Real-Time Tracking Method of Students' Targets in Wushu Distance Teaching Based on Deep Learning

Jie Zhang[1(✉)] and Na Ma[2]

[1] Shaanxi University of Chinese Medicine, Xianyang 712046, China
zhaoliangyuan526@sina.com
[2] Sports Department, Tianjin Foreign Studies University, Tianjin 300204, China

Abstract. The development of information technology has promoted the development of distance education, and Wushu teaching has gradually changed from traditional face-to-face teaching to distance education. In order to improve the teaching effect, it is necessary to track the teaching objects in real time. In order to improve the real-time tracking of students' goals in Wushu distance learning, a real-time tracking method of Wushu distance learning goals based on deep learning is designed. The experimental results show that the designed real-time tracking method has a short tracking delay, which proves its effectiveness.

Keywords: Deep learning · Wushu · Distance teaching · Goal · Real-Time tracking

1 Introduction

As one of the important directions of computer vision, the main task of target tracking is to extract the motion information and position information of the target [1], which provides the basis for the semantic analysis of later behavior. Target tracking can be defined as: selecting an area as a target in the first frame of a video stream, automatically finding the target in the next several frames, and outputting the target position. With the development of society [2, 3], people are increasingly pursuing a more intelligent life. Video surveillance exists in many places in our lives, including criminal investigation monitoring, traffic vehicle monitoring, community security monitoring, etc. At present, many of these surveillance videos are artificially tracked. Due to the uncertain factors of manual operation, important information may be missed [4], so the research on target tracking algorithm has strong research value and wide application fields.

At present, many tracking algorithms are in the stage of simulation experiment, and there are few real applications. Target tracking faces a series of challenges, such as complex and changeable environment, non-rigid deformation of target, occlusion of target, illumination change, long-term tracking and so on, which leads to great limitations in the research of moving target recognition and tracking technology [5]. Therefore, the

W. Fu and G. Sun (Eds.): eLEOT 2022, LNICST 453, pp. 695–708, 2022.
https://doi.org/10.1007/978-3-031-21161-4_53

research on tracking and recognition based on extended targets is of great value in practical application and indispensable significance in theoretical research.

The main tasks of target tracking include acquiring video sequences, preprocessing the video sequences, giving the target to be tracked, extracting target features, matching or binary classification, and finally giving the location or behavior track of the target. The target tracking task needs to meet the requirements of stability, accuracy and real-time. Because of the limitations of algorithms, hardware and large amount of calculation, target tracking is not widely used in commercial applications. Therefore, it is of great significance to study the real-time and stable tracking of extended targets in complex background.

Inspired by convolutional networks, some scholars use this advantage of convolutional neural networks to use layers with switch mechanism in the tracking process. Ma et al. have done similar work [6, 7]. They use training CNN [8] on ImageNet to improve the tracking accuracy and stability. This algorithm uses switching between layers with semantic information and fine-grained information, and fuses information from layers for tracking strategy from coarse to fine. However, these algorithms mentioned above are all pre-trained offline on ImageNet. And then directly used for online tracking, and the network will not be fine-tuned in the tracking process. This tracking network that only uses target image data for pre-training is unreliable, because the target in one video can be the background in another video. Therefore, some scholars have proposed a new tracking network [9] to solve the above problems. The network uses video data pre-training-a two-layer tracker based on CNN. This method effectively adjusts the pre-learned features according to specific targets during online tracking. Nam proposed a video training CNN network [10] with common network and multiple branches to distinguish the target from the background. However, these video training trackers don't explicitly use the semantic information of the target, that is, they don't know the category of the object. Without knowing the category of the object, the tracker will probably fail to track.

With the advent of the Internet age, many scholars began to innovate Wushu teaching methods by means of the Internet, improve teachers' teaching quality and optimize students' learning mode. The traditional real-time tracking method of Wushu distance teaching students' goals has poor tracking effect and can't meet the real-time demand of tracking. Therefore, this paper designs a new real-time tracking method of Wushu distance teaching students' goals based on deep learning, which can be used as a reference for improving the subsequent Wushu teaching effect. With the improvement of hardware capability and the rapid progress of computing capability, the society is developing towards intelligence. It is imperative to introduce deep learning into the target tracking task, extract target features through convolutional networks, and study the use of deep learning algorithm to complete the target tracking task. By training a large number of different environments and different targets, we can adapt to the tracking of different targets in different environments and get more robust features.

2 Design of Real-time Tracking Method for Students' Targets in Wushu Distance Education Based on Deep Learning

2.1 Determine the Tracking Characteristics of Wushu Distance Teaching Goals

The most important step of target tracking is to extract the features of the target object. Feature extraction is to convert pixel features into semantic features, and this semantic feature can be divided into traditional artificial features and learning-based features in deep learning. Whether the extracted target features are more abundant and can represent the target determines the accuracy and stability of target tracking, while the amount of calculation in the process of feature extraction determines the speed of target tracking.

The traditional feature extraction method is artificially designed according to subjective judgment, which can be divided into generative model features and discriminant model features according to mathematical models. Generative model feature is that the target is represented by template, and tracking is realized by matching the template feature with the candidate target feature in the search area, such as sparse representation, and the candidate target with the sparse coefficient and the smallest error is selected as the result. Discriminant model features refer to binary classification features, which distinguish the target from the background, such as TLD. According to the different calculation of visual features, it can be divided into pattern features, gradient features, shape features and color features. Pattern features include sliding window features obtained by Gabor filter to simulate human eye receptive field, dot-line features expressed by the sum of pixel differences in adjacent matrix areas, LBP local features for extracting illumination invariant characteristics, etc. Gradient features include SIFT features of moving targets by obtaining gradient information around key points, and HOG features of moving targets by gradient direction and intensity of spatial distribution areas.

Because the traditional feature is a certain target feature representation method set artificially, the robustness is limited, and the representation effect is better for specific scenes or specific objects, but there are some limitations for complex background and transformed targets in natural environment. The method based on deep learning generally uses convolution network to extract features, which is a learning-based feature. By modifying the weights of convolution layer through the training process, a convolution network with feature extraction ability is finally obtained. This feature is different from the traditional manual setting feature, it can't be presented in a way that can be understood by people, and it is a feature that can be understood by the network. Generally, the shallow network extracts the edge features and location information of the target, while the deep network extracts the semantic information, which is used for tasks such as classification and identification.

The advantages of traditional feature representation are small amount of calculation and high speed, while the disadvantages are that feature representation ability is not strong enough and features are not rich enough. The features extracted by convolution network have the advantages of good feature robustness, rich features and strong representation ability, and are suitable for the transformation target of complex scenes. However, the amount of calculation is large, and a large enough data set is needed for training. With the development of computer technology, the computing power of hardware is gradually enhanced, and new network frameworks are put forward one after

another. The future development prospect of deep learning is immeasurable, so the algorithm based on deep learning has room for development and application value. As far as the current development trend is concerned, how to enrich the features and reduce the amount of calculation is a problem that needs to be solved at present. Based on this, a neural network for judging the features of Wushu distance teaching target tracking can be designed, as shown in Fig. 1 below.

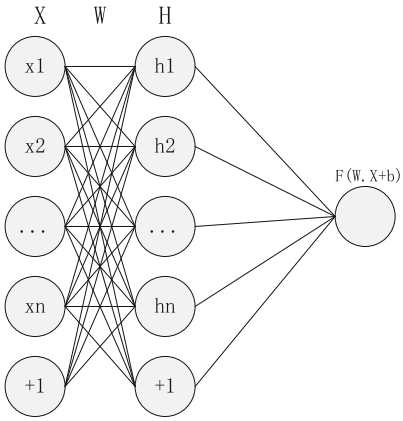

Fig. 1. Tracking feature judgment neural network

It can be seen from Fig. 1 that the multilayer neural network, that is, the number of hidden layers increases, is used for nonlinear classification problems. In multilayer neural network, the leftmost layer is the input layer, the rightmost layer is the output layer, and the middle layer is collectively called the hidden layer. The layers are connected with each other, and the data is transmitted backward in turn, and the error is transmitted backward and forward [11]. The so-called training is the process of inputting the training data set into the network, comparing the output layer with the training label, and propagating the error back, so as to constantly update the connection weights.

In the neural network, the activation function exists between the output of the upper layer and the input of the lower layer, and its function is to apply the neural network to nonlinear relationship fitting. The commonly used activation functions are Sigmoid function, tanh function and modified linear function ReLU. Sigmoid function has advantages in compressing data amplitude, but there is a problem of gradient disappearance in deep neural network. At the same time, due to the exponential operation, when the training process consumes more, draw the real-time tracking decision function change image as shown in Fig. 2 below.

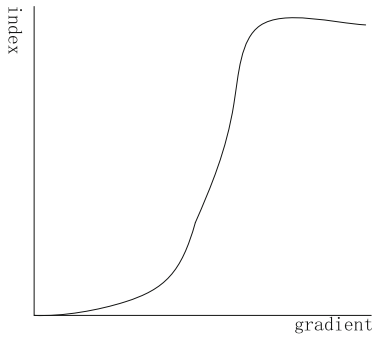

Fig. 2. Real-time tracking function change image

As can be seen from Fig. 2, since the peak value of the gradient function is 0.25, the gradient decreases by a factor of 0.25 during the propagation, which easily leads to gradient dispersion. Sigmoid function has advantages in compressing data amplitude, but there is a problem of gradient disappearance in deep neural network. At the same time, because of exponential operation, the training process consumes more time.

On the basis of the above analysis of the tracking characteristics of Wushu distance teaching objectives, a real-time tracking model of Wushu distance teaching objectives is constructed by deep learning.

2.2 Design of Real-Time Tracking Model of Wushu Distance Teaching Objectives Based on Deep Learning

Deep learning is to learn the inherent laws and representation levels of sample data, and the information obtained in the learning process is of great help to the interpretation of data such as words, images and sounds. Its ultimate goal is to make the machine have the ability to analyze and learn like a human being, and recognize data such as words, images and sounds. Therefore, this paper designs a real-time tracking model of Wushu distance teaching goals by using deep learning neural network. First, it is necessary to determine the training index of the model, and the calculation formula is as follows (1):

$$E = \frac{f}{a_t t + I} \tag{1}$$

where at represents the compression amplitude of neural network, f represents the value, t represents the gradient of deep neural network, and I represents the training index. Tanh function is a zero mean function, which will not lead to all positive or all negative local gradients, so the convergence speed will be accelerated and the effect is better than Sigmoid function. However, the gradient of tanh function also disappears, and the calculation amount of exponential operation still exists. At this time, the design index optimization formula is shown in the following (2).

$$\tanh = \frac{e - e_0}{e + e_0} \tag{2}$$

In formula (2), e represents the input value of neural network, e0 represents the optimization parameter, and void convolution is often used in image segmentation, which can keep information better than pool layer in the process of down-sampling, and can restore images more completely after up-sampling. The role of void convolution in the image field is not only in image segmentation, but also can be considered when it is necessary to expand the receptive field, reduce the amount of calculation and obtain small features. The designed real-time tracking method of distance education targets adds void convolution calculation, and the schematic diagram of convolution kernel division is shown in Fig. 3 below.

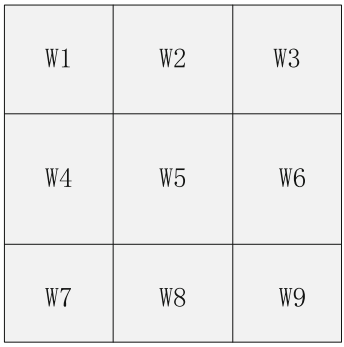

Fig. 3. Schematic diagram of convolution kernel division

It can be seen from Fig. 3 that hole convolution can increase the receptive field and reduce the feature size with the same amount of calculation. Before the empty convolution was put forward, the downsampling in the convolution network was usually done through the pooling layer. The role of pooling is to reduce the image size and increase the receptive field, but the process of size reduction is accompanied by the loss of information.

At present, there are many algorithms in the field of target tracking, and the performance evaluation of the algorithms is reflected in the accuracy, real-time performance and stability of the algorithms. In order to make a fair comparison of various algorithms, the academic community has given public test sets and evaluation indicators, and the tracking model based on this design is shown in (3) below.

$$overlap = \frac{R \cup G}{R \cap G} \tag{3}$$

In the tracking model (3), R represents the real tracking logo and G represents the logo box. In the tracking field, different application scenarios have different requirements for real-time performance, so the limitation of real-time performance depends on specific occasions. In this paper, the basic requirement of real-time performance is 20FPS, and improving the speed on the basis of setting is the requirement to improve the performance of the algorithm.

Therefore, the real-time tracking model of Wushu distance teaching target is constructed, and on this basis, the model is further optimized through offline training and online updating.

2.3 Optimizing FDT Net Wushu Distance Teaching Goal Tracking Algorithm

MDNet is a tracking network structure, using convolution layer to extract features, and full connection layer as classification layer. Different from ordinary classification, it is a binary classification, that is, it is enough to distinguish the target from the background. Multi-domain structure exists to adapt to different tracking scenes. QE trains one FC6 for each input video, thus training multiple FC6 for each calibrated tracking video. The previous Conv1-Conv3 and FC4-FC5 are shared layers.

Conduct off-line training and online update on the network. The parameters of convolution layer are obtained during off-line training, while in the tracking process, the parameters of convolution layer are kept unchanged, and the parameters of full connection layer FC4-FC6 are fine-tuned online. The renewal strategy also takes into account the long-term and short-term, and at the same time, it increases the intensive training of hard-to-distinguish samples. The network structure of. NET is designed for tracking tasks. The update strategy is considered carefully and the tracking accuracy is high. However, due to online update and scoring mechanism of multiple candidate boxes similar to the detected pattern, the calculation of tracking process is too large, which reduces the tracking speed. With OTB and VOT2014 as test data sets, it can only reach 1FPS on GPU platform.

The reasons for the low speed of MD. NET algorithm can be analyzed as follows: the features of the full connection layer are fully connected, resulting in a large amount of calculation; Multi-domain FC6 layer, learning parameters increase, on the one hand, the convergence speed slows down, on the other hand, a new FC6 is needed for fine-tuning when tracking; It takes time to track and update the parameters of the full connection layer online. At this time, the tracking algorithm of Wushu distance teaching goal is designed as shown in (4) below.

$$P = \frac{\sqrt{overlap}}{\frac{1}{s(t)}} \tag{4}$$

When the algorithm is optimized, the low-level network is trained first, and then the high-level network is initialized with the weights of the low-level network. The purpose of this measure is to shorten the network training time and make the network converge faster. The reason why VGGNet performs well in the positioning task lies in the use of multiple 3*3 convolution kernels. Two 3*3 convolution layers can be superimposed, which is equivalent to a 5*5 convolution kernel effect, but there are fewer parameters to learn. And after the superposition of several convolutions, many nonlinear transformations have been completed, which has stronger feature learning ability.

The reason why VGGNet performs well in classification tasks lies in its structural features. With the deepening of network layer, the number of channels of output features increases and the resolution decreases. The advantage of this design is that the multi-dimensional features are converted into classification vectors, which is convenient to

connect the full connection layer as the output. Conv1-Conv3 convolution layers come from VGGNet, the input and output of convolution layers. Unlike VGGNet, the input size has been modified. The input size of FDTNet is 107*107, which is set to get the 3*3 feature map after Conv3 layer. The convolution layer is followed by the full connection layer FC4-FC6. FC4 and FC5 each have 512 output units, while FC6 is the second classification layer, which is used to distinguish the target from the background. For the selection of convolution layer, the factors to be considered are the amount of calculation and feature sufficiency. The low convolution layer extracts the line features and edge features in the image, which belongs to the bottom information, while the high convolution layer extracts the semantic information in the image, which belongs to the high features.

In the operations of target detection and target recognition, the convolution layer is generally deeper, which is used to extract enough features. However, the deeper the convolution layer is, the larger the calculation amount is, and the worse the real-time performance is. For the subject of target tracking in this paper, it is only necessary to judge the target and background, that is, the requirements for classification tasks are not high. In this paper, three convolution layers are selected, the first two layers are used to extract the edge information of the target, and the third layer is used to extract the global information of the target and increase the receptive field of features. Of course, if you choose a deeper convolution layer, the feature expression ability of the target will be stronger, which is beneficial to the accuracy improvement, but it will increase the computational complexity to a certain extent. In this paper, the selection of convolution layers takes into account the feature sufficiency and computational complexity.

According to the analysis of the network structure of MDNet and the reasons for its low speed, the improvement of FDTNet can be summarized as follows: the full connection layer Fc6 is modified into a single domain, the convolution layer Conv1-Conv3 extracts the target features, and the full connection layer performs the target and background binary classification operation. ILSVRC2015 detection data set is used as training data set, and the types and quantity of training data sets are increased, so as to ensure that convolution layer has enough feature extraction ability in the training process. The fully connected layer of a single domain reduces the amount of computation and improves the training speed and convergence. Enlarge the training data set to ensure that the improvement does not affect the accuracy. The parameters of the full connection layer are not updated online, the parameters of the convolution layer and the full connection layer are kept fixed, the parameters are not modified, the calculation memory is not occupied, and only the feedforward network calculation is performed. This improvement improves the tracking speed. A new linear update rule is added, and the final position is determined by the weight of the regression position and the position of the previous frame. Adding a superparameter increases the selection improvement of the tracking precision tracking candidate frame, limiting the selection position of the tracking frame to 2*2 times that of the target frame of the previous frame, reducing the number of selection frames, reducing the amount of calculation and improving the tracking speed, but losing the stability.

After the training, keep the network parameters unchanged, and give the target frame of the first frame of video. For each subsequent frame, in the area of 2*2 times of the

target position of the previous frame as the center, 32 candidate frames are selected by multi-dimensional Gaussian distribution in three dimensions of width, height and scale. The candidate frames are input into the network with a uniform size of 107*107, and the output of the network is a two-dimensional vector, which indicates the probability of the target and background corresponding to the input candidate frames. The three candidate boxes with the highest goal scores are the target candidate boxes.

For each tracking video, the location of the target in the first frame is known, and the candidate frame of the first frame, the network output corresponding to the first frame and the actual location of the target in the first frame are taken as the data of regression network training, and the trained regression network is used to predict the frame where the target is located in the video. 800 samples are selected from the first frame of the tracking test video sequence, and the selection rules of samples are consistent with those of tracking candidate frames. These 800 samples are taken as the training data set of regression network, and the real position of the target in the first frame of the test video is taken as the label of regression training, and the trained regressor can be directly used for the regression positioning operation of the subsequent frames of the video.

Training video tracking through selected data sets, and tracking students' goals in real time in distance Wushu teaching according to the constructed model.

2.4 Real-Time Tracking of Students' Goals in Wushu Distance Teaching

Real-time tracking of students' targets in Wushu distance education can also improve the existing tracking algorithms. Traditional tracking algorithms such as KCF are excellent in speed, and correlation filtering is used to obtain the correlation of targets, which has the advantage of small amount of feature calculation. The disadvantages are that there is a boundary effect, the detection range is limited, and the adaptive scene is limited due to the limited feature extraction ability. Combined with KCF's correlation thought, the Siam-FC tracking network with cross-correlation is considered to be improved. Cross-correlation is used to obtain target correlation, which belongs to sliding window detection, and has the efficient realization of full convolution. The calculation amount is mainly determined by the number and size of features involved in sliding window detection. Siam-FC has relatively good real-time performance, and its accuracy needs to be improved. Then the improvement idea based on Siam-FC is to reduce the amount of feature calculation, improve the robustness of features and obtain richer semantic features. Siam-MF is an extended target tracking network based on Siam-FC, which aims to improve the accuracy and real-time performance while ensuring the tracking stability, so that the tracking algorithm can adapt to various application environments.

The convolution layer of Siam-FC comes from AlexNet, and the tracking task is completed by convolution feature correlation. The structure of the network is easy to understand. After feature extraction of the target template and the search area, similarity measurement is carried out, and the location of the target in the search area is located according to the similarity score map. The feature extraction task is completed through the convolution layer of the network, and the whole convolution layer is used as the similarity measurement function. The training set is ILSVRC2015 data set, and the shallow convolution layer is selected to extract features. The training process mainly completes the determination of convolution layer parameters, while the tracking process

does not need to update the parameters. Compared with MDNet, the comprehensive performance is stronger and the accuracy is slightly lower than that of MDNet, but the speed of OTB test set reaches 58FPS. Siam-FC has strong comprehensive performance, but it needs to be improved in speed and accuracy. In-depth analysis of Siam-FC network shows that the extraction ability of convolution layer affects the tracking accuracy, and it can be considered to improve the accuracy by improving the feature extraction ability of convolution layer. The reason that affects the stability lies in the selection of tracking template. Siam-FC chooses the first frame target as the template and does not update it, and the subsequent changes of target shape will affect the stability. The main reason that affects the speed lies in the whole convolution layer, and the more characteristic channels for cross correlation, the greater the amount of calculation.

The main line part of the convolution layer of Siam-MF comes from five convolution layers Conv1-Conv5 of AlexNet, and the empty convolution layers Skip1 and Skip2 are added for the output of Conv1 and Conv3 and the feature fusion of Conv5 respectively. For the selection of convolution layer, the calculation amount and feature sufficiency are considered. The lower convolution layer represents edge information and position information, extracts line features and edge features from images, and belongs to the bottom information. The high-level convolution layer extracts the semantic information in the image, which belongs to the high-level features. In the operations of target detection and target recognition, the convolution layer is generally deeper, and enough features are extracted for classification. However, for the network with deeper convolution layer, the greater the computation, the worse the real-time performance. For the subject of target tracking in this paper, only target features need to be extracted, which is not used for classification, that is, the semantic features of the target are not high. In this paper, five layers of convolution are selected to fuse the features of the outputs of Conv1, Conv3 and Conv5, so as to obtain richer target features.

The network parameters of Siam-MF are obtained through training. The input of the network is preprocessed pictures, including the target template and the search area. The search area is 2*2 times of the location of the target in the previous frame. After cutting out the candidate box and transforming the size of the original picture, the template with the size of 127*127*3 and the search area with the size of 255*255*3 are obtained. After feature extraction by convolution layer, the correlation between the target and the search area is finally obtained through full convolution. For the tracking process, the feature of the target template and the search area is extracted by convolution layer, and then the cross-correlation analysis is carried out by full convolution layer to get the score map of the target in the search area, which can be considered that the position where the maximum value is located is the target position.

In view of the in-depth analysis of Siam-FC, the improvement of the original network can be summarized into the following four points: using feature fusion to obtain more comprehensive features. Output Conv1 through Skip1 layer and Conv3 through Skip2 layer. At the same time, feature fusion is carried out with the output of Conv5 to obtain richer target features and improve tracking accuracy. Introduce void convolution. Void convolution increases the receptive field, reduces the amount of calculation and improves the tracking speed and accuracy. Using void convolution in Skip1 and Skip2 layers can keep the receptive field of features and reduce the amount of calculation. Number of

channels is reduced. Reducing the number of characteristic channels entering the whole convolution layer will greatly reduce the amount of calculation. The main calculation amount of cross correlation comes from sliding window detection of full convolution network. For Conv5 layer, the number of feature channels is reduced, the reduced features are compensated by feature fusion, and the final features are combined with different levels of data. This operation reduces the amount of calculation and improves the feature extraction ability.

The convolution layer extracts the template features and extracts the search area features of each frame. The void convolution is added to the connection layers Skip1 and Skip2 to reduce the number of characteristic channels of Conv5 in AlexNet, and the output of Conv1 through Skip1 is matched with the output of Conv3 through Skip2 and the output of Conv5, so that the receptive field is increased and the amount of calculation is reduced. The process of Siam-MF network is as follows: the target and search area pass through the same convolution network to extract the features of the target and search area; The feature layer of the target and the feature layer of the search area complete the cross correlation through full convolution, and the correlation graph of the target in the search area is obtained, and the position of the maximum value of the correlation graph is the central position of the target in the search area. According to the motion orientation of the extended target in the video, the search area is set to be 2*2 times the size of the target frame. At the same time, the deep convolution network leads to the loss of position information, which is undesirable in the field of tracking. Therefore, in this paper, the output features of different layers of convolution layer are fused, combining the position information of the shallow layer, the extracted features of the middle layer and the semantic information of the deep layer.

3 Test

In order to verify the tracking effect of the real-time tracking method of students' targets in Wushu distance education based on deep learning designed in this paper, this paper compares it with the traditional real-time tracking method of students' targets, and carries out experiments as follows.

3.1 Prepare for the Experiment

The experiment is carried out on ILSVRC2015 dataset. This data set includes 3862 snippets for training, 555 snippets for verification and 937 snippets for testing. Each snippet includes 56 ~ 458 frames of images. Conv1-Conv5 used in feedforward network uses the Conv1-Conv5 layer of AlexNet. In the tracking video sequence, the target is generally not too large, so the input size of the target is set to 127*127, and the input size of the search area twice as large as the target template is set to 255*255. After convolution layer, the features of 6*6 and 22*22 are obtained respectively, and the correlation graph of 17*17 is obtained after full convolution. In order to save the debugging time of parameters, the training parameters of Simese-FC are used as the initial values of the training parameters of this network. After parameter debugging, the stochastic gradient descent method Stochastic Gradient Descent(SGD) method is finally determined to be

used for training. The quantitative evaluation criteria and evaluation parameters designed at this time are shown in Table 1 below.

Table 1. Quantitative evaluation criteria and evaluation parameters

Teacher/Student	Overlap	Accuracy	FPS
0034004	43.54	92.65	42.56
0034005	42.62	91.54	43.54
0034006	50.34	90.26	41.12
0034007	59.33	92.45	42.54
0034008	69.45	94.16	43.15
0034009	72.33	92.49	44.66
0034010	59.41	94.56	42.41
0034011	49.38	92.61	46.65
0034012	47.64	93.46	45.34

It can be seen from Table 1 that the accuracy and stability of Siam-MF and Siam-FC are the same for the same tracking video, and the real-time performance of Siam-MF algorithm is improved to a certain extent compared with Siam-FC. According to this evaluation standard and index, subsequent real-time target tracking experiments can be carried out.

3.2 Experimental Results and Discussion

In the preparation of the above experiment, we use the real-time tracking method of students' targets in Wushu distance education based on deep learning designed in this paper and the traditional real-time tracking method of students' targets to track, and record the delay of the two methods. The experimental results are shown in Table 2 below.

From Table 2, it can be seen that the designed real-time tracking method of students' targets in Wushu distance education has a short tracking delay, which proves that it has good tracking effect, effectiveness and certain application value.

On the basis of the above experiments, the methods of references [9] and [10] are introduced as comparison methods, and the tracking accuracy is taken as the index. The results are shown in the following Table 3:

From the analysis of Table 3, it can be seen that this method has the highest real-time tracking accuracy, with an average of 95.08%, while the real-time tracking accuracy of the methods in reference [9] and reference [10] are 90.02% and 88% respectively, which indicates that this method has more reliable real-time tracking results.

Table 2. Experimental results/ms

Tracking times	This paper designs a real-time tracking method for students' targets in Wushu distance education based on deep learning, tracking delay	Tracking delay of students' target real-time tracking method based on traditional Wushu distance teaching
1	0.56	1.65
2	0.44	1.34
3	0.43	1.69
4	0.39	1.47
5	0.61	1.33
6	0.25	1.48
7	0.44	1.94

Table 3. Tracking accuracy of different methods/%

Group	Methods of this paper	The method in reference [9]	The method in reference [10]
1	96.1	90.6	88.5
2	95.3	91.1	87.6
3	94.7	89.6	89.1
4	93.8	88.7	87.3
5	95.5	90.1	87.5

4 Conclusion

Wushu bears the heavy traditional Chinese culture, and schools are an important territory for inheriting Wushu culture. There are many problems in Wushu teaching. Yu Jing scholars believe that the weak teachers and single teaching content are the main problems in Wushu teaching today. The lack of teachers leads to the inability of our Wushu learning to achieve high-quality and high-level exercises, thus failing to satisfy the students' awareness of lifelong physical education; Li Benyi, a scholar, explored the dilemma of Wushu teaching from the perspective of human studies, that is, ignored the dominant position of human beings; In the process of Wushu teaching, it is necessary to meet students' individual needs, respect students' dominant position and pay attention to students' overall development. Therefore, it is necessary to build an online teaching platform to realize Wushu distance teaching. In order to ensure the effect of Wushu distance teaching, this paper designs a real-time tracking method of distance teaching objectives, and carries out experiments. The results show that the tracking time delay of the designed distance tracking method is short, which proves that its tracking effect is good and effective, and it has certain application value and can be used as a reference for subsequent Wushu teaching. However, there are few experimental data in the application

of this method, so it is not certain that it is suitable for a large number of college martial arts teaching. In the future research, it is necessary to conduct more extensive experiments and improve them to improve their applicability.

References

1. Feng, J., Zhao, H.: Dynamic nodes collaboration for target tracking in wireless sensor networks. IEEE Sens. J. 99, 1 (2021)
2. Gong, Y., Cui, C.: A measurement set partitioning algorithm based on CFSFDP for multiple extended target tracking in PHD Filter. Radioengineering, 30(2), 407–416 (2021)
3. Li, Z., Chen, X., Zha, Z.: Design of standoff cooperative target-tracking guidance laws for autonomous unmanned aerial vehicles. Math. Probl. Eng. **2021**(2), 1–14 (2021)
4. Wang, X., Xie, W., Luo, J., et al.: Labeled multi-bernoulli maneuvering target tracking algorithm via TSK iterative regression model. Chin. J. Electron. **31**(2), 227–239 (2022)
5. Li, S., Feng, X., Deng, Z., et al.: Minimum error entropy based multiple model estimation for multisensor hybrid uncertain target tracking systems. IET Signal Processing, 14(3) (2020)
6. Sun, C., Wan, Z., Huang, H., et al.: Intelligent target visual tracking and control strategy for open frame underwater vehicles. Robotica, 1–15 (2021)
7. Ma, C., Huang, J.B., Yang, X., et al.: Hierarchical convolutional features for visual tracking. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society (2015)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Sci. (2014)
9. Wang, L., Liu, T., Wang, G., et al.: Video tracking using learned hierarchical features. IEEE Trans. Image Process. **24**(4), 1424–1435 (2015)
10. Nam, H., Han, B.: Learming multi-domain convolutional neural networks for visual tracking (2015)
11. Tianhua, C., Siqun, Z., Yuxiao, L.: Semantic segmentation of remote sensing images based on improved deep neural network. Comput. Simul. **38**(12), 27–32 (2021)