Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5402410, 10 pages
https://doi.org/10.1155/2022/5402410

Research Article

Object Extraction of Tennis Video Based on Deep Learning

Huadong Huang

Institute of Physical Education and Health, Yulin Normal University, Yulin, 537000 Guangxi, China

Correspondence should be addressed to Huadong Huang; huanghuadong@ylu.edu.cn

Received 23 January 2022; Revised 2 March 2022; Accepted 5 March 2022; Published 20 March 2022

Academic Editor: Kalidoss Rajakani

Moving target detection and tracking technology is the core technology in the field of computer vision. It integrates image processing, pattern recognition and intelligence, and artificial intelligence and automatic control are the keys to an intelligent video surveillance system. The method acquires video image signals through visible light or infrared sensors, performs digital image processing on the video images, detects moving targets, and then extracts moving targets for target recognition. Then, the moving target is predicted and tracked according to the image features and spatiotemporal features, and the contour shape, position, and motion trajectory of the target are obtained, which provides data support for subsequent tasks. This paper uses the convolutional neural network model HyperNet as the technical support to study the deep learning (DL) tennis video target extraction. The final experimental results show that the loss value of the loss function in the training process of this method is stably maintained between 1.5% and 2.5%, and the speed performance is also greatly improved. The number of boxes for extracting candidate regions is significantly reduced, the calculation time for each frame will not exceed 1.8 s, the orientation accuracy of target extraction is 96.32%, and the size accuracy is 91.05%.

1. Introduction

In the era of big data, the number and scale of videos are increasing, which poses challenges to video object detection. Improving the efficiency and speed of video object detection is of great significance to object detection and recognition. Especially in the field of images, computer vision technology based on artificial intelligence has made great progress and has gradually become one of the key technologies in the field of image and video processing. The video target area extraction is used to determine the target position in each frame of the video image and generate the moving target trajectory. However, data mining and detection of small objects are very rare, which is a difficult problem in computer vision. At present, there is still a lot of research space for this research content. Although DL has been involved in many fields, there is little research on small object detection. However, with the development of science and technology, some imaging instruments can capture the rare features of some small targets, which lay a good foundation for in-depth study of small targets.

The concept of DL emerged in the 1980s. It is generally believed that it is developed from artificial neural networks on the basis of various complex multilayer neural networks. By optimizing algorithms, it can automatically learn and extract nonlinear features. In recent years, DL based on computer vision has made great progress. The Alex-net algorithm based on deep convolutional neural network has achieved good results in image classification, which is 11% higher than the previous optimization algorithm, providing a new perspective for DL. The optimization of various computer vision models based on DL, the improvement of various public data sets, and the improvement of computer hardware performance, DL has also ushered in a golden age of development and entered the public eye.

The innovation of the research on target area extraction and detection tracking in this paper is first, it summarizes the classic image target extraction algorithm and analyzes the characteristics. The second is a brief overview of several typical image target extraction models based on convolutional neural networks, and their performance is analyzed. The third is to improve the HyperNet video target extraction

model and achieve good results, and it has a good performance in target extraction of various data sets.

2. Related Work

With the establishment of massive data, the development of computer hardware, and the breakthrough of in-depth learning technology, the development of computer vision is changing with each passing day, and all of these have accelerated the development and application of new computer vision technologies. Many scholars have made useful attempts and made breakthroughs. Li first analyzed the positioning principle of robot binocular vision and then used the accelerated robust feature (SURF) method to extract target features. He adopted the BP neural network (BPNN) method for localization and verified the method through experiments [1]. Although the experimental results verify the effectiveness of the machine learning method, the actual application error is still unknown. Tang and Huo optimize the live broadcast synchronization of tennis professional league based on wireless network planning. The scheme first establishes a three-dimensional detection target model, extracts the background from the moving video image, uses the interframe difference elimination algorithm to extract, and predicts the target motion trajectory. Then, the scheme uses the Hilbert transform to analyze the phase difference characteristics of the foreground trajectory of the image and detects the missing points of the target [2]. To achieve higher processing efficiency and accuracy, Fan applies DL to object class detection. In the background extraction step, an improved meaning-based background extraction algorithm and a region-of-interest extraction algorithm that reduce image pixels are proposed. The test results show that the algorithm has good video object detection performance [3]. Pang et al. incorporate SE block and temporal attention mechanism (TAM) in the framework of Siamese neural network [4]. Li et al. proposed a domain-adaptive diagnostic model based on improved deep neural networks and raw vibration signal transfer learning [5]. In order to better solve the problem of low tracking accuracy caused by target scale mutation, Yang designed and proposed an adaptive scale mutation tracking algorithm based on DL network. The target is detected first, then the kernel correlation filtering method is used to track it, and the validity of the model is verified by experiments [6].

3. Tennis Video Target Extraction Based on DL

3.1. Image Object Extraction Algorithm

(1) Frame difference method

The interframe difference method is the simplest and most intuitive moving target detection method. The basic idea is the moving target will cause huge changes in image pixels due to motion, and the pixel positions of these important changes are detected. The frame difference method is the fastest and most effective method in moving object detection. However, the environmental requirements are also the most demanding. You need a fixed camera, otherwise, the image will shake and the background will be detected by the objects in front [7, 8]. Therefore, it will be used in conjunction with morphological processing methods.

First using the difference between the current frame and the previous frame and obtaining the absolute value of the result:

$$C_t = |P_t - P_{t-1}|. \tag{1}$$

P_t represents the current frame of the video image, and C_t is the difference result of the previous frame image and the current frame image, which is thresholded:

$$c_t = \begin{cases} 0, C_t \le K, \\ 1, C_t > K. \end{cases} \tag{2}$$

K is an adaptive threshold, which is obtained from the segmentation map and will change dynamically. c_t is the final detection map. After morphological processing to remove outliers and fill in the target, the result is shown in Figure 1. However, the frame difference method also has its shortcomings. When the video background shakes, the background may also be judged as the foreground target [9].

(2) Image segmentation algorithm

Dividing the video input image into multiple small areas, and the input image containing the target to be detected is

$$I = (S, A), (s_i, s_j) \in A. \tag{3}$$

S is a set of pixel points, and the connection between two adjacent pixels is an edge. The larger the difference between the pixels, the lower the similarity between the pixels.

I is represented as a minimum spanning tree, and B is obtained after segmentation. The connection of each vertex in B is called an edge, and the difference between the pixels connected by the largest edge is used as the internal difference of the segmentation area:

$$\mathrm{In}(B) = \max_{a \in \mathrm{MST}(B,A)} \omega(a). \tag{4}$$

In the formula, $\omega(a)$ represents the difference between two pixels connected by edge a, which is called the weight in this paper.

Using the minimum weight of the connection between the two segmentation regions to measure the difference between the segmentation regions:

$$\mathrm{Dis}(B_1, B_2) = \min_{s_i \in B_1, s_j \in B_2 (s_i, s_j) \in A} \omega(s_i, s_j). \tag{5}$$

$\omega(s_i, s_j)$ is the minimum weight between the segmented regions s_i and s_j. When there is no connecting edge between

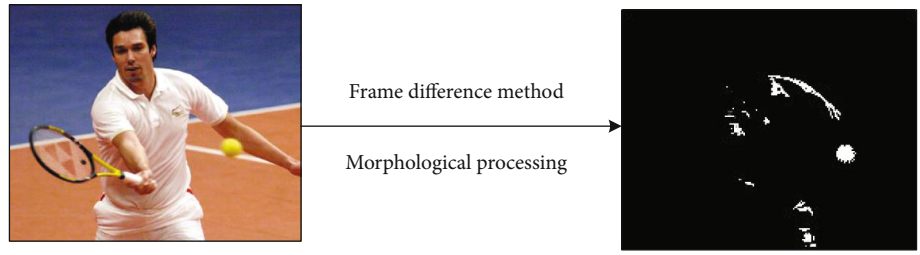

Figure 1: Detection results of frame difference method.

the two partitioned regions, the difference between the two partitioned regions is considered to be infinite:

$$\mathrm{Dis}(B_1, B_2) = \infty. \tag{6}$$

Then judging whether there is a clear boundary between the two divided regions.

$$Q(B_1, B_2) = \begin{cases} \text{true, if } \mathrm{Dis}(B_1, B_2) > \mathrm{In}_{\min}(B_1, B_2), \\ \text{false,} \quad \text{else,} \end{cases} \tag{7}$$

$$\mathrm{In}_{\min}(B_1, B_2) = \mathrm{Min}\left[\mathrm{In}(B_1) + \tau(B_1), \mathrm{In}(B_2) + \tau(B_2)\right], \tag{8}$$

$$\tau(B) = \frac{t}{|B|}. \tag{9}$$

t is the parameter, τ is the threshold, and $\mathrm{In}_{\min}$ is the minimum segmentation internal difference. The segmentation steps are as follows: first, arranging all edges in the graph from small to large based on the edge weights. Second, each vertex in the graph is regarded as a partition region. Then building a model for each edge, determining whether the two vertices are in the same segmentation domain, merging if they belong, otherwise, skipping. Repeating these steps, traversing all edges, and finally getting a segmented image [10, 11].

The advantage of this algorithm is that in the image segmentation part, the predivided regions in advance meet the multiscale requirements of the target, and the oversegmented regions will be merged later by the region merging algorithm.

(3) Background difference algorithm

The premise of the interframe difference method is the invariance of the image background. If the background moves, this method is not suitable, and the background difference method can avoid this problem. The basic principle of the background difference method is to extract the static background from the video sequence and then use the difference between the current frame and the background to obtain the moving foreground. In static scenes, the background model can be preimaged without moving objects or noise. Different processing is performed on the current image frame and the background reference model, and the region of the moving object is determined by counting the change information in the histogram [12]. Therefore, the size, position, shape, and other related information of the moving object can be known. Its schematic diagram is shown in Figure 2. The background difference method can effectively extract moving targets, and the commonly used background modeling method can be Gaussian modeling. When the background reference model is given, the background difference method is an efficient moving object detection method.

The way to initialize the background model is generally to calculate multiple frames of images from the image sequence and then take the average. In formula (10), $B(\cdot)$ is the background pixel value at the (x, y) position at time t, and $P(\cdot)$ represents the image information of the k-th frame. The background is extracted from the previous N frames of graphics. After the background image is stored, the difference graphics between the current frame and the background image is used to extract the moving target.

$$B(x, y, t) = \frac{\sum_{k=t-N}^{t-1} P(x, y, k)}{N}. \tag{10}$$

Letting $G_k(x, y)$ be the current frame, $B_k(x, y)$ be the corresponding background frame, and $D_k(x, y)$ be the difference result, then

$$D_k(x, y) = |B_k(x, y) - G_k(x, y)|, \tag{11}$$

$$T_k = \begin{cases} 1, D_k(x, y) \geq T, \\ 0, D_k(x, y) < T. \end{cases} \tag{12}$$

When T_k is 0, it means the background, and when T_k is 1, it means the moving target area. T is the threshold value. When the difference image point value is greater than the set T value, we consider this point to be a point on the moving target, otherwise, it is considered to be an image background point.

An example of background differential detection results is shown in Figure 3.

(4) Region merging algorithm

After the regions are divided according to the features of the input images, they are generally merged according to the similarity features. The reference similarity algorithms

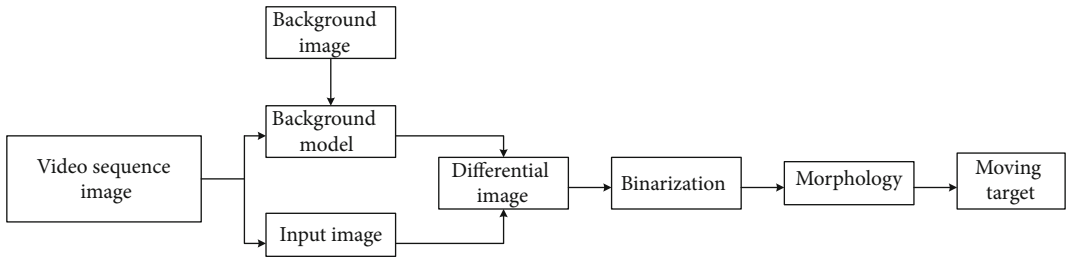

Figure 2: Principle of background difference method.

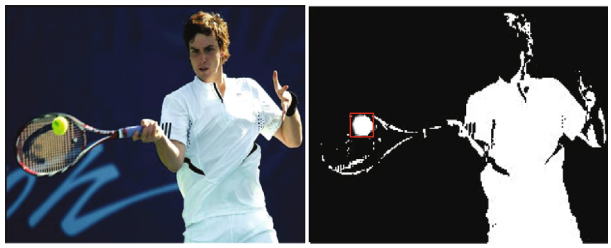

Figure 3: Detection results of background difference method.

include color similarity, shape similarity, size similarity, and texture similarity to measure the similarity between regions [13, 14].

$$S = m_1 * SC(t_i, t_j) + m_2 * SF(t_i, t_j) + m_3 * SS(t_i, t_j) + m_4 * ST(t_i, t_j). \tag{13}$$

The color similarity algorithm calculates the bins (bins = 25) color histogram of each channel for each area of the segmented image. It takes the minimum value of each corresponding histogram and then divides it by the area size for normalization. The calculation formula is

$$SC(t_i, t_j) = \sum_{k=1}^{n} \min\left(x_i^k, x_j^k\right). \tag{14}$$

Calculating the color histogram of the merged area as

$$C_q = \frac{\text{size}(t_i) * C_i + \text{size}(t_j) * C_j}{\text{size}(t_i) + \text{size}(t_j)}. \tag{15}$$

The new merged region size is

$$\text{Size}(t_q) = \text{size}(t_i) + \text{size}(t_j). \tag{16}$$

Shape similarity examines the proportion of overlapping areas, which is a method to measure the degree of occlusal boundary, which is more convenient and similar. It defines merged regions with shape-consistency similarity to maximize the overlap of outer rectangular regions. The smaller the bounding box after the divided regions are merged, the higher the compatibility of the shapes.

$$SF(t_i, t_j) = 1 - \frac{\text{size}(BB_{i,j}) - \text{size}(t_i) - \text{size}(t_j)}{\text{size}(im)}. \tag{17}$$

$BB_{i,j}$ represents the smallest rectangular bounding box that can wrap the two regions after merging. Dimensional similarity is to divide the number of pixels in an area to prevent a large area from continuously swallowing other small areas around it. During the calculation, a higher weight is assigned to the small area. Specifically, a similar size is obtained by subtracting the ratio of the size of one of the two regions to the size of the entire region [15]. The formula for calculating dimensional similarity is as follows:

$$SS(t_i, t_j) = 1 - \frac{\text{size}(t_i) + \text{size}(t_j)}{\text{size}(im)}. \tag{18}$$

The texture similarity class calculates the fast SIFT features of each region and uses the Gaussian distribution of variance 1 to calculate the gradients of each channel in eight different directions to extract texture features. The calculation formula is

$$ST(t_i, t_j) = \sum_{k=1}^{N} \min\left(T_i^K, T_j^K\right). \tag{19}$$

3.2. DL Algorithms. The concept of DL originated from the study of artificial neural networks. A multilayer hidden layer perceptron is a DL structure. By combining low-level features, DL can form more abstract high-level representations. Currently, learning methods such as classification and regression are mostly shallow-structured algorithms. The disadvantage is that in the case of limited samples and large amount of computation, the generalization ability for complex problems is limited. DL expresses important features of input samples through DL of nonlinear networks.

(1) Convolutional neural network structure

Convolutional neural networks are the first learning algorithms to train multilayer network structures inspired by the structure of the visual system. The former can get the transformed invariant features. This network can directly perturb the original image and avoid complex preprocessing, so it has been widely used [16, 17]. Convolutional neural networks mainly adopt three structural ideas: local receptive field, weight sharing, and downsampling. Each neuron in the CNN network defines a corresponding local receptive field and only accepts the signal transmitted from the local receptive field. The local receptive fields of

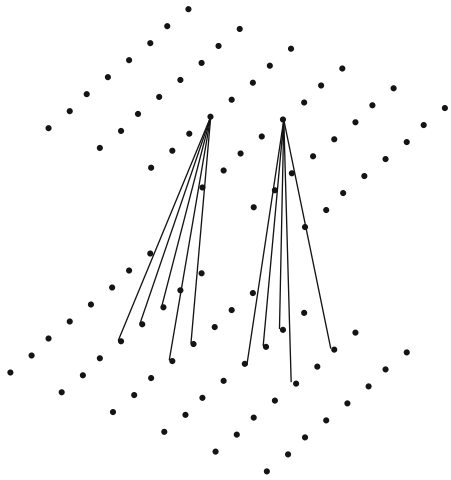

Figure 4: Convolutional layer local receptive field.

neurons on the same feature plane have the same size, as shown in Figure 4.

A typical CNN network structure is shown in Figure 5, which mainly includes a convolution layer, a pooling layer (also called downsampling), a fully connected layer, and a softmax layer. The task of CNN extracting features is done by convolutional layers, and neurons in the same feature plane are connected by weights shared by weights. Generally speaking, the deeper the network is, the greater the number of convolutional layers and nonlinear units, the greater the network capacity, and the stronger the corresponding nonlinear modeling ability. If it trains on large datasets, it usually gets good generalization ability. However, if a large-capacity model is trained on a small dataset, no matter how strong the nonlinear modeling ability is, it will not be able to obtain good generalization performance, but will lead to serious overfitting. Therefore, the choice of network structure and topology should be determined according to the size of the dataset and the actual application scenario.

The specific convolution and subsampling process for graph feature extraction is shown in Figure 6. Each layer of convolution is followed by a pooling layer, because the data dimension will rise after convolution, and if the convolution is continuous, it will inevitably fall into dimensionless mutation. The subsampling layer is similar to the convolutional layer, the neurons on each feature plane also share the connection weight, and each neuron only accepts the data in its own sensing area. The feature planes in the subsampling layers correspond to the feature planes in the convolutional layers. The appearance of downsampling is to reduce the dimension of the feature map after convolution, reduce the amount of subsequent computation, and facilitate the classification of image features.

(2) CNN-based target detection model

The R-CNN detection model is the first real object detection model based on CNN. First, the R-CNN detection model uses a candidate region extraction algorithm to extract about 2000 candidate regions that may contain objects. Then, the scale resolution of each candidate region is uniformly processed and input into the CNN network model to extract image features. The features extracted from the image are then fed into the SVM classifier for classification. The overall structure of the R-CNN target detection model is simple and clear, but there are two shortcomings that limit the efficiency of the model. One is that the entire process of extracting candidate regions from the input image is all run on the CPU. The second is that there are a lot of repeated operations in image feature extraction.

The HyperNet object detection model fuses the image features extracted from the first, third, and fifth layers of the CNN model with the local response normalization of DL to generate a high-quality fusion feature [18]. For the different resolutions of the image features extracted from the first layer, the third layer, and the fifth layer, HyperNet performs maximum clustering processing on the image features extracted from the eighth layer to generate $5 * 5 * 42$ image features. The image features extracted from the fifth layer are deconvolved. Through a series of processing of image features, the image features extracted in the third layer of convolution have the same dimension, which makes the LRN method easy to use for image feature fusion. The HyperNet detection model can obtain the detection accuracy of the fast CNN detection model with about 2000 candidate regions when only 100 candidate regions are selected. Furthermore, HyperNet has good performance in small object detection.

(3) Performance comparison

The performance of several main algorithms for CNN convolutional neural network mentioned in this section is compared on the VOC2007 test set, and the results are shown in Table 1. It can be seen that each improvement to the R-CNN model has achieved a better performance improvement. Among these algorithms, HyperNet has the best performance, but it still has some shortcomings and needs to be further improved.

4. Experiment and Result Analysis of Tennis Video Target Extraction Based on DL

4.1. HyperNet Model Improvement and System Structure Design. Since the training and operation of the DL network will perform a large number of operations, it requires high memory and processor performance. The hardware configuration used in the experiments in this paper is shown in Table 2.

DL is a developing field. In order to quickly and effectively build and train various DL models, many open source DL model frameworks have been proposed. The current open source mainstream DL frameworks and their characteristics are shown in Table 3. These DL frameworks facilitate the rapid translation of DL from theory to practical systems [19]. In this paper, Pytorch is selected as the DL framework for building video target extraction system. It is an end-to-end neural network framework open sourced by Facebook. It is simple and easy to use, has a flexible interface, and provides an active community, so it has been

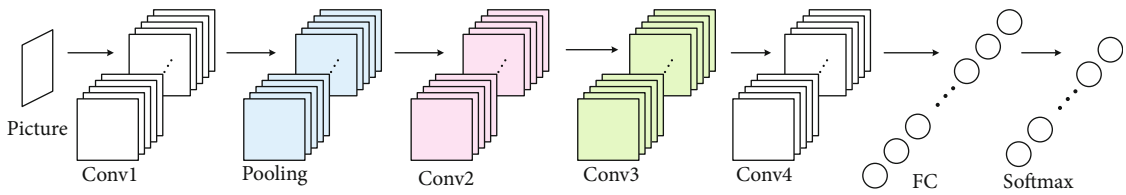

Figure 5: CNN structure.

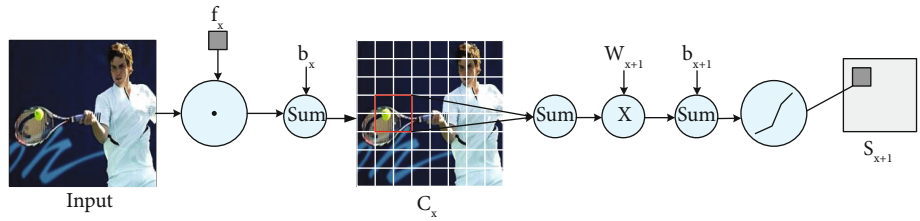

Figure 6: Convolution and subsampling process.

Table 1: Algorithm performance comparison.

Category	VOC2007mAP (%)	FPS (titan X)	Number of generated boxes	Enter picture resolution
R-CNN	65.7	5	~6000	~1000∗600
Faster R-CNN	74.1	7	~6000	~1000∗600
HyperNet	80.6	45	128	512∗512

Table 2: Experimental environment configuration.

Operating system	Ubuntul14.04
Graphics card	Intel core i7 8550U
CPU	NVDIA MX450
Memory	32 G
Hard disk	256 G

Table 3: Comparison of mainstream open source DL frameworks.

Frame name	Core language	Other interfaces	CPU supported	GPU supported
Pytorch	C++	Python	Yes	Yes
Caffe	C++	Python Matlab	Yes	Yes
TensorFlow	C++	Python	Yes	Yes
MxNet	C++	Python R	Yes	Yes
Cuda-convnet	C++	Python	No	Yes

favored by many researchers. Similarly, Caffe is concise, efficient, convenient, and easy to use, and can design, train, and deploy DL models with minimal coding. Its excellent computing performance has been sought after by the industry and researchers.

The HyperNet target extraction model mainly includes two parts, namely, the position prediction and the category prediction part. The number of layers and parameters newly added in these two parts are shown in Table 4. Conv8_2 and Conv9_2 correspond to the feature cascade obtained by dilated convolution pooling and BN layer adjustment in the original model. Then performing the convolution operation to generate Conv8_2_Conf and Conv9_2_Conf, completing the fusion of different levels of features, and then using the convolution kernel of size 3∗3 to operate to generate the position prediction of the corresponding scale default box.

Then, there is the choice of the kernel function, also known as the activation function. Its main role is to increase the expressive power of neurons and neural networks so that they can handle complex nonlinear problems. Commonly used activation functions include hyperbolic tangent function Tanh, Sigmoid function, ReLU function, and Leaky ReLU function. Their function images are shown in Figure 7. Tanh maps real numbers to a number between [0, 1], and the Sigmoid function maps real numbers to a number between [-1, 1]. The ReLU function outputs 0 when $x < 0$, and x when $x > 0$. Compared with ReLU, Leaky ReLU is no longer constant 0 when $x < 0$, but ax, where a is a constant, which is an artificially set hyperparameter. Tanh and Sigmoid functions have problems such as large amount of computation and gradient dispersion in the process of backpropagation, which make the training efficiency of multilayer neural networks low or impossible to train at all. ReLU may make neural network parameters not updated and become dead neurons. This paper chooses Leaky ReLU as the activation function of the model.

In order to analyze the performance of the improved target detection algorithm, an image target extraction system based on the improved HyperNet model is constructed. The main functions of the system are image input, target recognition and classification, marked image output, human-computer interaction interface, etc. The target extraction algorithm completes the process of region

Table 4: Newly added layers and parameters to improve the HyperNet network.

	Input layer	Output layer
Category forecast	Conv4_3 38 * 38 * 512	Conv4_3_Conf 38 * 38 * 512
	FC7 19 * 19 * 1024	FC7_Conf 19 * 19 * 1024
	Conv8_2 10 * 10 * 512	Conv8_2_Conf 10 * 10 * 512
	Conv9_2 5 * 5 * 256	Conv9_2_Conf 5 * 5 * 256
	Conv3_1 75 * 75 * 128	Conv3_1 75 * 75 * 128
Location prediction	Conv4_3 38 * 38 * 512	Conv4_3_loc_pl 38 * 38 * 512
	FC7 19 * 19 * 1024	FC_7_loc_pl 19 * 19 * 1024
	Conv8_2 10 * 10 * 512	Conv8_2_loc_p 10 * 10 * 512
	Conv9_2 5 * 5 * 256	Conv9_2_loc_p 5 * 5 * 256
	FC7_loc 19 * 19 * 1024	FC7_f 19 * 19 * 1024

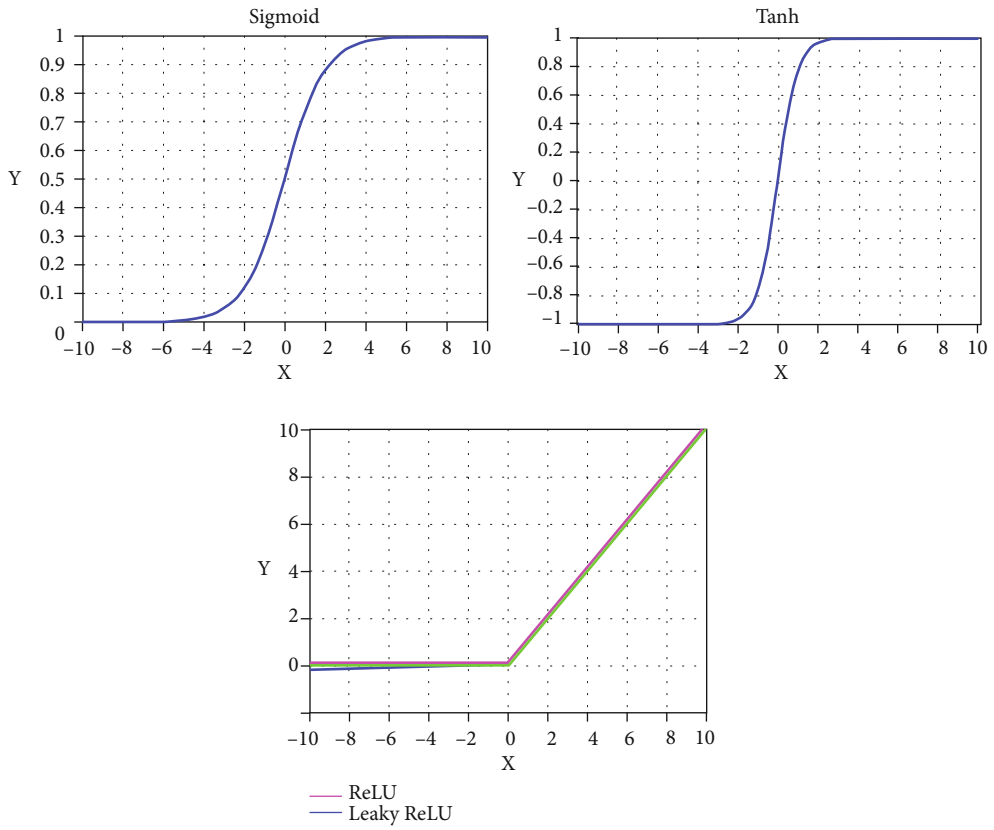

Figure 7: Several activation function images.

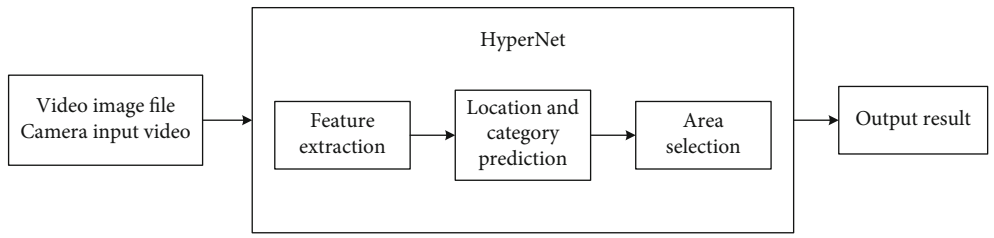

Figure 8: Object extraction system structure.

selection, feature extraction, location, and category prediction. The overall system structure is shown in Figure 8 [20].

4.2. HyperNet Model Training and Performance Testing. To verify the effectiveness of the improved HyperNet target extraction model proposed in this paper in extracting candidate regions and tracking targets, this paper selects the PASCAL VOC 2012 dataset and the Video53 dataset. Among them, the PASCAL VOC dataset is one of the most widely used evaluation datasets in the field of object classification and detection, containing 5751/5823 training/validation samples and 10991 testing samples. The Video53 dataset is a continuously shot video. The main training hyperparameters involved in the training process are the training process is performed for 1000 iterations. The basic learning rate is 0.0001 for the first 1500 training sessions, 0.00001 for the last 500 training sessions, and the learning rate decay weight is 0.0005. In the network training part, forward propagation calculates the prediction result and passes the prediction result and the true value through the loss function to calculate the loss value. The training loss curves of the model on the two types of datasets are shown in Figure 9. It can be seen that the improved HyperNet target extraction model proposed in this paper quickly converges on the PASCAL VOC 2012 dataset and the Video53 dataset training set after the data augmentation method. The loss value of loss function is stable between 1.5% and 2.5%, the performance is relatively stable, and the convergence speed is fast.

Comparing this model with other target extraction models (mainly YOLOv2 model, Faster R-CNN detection model, and SSD300 model). SSD300 means that the SSD target detection model uses an image with a resolution of 300×300 as input, and the YOLOv2 target detection model selects an input image resolution of 416×416 in this experiment. The data set selected for comparison is the PASCAL VOC2007 test set, and the evaluation standard used is the average detection accuracy of the area enclosed by the P-R curve and the coordinate axis. Among them, the vertical axis coordinate of the P-R curve is the precision rate, which measures the proportion of the actual target in all the detection results. The horizontal axis of the P-R curve is the recall rate, which measures the proportion of the actual target in the detection result, reflecting the recall ability of the detection model. Figure 10 shows the average detection accuracy value calculated by the P-R curve for the 20 types of targets to be detected in the VOC2007 test data set. On the whole, the average detection accuracy of the HyperNet target extraction model in this paper is improved compared with other target extraction models in 7 categories. The mPA value of the HyperNet target extraction model is 78.36%, the mPA value of the YOLOv2 model is 76.93%, the mPA value of the Faster R-CNN detection model is 75.95%, and the mPA value of the SSD300 is 73.91%. This can prove the feasibility of the improved HyperNet model in this paper for target detection tracking and region extraction. For the case of high complexity of the background and target environment, due to the increased detection difficulty, the discrimination and extraction of several models are not very optimistic. However, for tennis video target extraction, the background in tennis is relatively simple and not very complicated. Therefore, it is feasible to apply the model proposed in this paper to target extraction from tennis videos.

4.3. Application of Tennis Video Object Extraction of HyperNet Model. In the experiments in this section, the data set selected for the experiment comes from a video of a tennis match. The video set is formed by 6 high-speed cameras placed in different directions of the competition venue, and the resolution of each camera is 1080∗1920HD. Using this dataset to detect and analyze the improved HyperNet target extraction model proposed in this paper, and comparing the model target extraction results with the real situation of the target to be detected. This experiment uses this data set to detect and analyze the improved HyperNet target extraction model proposed in this paper and compare the model target extraction results with the real situation of the target to be detected. The number of frames of the video is randomly selected for detection, the accuracy rate is averaged, and the results are as follows: the orientation accuracy value is 96.32%, and the size accuracy is 91.05%, which basically meets the needs of video target extraction. The extracted region speed and the number of region candidate boxes detected by this algorithm are shown in Table 5. Macroscopically, the three algorithms are evaluated in terms of the number of region candidate boxes and the time it takes to extract regions. On the whole, the number of boxes for extracting candidate regions is small, which means that the extraction accuracy may be higher. Moreover, the number of frames processed per second also has a good performance, the calculation time is greatly reduced, and the calculation time for each frame will not exceed 1.8 s.

5. Discussion

The research direction of this paper is "Tennis Video Target Extraction Based on DL," which belongs to the research field of artificial intelligence and computer vision. Object extraction is a challenging topic in the field of computer vision, and its main purpose is to detect and localize specific objects from still pictures or videos. This paper firstly sorts out the related research and summarizes the main framework of the research. Second, several traditional algorithms of image target extraction are introduced, including frame difference method, image segmentation method, background difference method, and region merging algorithm. Then, it summarizes the image target advance algorithm of DL, mainly focusing on the analysis and introduction of several models generated by convolutional neural networks. And with the average extraction accuracy as the indicator, the performance comparison of R-CNN, Faster R-CNN, and HyperNet on the VOC 2007 dataset is summarized. Then, the HyperNet model is adjusted, the commonly used activation functions are introduced, and an improved HyperNet image target extraction model is designed based on the previous foundation. The model contains processes such as region selection, feature extraction, location, and class prediction. Then, the HyperNet model was trained, learned, and performance tested. First, the VOC 2012 data set and the Video 53

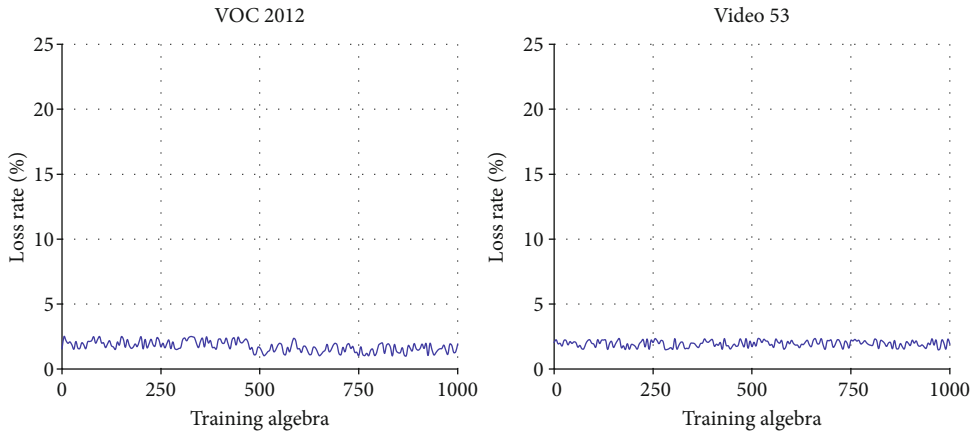

Figure 9: Model training loss rate curve.

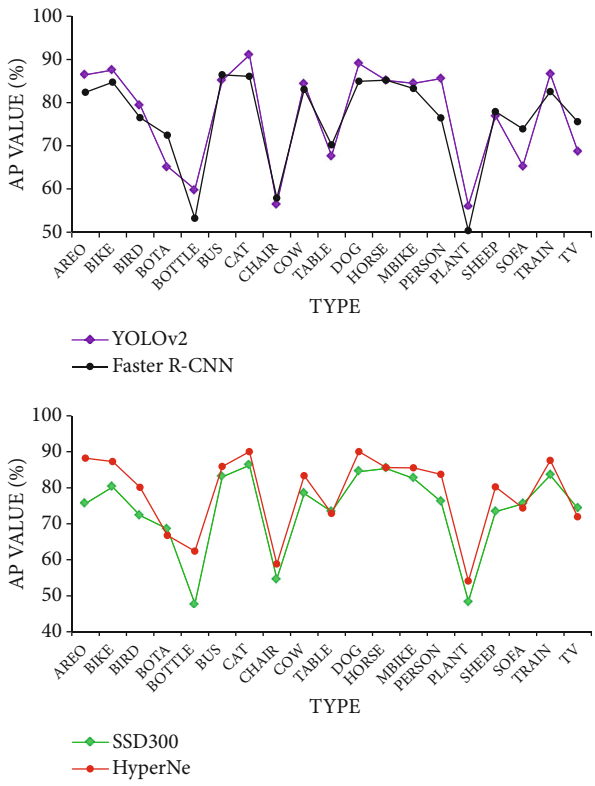

Figure 10: Performance test of each model on VOC2007.

data set are used as the training and test content, and the data enhancement method is tested. The results show that the performance of the model algorithm is relatively stable, and the convergence speed can meet the application requirements. Then, the PASCAL VOC2007 dataset is used as the experimental object to compare the performance of the model proposed in this paper and several other models with better performance. The models used for comparison mainly include YOLOv2 model, Faster R-CNN detection model, and SSD300 model. Compared with the average extraction accuracy of mpA, the results show that except for the complex background environment, the target extraction accuracy in other scenes can meet the requirements. And among several methods, the average accuracy of the improved HyperNet video target extraction model in this paper is the highest. Finally, using an actual tennis match video as the research data set, the number of candidate regions and the extraction speed of the model proposed in this paper are tested, the results show that the number of candidate boxes extracted by the model is greatly reduced, and the calculation speed is also in line with expectations.

Table 5: Extraction candidate area and extraction speed of HyperNet model.

Frame number	Bounding box	Speed
Frame 26	31	1.462 s
Frame 39	26	1.261 s
Frame 137	44	1.572 s
Frame 294	16	1.021 s
Frame 96	35	1.417 s
Frame 319	26	1.258 s
Frame 283	29	1.342 s

6. Conclusion

The main task of object recognition is to identify the types of objects in an image, and the detection problem can usually be boiled down to the problem of multiobject recognition in a given image. Although feature learning based on convolutional neural network can achieve good results in the task of video image object extraction. But on the other hand, in the field of long-term object detection, even in the field of computer vision, the more effective the design of the network structure, the higher the efficiency of feature representation. Starting from the feature learning method based on convolutional neural network, this paper looks forward to the next work: first, improving the nonlinear unit to guide more effective feature representation. The second is the research on the initialization of affine transformation parameters. The third is the in-depth study of small target detection methods. Improving the accuracy of small target

detection is valuable for many practical application scenarios.

Data Availability

No data were used to support this study.

Conflicts of Interest

The author declares that there is no conflict of interest with any financial organizations regarding the material reported in this manuscript.

References

[1] L. Li, "Research on target feature extraction and location positioning with machine learning algorithm," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 429–437, 2020.

[2] K. Tang and L. J. Huo, "Optimizing synchronization of tennis professional league live broadcast based on wireless network planning," *Mobile Information Systems*, vol. 2021, no. 7, Article ID 8732115, 9 pages, 2021.

[3] T. Fan, "Research and realization of video target detection system based on deep learning," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 18, no. 1, article 1941010, 2020.

[4] H. Pang, Q. Xuan, M. Xie, C. Liu, and Z. Li, "Research on target tracking algorithm based on Siamese neural network," *Mobile Information Systems*, vol. 2021, no. 4, Article ID 6645629, 11 pages, 2021.

[5] J. Li, X. Li, D. He, and Y. Qu, "A domain adaptation model for early gear pitting fault diagnosis based on deep transfer learning network," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 234, no. 1, pp. 168–182, 2020.

[6] D. Yang, "Target tracking algorithm based on adaptive scale detection learning," *Complexity*, vol. 2021, no. 6, Article ID 9033912, 11 pages, 2021.

[7] C. Li-quan, L. You, F. Shen, Z. Shan, and J. Chen, "Pose recognition in sports scenes based on deep learning skeleton sequence model," *Journal of Intelligent & Fuzzy Systems*, vol. 3, pp. 1–10, 2021.

[8] H. Peng and Q. Li, "Research on the automatic extraction method of web data objects based on deep learning," *Intelligent Automation & Soft Computing*, vol. 26, no. 3, pp. 609–616, 2020.

[9] A. Raschke and M. Lames, "Video-based tactic training in tennis," *German Journal of Exercise and Sport Research*, vol. 49, no. 3, pp. 345–350, 2019.

[10] D. Demir Sahin, E. Isik, I. Isik, and M. Cullu, "Artificial neural network modeling for the effect of fly ash fineness on compressive strength," *Arabian Journal of Geosciences*, vol. 14, no. 23, pp. 1–14, 2021.

[11] X. Liu and Z. Zhang, "A vision-based target detection, tracking, and positioning algorithm for unmanned aerial vehicle," *Wireless Communications and Mobile Computing*, vol. 2021, no. 7, Article ID 5565589, 12 pages, 2021.

[12] C. Zhang and X. Liu, "Feature extraction of ancient Chinese characters based on deep convolution neural network and big data analysis," *Computational Intelligence and Neuroscience*, vol. 2021, no. 3, Article ID 2491116, 10 pages, 2021.

[13] Y. Liu and Y. Ji, "Target recognition of sport athletes based on deep learning and convolutional neural network," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 2253–2263, 2021.

[14] Y. Zhang, R. Alturki, H. J. Alyamani, M. A. Ikram, A. . Rehman, and M. Haleem, "Multilabel CNN-based hybrid learning metric for pedestrian reidentification," *Mobile Information Systems*, vol. 2021, no. 7, Article ID 5512382, 7 pages, 2021.

[15] J. Peng, K. Fu, Q. Wei, Y. Qin, and Q. He, "Improved multi-view decomposition for single-image high-resolution 3D object reconstruction," *Wireless Communications and Mobile Computing*, vol. 2020, no. 5, 14 pages, 2020.

[16] H. Li, X. Han, and Z. Fang, "A visual model of welding robot based on CNN deep learning," *Acta Welderica Sinica*, vol. 40, no. 2, pp. 154–160, 2019.

[17] D. Sun, J. Wu, H. Huang, R. Wang, F. Liang, and H. Xinhua, "Prediction of short-time rainfall based on deep learning," *Mathematical Problems in Engineering*, vol. 2021, no. 5, Article ID 6664413, 8 pages, 2021.

[18] J. Brooke, A. Hammond, and G. Hirst, "Using models of lexical style to quantify free indirect discourse in modernist fiction," *Digital Scholarship in the Humanities*, vol. 32, no. 2, pp. 234–250, 2017.

[19] X. Fan, S. Hu, and J. He, "A dynamic selection ensemble method for target recognition based on clustering and randomized reference classifier," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 3, pp. 515–525, 2019.

[20] J. Gu, T. Su, Q. Wang, X. Du, and M. Guizani, "Multiple moving targets surveillance based on a cooperative network for multi-UAV," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 82–89, 2018.