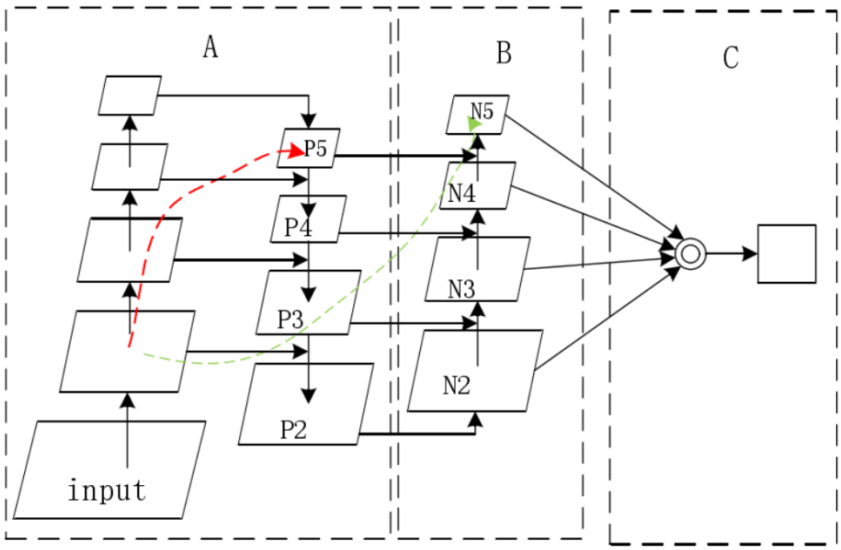


Fig.12. Feature Fusion Network Structure Diagram

The table tennis vision system needs to quickly and accurately identify and locate table tennis targets in a certain interference environment to meet the task requirements of the robot system for real-time sparring. Traditional target detection methods such as color segmentation and contour search algorithms are widely used and have high speed, but the problems are also obvious, that is, the anti-interference ability is very poor, and they are easily affected by environmental changes. And once the color of the ping pong ball is changed, the detection accuracy will be greatly reduced. The target detection algorithm based on deep learning has strong anti-interference ability, but the large amount of network calculation results in a low detection speed, and the detection effect for small targets is not ideal. In order to balance the speed and accuracy of target detection, this paper constructs a target detection network based on feature fusion network. For small targets such as table tennis, a feature fusion network is added to combine the upper layer with rich semantic information and the lower layer with rich object location information. The fusion of the characteristic layers of the network improves the detection accuracy of the network for small targets, and trains and tests in the data set constructed by itself, and continuously optimizes the network structure to balance the detection speed and accuracy of the network.

5 Conclusion

The main work of this paper is to study the target detection and trajectory prediction for the rotating table tennis ball. Constructed a table tennis target detection network based on feature fusion and trained and tested it on the table tennis data set made by ourselves; constructed a table tennis trajectory prediction network based on LSTM, and distinguished the type of table tennis rotation based on human experience to realize the prediction of the trajectory of table tennis under a variety of rotating bodies. The main contributions of this article are as follows:

(1) Based on the seven-degree-of-freedom KUKA robot system, binocular vision is used to collect table tennis pictures in a variety of environments, and the table tennis targets in each picture are marked. Image processing operations are performed on part of the pictures in the data set, which increases the richness of the data set. A data set containing 8000 images of table tennis was constructed. Laid a foundation for the learning of neural networks.

(2) In order to solve the problem of low detection accuracy caused by the interference of environment and light in the task of table tennis target detection by traditional methods, this paper proposes a target detection method based on feature fusion network. In the feature extraction network, the structure of the cross-layer link network (CSPNet) is used to optimize the

gradient in the network learning process and strengthen the network learning ability. In order to solve the problem that the detection ability of small targets decreases with the increase of network depth, a feature fusion network is constructed, which combines the feature layer with rich semantic information in the upper layer and the feature map with rich object position information in the lower layer, so that the increase in network depth can still be achieved Learn the location information of the object. Improved the ability of the network to detect small targets. The network in this paper can adapt to factors such as different illumination, environmental interference, etc., the detection accuracy is higher than the traditional target detection algorithm, and the detection speed meets the requirements of the vision system of the table tennis robot.

(3) Aiming at the problem of low prediction accuracy of traditional trajectory prediction methods based on table tennis physical models and motion state analysis, this paper constructs a table tennis trajectory prediction network based on LSTM; using LSTM's ability to predict sequences, stack them continuously to achieve In the case of receiving a piece of table tennis trajectory position information, the long-term prediction of the table tennis trajectory; in addition, the general rotation type of the ping pong ball is reversed through the flight trajectory of the ping pong ball, and the rotation type of the table tennis under different rotation states is proposed. The experiment shows that the prediction points are concentrated in the sweet spot of the table tennis racket, the prediction accuracy is significantly improved and the prediction speed meets the visual system System requirements.

References

[1] Kalpesh Prakash Modi. Vision application of human robot interaction[microform]: development of a ping pong playing robotic rm[M].2005.

[2] Russel L.Andersson. A Robot Ping-Pong Player: Experiments in Real Time Control[M]. Cambridge MA: The MIT Press,1987.

[3] Katharina Muelling, Jens Kober, Jan Peters. A biomimetic approach to robot table tennis[J]. Adaptive Beehavior, 2010. 19(5):359-376.

[4] Jens Kober, Katharina Muelling, Oliver Kromer, Christoph H Lampert, Bernhard Scholkopf, Jan Peters. Movoment templates for learning of hitting and batting[C]//Robotics and Automation(ICRA), 2010 IEEE International Conference on Simulation of Adaptive Behavior: From Animals To Animats. 2010:273-282.

[5] Jan Peters, Katharina Muelling, Jens Kober, Duy Nguyentuong, Oliver Kromer. Towards motor skill learning for robotics[J]. Springer Tracts in Advanced Robotics, 2011. 70:469-482.

[6] Katharina Muelling, Jens Kober, Oliver Kroemer, Jan Peters. Learning to select and generalize striking movements in robot table tennis[J]. International Journal of Robotics Research, 2013. 32(3):263-279.

[7] Hartley J.Toshiba progress towards sensory control in real time [J]. Indust. Robot,1987,14(1):50-52.

[8] Jens Kober, Katharina Muelling, Jochen Peters. Learning throwing and catching skills[C]. In Intelligent Robots and Systems(IROS), 2012 IEEE/RSJ International Conference on IEEE, 2012,5176-5168.

[9] Katharina Muelling, Jens Kober, Jan Peters. Learning table tennis with a mixture of motor primitives[C]//Humanoid robots(Humanoids), 2010 10TH IEEE-RAS International Conference on IEEE, 2010:411-416.

[10] Akira Nakashima, Yuki Ogawa, Yosuke Kobayashi, Yoshikazu Hayakawa. Modeling of rebound phenomenon of a rigid ball with friction an.d elastic effects[C]. In American Control Conference(ACC), 2010. IEEE, 2010,1410-1415.

[11] Christoph H Lampert, Jan Peters. Real-time detection of colored objects in multiple camera streams with off-the-shelf hardware components[J]. Journal of Real-Time Image Processing, 2012, 7(1):31-41.

[12] Jan Peters, Katharina Mulling, Jens Kober. Experiments with motor primitives in table tennis[C]. In Experimental Robotics. Springer, 2014, 347-359.

[13] Akira Nakashima, Yuki Ogawa, Chunfang Liu, Yoshikazu Hayakawa. Robotic table tennis based on physical models of aerodynamics and rebounds[C]. In Robotics and Biomimetics(ROBIO), 2011 IEEE International Conference on. IEEE, 2011, 2348-2354.

[14] Akira Nakashima, Junko Nonomura, Chunfang Liu, Yoshikazu Hayakawa. Hitting back-spin balls by robotic table tennis system based on physical models of ball motion[C]. In Robot Control, 10th IFAC Symposium on. 2012, 834-841.

[15] Shuzhi Sam Ge, Frank L.Lewis. Autonomous mobile robots: sensing, control, decision making and applications[M]. volume 22. Boca Raton, FL, USA: CRC Press Taylor&Francis Group, 2006.

[16] Michiya Matsushima, Takaaki Hashimoto, Mashiro Takeuchi, Fumio Miyazaki. A learning approach to robotic table tennis[J]. IEEE Transactions on Robotics, 2005. 24(4):767-771.