SIAMESE NETWORK COMBINED WITH ATTENTION MECHANISM FOR OBJECT TRACKING

Danlu Zhang , Jingguo Lv*, Zhe Cheng, Yingqi Bai, Yifei Cao

Beijing University of Civil Engineering and Architecture, Beijing, China

KEY WORDS: Object Tracking, Deep Learning, Siamfc, Attention Mechanism, Siamese Network

ABSTRACT:

After the development of deep learning object tracking methods in recent years, the fully convolutional siamese network object tracking algorithm SiamFC has become a more classic deep learning object tracking algorithm. In view of the problem that the accuracy of the tracking results of SiamFC will be reduced in the case of complex backgrounds, this paper introduces the attention mechanism based on the SiamFC, which performs channel and spatial weighting on the feature maps obtained by convolution of the input image. At the same time, the backbone network model of CNN in the algorithm is adjusted, then the siamese network combined with attention mechanism for object tracking is proposed. It can strengthen the effectiveness of the results of feature extraction and enhance the ability of the network model to discriminate targets. In this paper, the algorithm is tested on the OTB2015, VOT2016 and VOT2017 datasets, and compared with multiple object tracking algorithms. Experimental results show that the algorithm in this paper can better solve the complex background problem in object tracking, and has certain advantages compared with other algorithms.

1. INTRODUCTION

In recent years, with the development of computer vision technology, visual object tracking technology has also developed rapidly. The so-called object tracking is to give the position and size of the object in the first frame of a given video sequence, and predict the position and size of the object in the subsequent video frame through an algorithm (Granström, Baum, 2017). Visual object tracking has always been an important research topic in the field of computer vision and one of the current research hotspots. Object tracking is widely used in many fields such as video automation monitoring, human-computer interaction, traffic monitoring, virtual reality, robot visual navigation and positioning, medical diagnosis, military applications, etc.

However, visual object tracking is actually a challenging task. In the tracking process, there are still a series of difficulties, such as the variability of moving target features, the scale change of target, the inconsistency of light intensity, occlusion, and the interference of complex backgrounds. These problems still have constraints on the performance and speed of the object tracking algorithm. Therefore, it is necessary to design a robust algorithm for object tracking.

With the rapid development of computer technology, more and more scholars at home and abroad have studied the moving object tracking in video, and have proposed many effective object tracking algorithms, which can be divided into three categories. The first type is the traditional object tracking algorithm, mainly including Kalman filter (Comaniciu et al., 2003), particle filter (Djuric et al., 2003), Meanshift algorithm (Cheng, 1995), Camshift algorithm (Allen et al., 2004) and optical flow method (Dowson, Bowden, 2007). The second type is the object tracking algorithm based on correlation filtering, mainly including MOSSE (Bolme et al, 2010), CSK (Henriques et al, 2012), KCF (Henriques et al, 2015), DSST (Danelljan et al, 2014), C-COT (Danelljan et al, 2016a) and ECO (Danelljan et al, 2017). The third category is based on deep learning object tracking algorithms, mainly including DLT (Wang, Yeung, 2013), FCNT (Wang et al, 2016), MDNet (Nam, Han, 2016), SINT (Tao et al, 2016), SiamFC (Bertinetto et al, 2016a), GOTURN (Held et al, 2016) and ATOM (Danelljan et al, 2019).

In the object tracking algorithm based on deep learning, SiamFC is a classic tracking algorithm. In this paper, aiming at the problem of poor tracking effect of SiamFC in complex background, the object tracking algorithm combined with attention mechanism is proposed in this paper to improve the performance of the tracking algorithm.

The main contributions of this paper are:

(1) A siamese network object tracking method combining spatial attention and channel attention is proposed, which increases the ability of the siamese network to discriminate the target, and improves the problem of poor tracking effect of SiamFC in complex backgrounds.

(2) Replace the CNN backbone network model in the siamese network object tracking algorithm from AlexNet to VGG, which increases the depth of the network and improves the algorithm's ability to express features of object.

* Corresponding author

(3) The algorithm is tested using multiple data sets and compared with various methods. The results show that the method in this paper has a certain degree of advancement.

2. ANALYSIS FOR SIAMFC

The object tracking algorithm based on the siamese network is first appeared in the SINT (Tao et al, 2016) algorithm in 2016. In the same year, Bertinetto et al. proposed the SiamFC (Bertinetto et al, 2016a) algorithm, which, like SINT, is also based on the siamese network, and the tracking problem is converted into a comparison problem of two images through the siamese network to solve the tracking problem. After the development of deep learning object tracking methods in recent years, SiamFC has become a classic deep learning object tracking algorithm.

SiamFC pioneered the application of the siamese network structure in the field of object tracking, significantly improving the tracking speed of the deep learning method, but it still has certain problems. According to the test results of the tracking effect of the SiamFC algorithm, it is found that the SiamFC algorithm will have a reduced tracking accuracy under the complex background, and may even cause tracking failure. Therefore, in the research work of this paper, the specific test and analysis of the SiamFC algorithm is carried out first, and then the problem is solved according to the conclusions drawn, and a more effective tracking algorithm is designed.

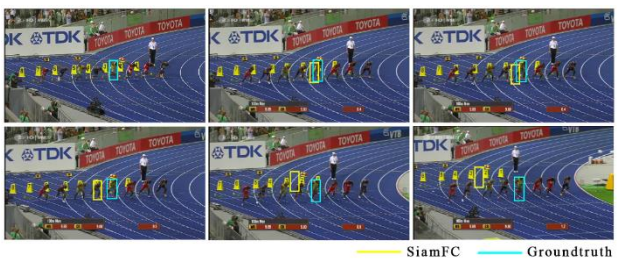

Figure 1. SiamFC algorithm tracking results in the first scene

In the research, the tracking results of the SiamFC algorithm in complex background tracking scenarios are tested. Among them, the background complexity can be defined as that the background near the tracking target has the color or texture similar to the target. Figure 1 and Figure 2 are the tracking results of two representative scenes, where the first picture of each figure is the first frame in the video. It can be clearly seen from the figure that when the background is complex, it is easy to cause interference to the tracking of the target.

Figure 2. SiamFC algorithm tracking results in the second scene

In the first scene, the background is more complicated, including not only the people around the target athlete, but also the track and the pass on the track. The tracked athlete is first offset from the athlete in the background. In the following tracking test, the tracking result is tracked in the background, completely deviating from the tracking target.

In the second scene, the back and forth running of the players on the basketball court also makes the background of the target tracking process relatively complex, and the tracking of the target is shifted to another player with similar clothing in the background.

It can be seen from the experimental results that in the case of complex background, using SiamFC algorithm for object tracking, the accuracy of tracking will be reduced. The main reason for this situation is that the SiamFC algorithm is an algorithm to judge the tracking target through similarity learning. The complex background with similar textures and colors around the target will interfere with the tracking process.

3. SIAMESE NETWORK OBJECT TRACKING ALGORITHM COMBINED WITH ATTENTION MECHANISM

In order to solve the problem of poor tracking effect of SiamFC algorithm in complex background, this paper introduces the attention mechanism to improve this algorithm. The attention mechanism enables the algorithm to focus on the original goal itself.

3.1 Algorithm Framework

Figure 3 shows the network structure of the proposed siamese network object tracking algorithm combined with attention mechanism. This network structure is improved on the basis of the SiamFC algorithm. The improvements are mainly in the following two aspects: 1) Embedded attention module, including channel attention module and spatial attention module; 2) The improved network structure replaces the CNN network backbone in SiamFC from the AlexNet to the VGG-16, increasing the depth of the network.

The basic principle of the improved algorithm is similar to the SiamFC algorithm, which uses similarity metric functions to determine the similarity of the target for the template image Z and the search image X. The similarity metric function here refers to a cross-correlation operation, that is, to use the feature map obtained by convolution of the template image Z to convolve the feature map obtained by convolution of the search image X. The feature image obtained by convolution of the template image Z is equivalent to the convolution kernel in the convolution process. Since the improved algorithm introduces the attention mechanism module based on the SiamFC algorithm, the calculation formula for calculating the similarity between the template image and the search image in the improved algorithm is:

$$f(z,x) = (\delta \odot \varphi(z)) * (\varepsilon \odot \varphi(x)) + b_1 \qquad (1)$$

where z = the value of a certain position on the template image Z

x = the value of a certain position on the search image X

φ= the function after CNN convolution operation

δ= the weight distribution of the template image Z obtained by the attention mechanism module

ε = the weight distribution of the search image X obtained through the attention mechanism module

b_1 = the value of each position in the score map

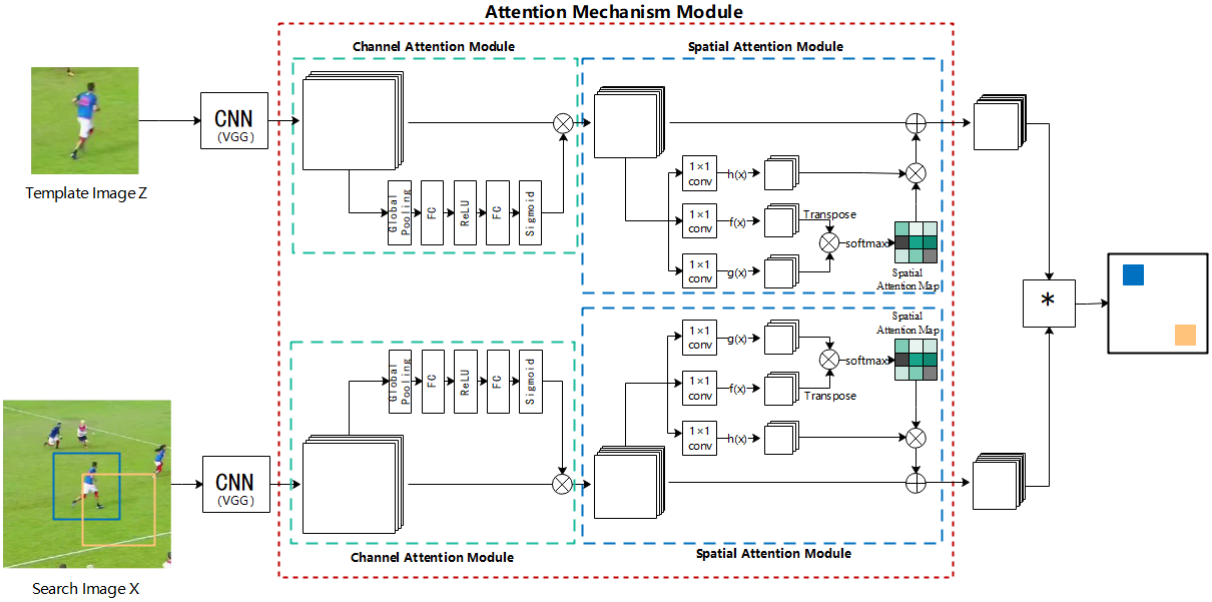

Figure 3. The network structure of improved algorithm based on SiamFC

The specific algorithm flow can be described as:

(1) Input template image Z and search image X. The size of the template image Z is 127 ×127 ×3, and the size of the search image X is 255 ×255 ×3.

(2) CNN convolution of template image Z and search image X respectively. This process is also the process of feature extraction. After convolution, the template image Z and the search image X respectively generate 6 ×6 ×512 and 22 ×22 ×512 feature maps.

(3) The extracted feature maps are sequentially weighted in channel and space. After extracting the features, the feature maps of the template image Z and the search image X are respectively input into an attention mechanism module, and the feature maps are first weighted on the channels to improve the feature expression ability between different channels, and then weighted on the space to highlight the importance of different locations.

(4) Calculate the response score map (score map). Perform cross-correlation operations on the features extracted from the template image Z and the search image X after channel and space weighting, respectively, and calculate and generate the response score map.

(5) Object tracking. When using this algorithm for object tracking, the search image centered on the previous frame of target position is generally used to calculate the response score map. Finally, the position with the largest score is multiplied by the step size to determine the current target position.

3.2 Attention Mechanism Module

The algorithm in this paper introduces the attention mechanism based on the SiamFC algorithm, and weights the channel and space of the features obtained by convolution of the input image. It strengthens the effectiveness of the results of feature extraction, enhances the discriminant analysis in the model, and improves the ability of the neural network model to discriminate against targets. The attention mechanism module includes channel attention module and spatial attention module.

3.2.1 Channel Attention Module

The introduction of the channel attention module on the basis of the SiamFC algorithm is mainly to allow the convolutional neural network to have a better adaptability to tracking the changes of the target's appearance semantics. The channel attention module can increase the proportion of feature channels related to the target, and reduce the proportion of other feature channels not related to the target. In this way, it can highlight the target that needs to be tracked. The channel attention module can also change the channel dependence between different channels. Each channel of the feature map obtained by the high convolution layer can be regarded as a response to a specific object category, and there is an interrelated relationship between the feature responses of different object categories. Therefore, the use of the interdependence between the feature maps of different channels can improve the ability to express the target features, for example, to enhance the interdependence between the feature maps of different channels.

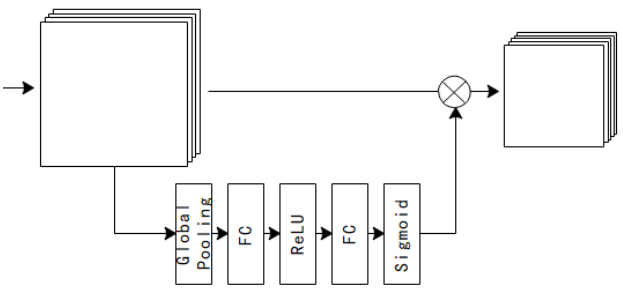

Figure 4. Channel attention module

The structure of the channel attention module is shown in Figure 4. First, the channel set of the input feature map is defined as:

$$A=[a_1, a_2, a_3, \dots, a_n] \quad (2)$$

where $a_k \in R^{H\times W}, k = 1,2,3,\dots,n_\circ$

Next, the input feature map A is global pooled, and the resulting feature vector is:

$$b = (b_1, b_2, b_3, \dots, b_n) \tag{3}$$

where $b_k \in R^{H \times W}, k = 1,2,3, \dots, n$

Then first pass the feature vector b through a fully connected layer FC, and then use the nonlinear activation function ReLU function to activate, so that the result has a nonlinear nature. After passing through the second fully connected layer FC, using the activation function sigmoid function to get the feature vector is:

$$\alpha = (\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_n) \tag{4}$$

where $\alpha_k \in R^{H \times W}, k = 1,2,3, \dots, n$

Finally, the obtained feature vector α is superimposed on the original feature map A, and the feature channels are rescaled to obtain the channel set of the finally generated channel attention feature map as:

$$\bar{A} = \alpha \cdot A = [\bar{a}_1, \bar{a}_2, \bar{a}_3, \dots, \bar{a}_n] \tag{5}$$

where $\bar{a}_k \in R^{H \times W}, k = 1,2,3, \dots, n$

3.2.2 Spatial Attention Module

The introduction of the spatial attention module on the basis of the SiamFC algorithm is mainly able to assign different weights to different spatial positions on the feature map, because different spatial positions have different importance in feature extraction. In the spatial attention module, the attention mechanism is introduced to establish the connection between any two positions in the feature map. For the feature of a certain position on the feature map, it can be calculated by weighting and summing the feature information of all positions on the feature map. Finally, the input feature and the spatial location feature are added to the element to further enhance the feature expression ability of the network model. Adding the spatial attention mechanism can increase the spatial position weight of important features to make the features more effective, and at the same time, it will not cause too much calculation and increase the calculation speed of the algorithm.

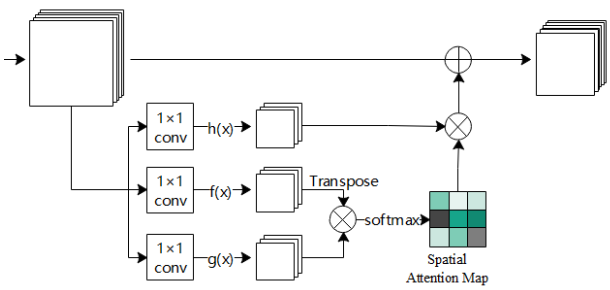

Figure 5. Spatial attention module

The structure of the spatial attention module is shown in Figure 5. First, the input feature map x is convolved using convolution kernels of size 1 × 1, respectively. Next, the three convolution results are converted using three transformation functions f (x), g (x), and h (x), respectively. Among them, the transformation functions f (x), g (x) and h (x) are:

$$f(x) = W_1 \cdot x, g(x) = W_2 \cdot x, h(x) = W_3 \cdot x \tag{6}$$

where W_1 = the weight of the function f (x)
W_2 = the weight of the function g (x)
W_3 = the weight of the function h (x)

Then, the result output by the function f (x) is transposed and matrix multiplied by the result output by g (x), and the obtained result can be calculated by using the softmax function to calculate the spatial attention map. The calculation formula of the spatial attention map is:

$$Y_{b,a} = \frac{e^{f(x_a)^T \cdot g(x_b)}}{\sum_{k=1}^{W \, H} e^{f(x_a)^T \cdot g(x_b))}} \tag{7}$$

where a = the a-th position on the input image
b = the b-th position on the input image

Then, the obtained spatial attention map and the function h (x) are subjected to matrix multiplication, and the obtained result is added to the input feature map x to calculate the feature map adjusted by the spatial attention module. The calculation formula of the final output is:

$$O_b = x_b + \beta \cdot (\sum_{a=1}^{W \, H} Y_{b,a} \cdot h(x_a)) \tag{8}$$

where x = the input feature map
β = the weight parameter

3.3 CNN Network Structure

In the SiamFC algorithm, the CNN backbone network structure used is AlexNet (Lecun, Bottou, 1998). In the algorithm of this paper, the backbone network used is the VGG-16 model (Simonyan, Zisserman, 2014) with deeper network layers, and some modifications have been made according to the algorithm.

The VGG-16 convolutional neural network model includes 16 layers (excluding the pooling layer), of which there are 13 convolutional layers and 3 fully connected layers. In the algorithm of this paper, the VGG-16 model has been modified to meet the needs of the algorithm. The main change is that the last three convolutional layers and the last three fully connected layers are removed, and the maxpooling layer before convolutional layer 4-1 is adjusted behind convolutional layer 4-1. Table 1 gives the specific CNN network structure parameters of the algorithm in this paper, including the size of the convolution kernel, stride of convolution, number of channels, template image size and search image size. The set CNN network structure contains 10 convolutional layers and 3 maxpooling layers, and no padding is used in the network. In addition to the last layer Conv4-3, each convolutional layer in the network uses the ReLU function for nonlinear activation. When training the network, batch normalization (BN) is performed after each convolutional layer.

4. EXPERIMENTS

4.1 Experimental Environment and Dataset

In the experiment, the operating system used was Linux (Ubuntu 16.04). During the experiment, CUDA was used for GPU acceleration. The GPU model is NVIDIA GeForce GTX 1060. The deep learning framework used in the experiment is pytorch, and the program implementation language is python.

In the experiment, the datasets used includes the training datasets and the test datasets. The training datasets used in the experiment are the Got-10k (Huang et al, 2018) dataset and the VID

(Russakovsky et al, 2014) dataset. The test datasets used in the experiment are OTB2015 (Wu et al, 2015), VOT2016 (Kristan et al, 2016) and VOT2017 (Kristan et al, 2017).

Layer	Stride	Size of Convolution Kernel	Number of Channels	Number of Feature Maps	Template Image	Search Image
Input layer	-	-	3	-	127×127	255×255
Conv1-1	1	3×3	96	96×3	125×125	253×253
Conv1-2	1	3×3	96	96×96	123×123	251×251
Maxpool1	2	3×3	96	-	61×61	125×125
Conv2-1	1	3×3	128	128×96	59×59	123×123
Conv2-2	1	3×3	128	128×128	57×57	121×121
Maxpool2	2	3×3	128	-	28×28	60×60
Conv3-1	1	3×3	256	256×128	26×26	58×58
Conv3-2	1	3×3	256	256×256	24×24	56×56
Conv3-3	1	3×3	256	256×256	22×22	54×54
Conv4-1	1	3×3	256	256×256	20×20	52×52
Maxpool3	2	2×2	256	-	10×10	26×26
Conv4-2	1	3×3	256	256×256	8×8	24×24
Conv4-3	1	3×3	512	512×256	6×6	22×22

Table 1. CNN network structure parameters used by the algorithm

4.2 Experiment Details

The training data needs to be preprocessed in the experiment. The siamese network structure requires training data to be image pairs, so the training data should be processed into image pairs (z, x). The template image Z and the search image X are both centered on the target and extracted from two frames of a video. The part beyond the image is filled with the RGB average value, and the target aspect ratio is kept unchanged. The specific category of the target is not considered during training, and the input image size of the network model is uniform.

The pre-trained model used in the experiment is the model trained on the ImageNet dataset. When training the network, use stochastic gradient descent (SGD) to train the network model. Among them, the momentum is set to 0.9; the decay mode of the learning rate is set to exponential decay, and the decay process starts from 10-2 to 10-8; the weight decay is set to 0.0005. The model was trained for 50 epoches, and the minimum number of mini-batch samples was 16.

Regarding the problem of scale transformation in the tracking process, the multi-scale test in SiamFC was followed in the experiment. In the multi-scale test, the target is tested in three scales, and the scale scaling factors are 1.025^{-1}, 0, and 1.025, respectively. Use these zoom factors on the image to be searched to search for the image.

4.3 Experimental Results

This paper tests the datasets and evaluates the proposed siamese network object tracking algorithm combined with attention mechanism. The training time in the experiment was about 26 hours. During testing, the average time required for each test data is 50 seconds. Figure 6, Figure 7, and Figure 8 are three representative experimental scenarios and their tracking results, where the first image in each image is the first frame of each video sequence, and the rest images are the tracking results of this algorithm and the tracking result of SiamFC.

Figure 6. Tracking result of the first scene

The first scene is a running competition. On the sports ground, there are multiple runners running on different tracks. One of the players is selected as the tracking target during tracking. During the running process, different players will stagger due to the difference in running speed. In the tracking process, because the tracking target person blocks the next person, the tracking result of the SiamFC algorithm will drift to the next person, resulting in tracking drift. The method proposed in this paper can focus on tracking the target.

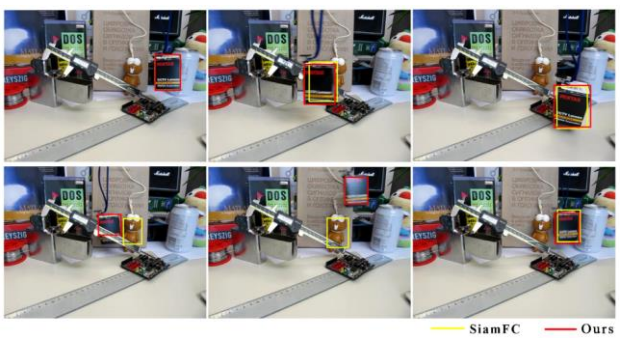

Figure 7. Tracking result of the second scene

In the second scene, there is a box that is controlled to move up, down, left, and right, and is in a place with a complicated background. During the movement, the box is partially blocked by surrounding objects. After occlusion, the tracking result of

SiamFC algorithm will drift to surrounding objects, and the method proposed in this paper can keep track of the target box.

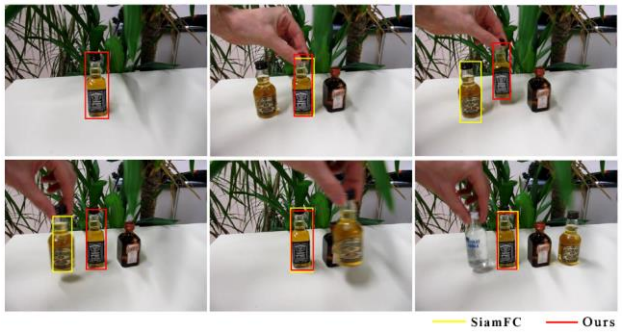

Figure 8. Tracking result of the third scene

In the third scene, various types of bottles place on the table are artificially added, and they are exchanged left and right respectively. In the process of bottle movement, other bottles will be blocked or blocked by other bottles. During the tracking process, the tracking results of the SiamFC algorithm will drift to non-target bottles, and the method proposed in this paper can keep track of the target bottles.

In summary, from the tracking results, it can be seen that when the complex background changes, the SiamFC algorithm may track other surrounding objects with semantic information, and the algorithm in this paper can focus on tracking the target itself, and the tracking effect is better.

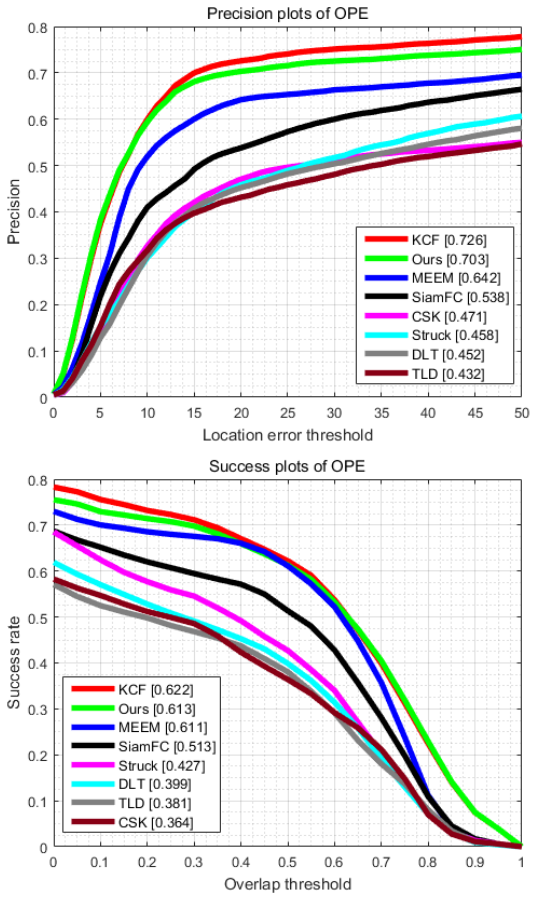

Figure 9. Precision plots and success plots of different trackers on OTB2015

4.4 Comparative Analysis

In order to further analyze the performance of the proposed algorithm to verify the effectiveness of the algorithm in this paper, a variety of methods were used on multiple test datasets and the tracking results of the method in this paper were compared.

4.4.1 Results on OTB2015 Dataset

Since each data on the OTB2015 dataset has attribute labels such as scale change, occlusion, and background clutters, 31 data with background clutters label are selected on the OTB2015 dataset for the experimental test. The experiment uses precision and success rate to evaluate the results, and the evaluation method is One-Pass Evaluation(OPE).

Figure 9 shows precision plots and success plots of our method and tracker SiamFC (Bertinetto et al, 2016a), Struck (Hare et al, 2016), KCF (Henriques et al, 2015), MEEM (Zhang et al, 2014), CSK (Henriques et al, 2012), DLT (Wang, Yeung, 2013), TLD (Kalal et al, 2013). In the precision plots, the algorithm in this paper ranks second, with the average precision of 0.703, which is 2.3% lower than the KCF algorithm, 6.1% higher than the MEEM algorithm, and 16.5% higher than the SiamFC algorithm. In the success plots, the algorithm in this paper is also ranked second, with the average success rate of 0.613, slightly lower than the KCF algorithm, which is 0.9% lower, which is higher than the success rate of algorithms such as MEEM algorithm and SiamFC.

4.4.2 Results on VOT2016 and VOT2017 datasets

On the VOT2016 and VOT2017 datasets, the method proposed in this paper is used for experimental testing. In the experiment, the three indexes of Accuracy, Robustness and Expected Average Overlap(EAO) are used to evaluate the results.

Trackers	VOT2016			VOT2017		
	A ↑	R ↓	EAO ↑	A ↑	R ↓	EAO ↑
DSST	0.53	-	0.181	-	-	-
MDNet	0.54	0.34	0.257	-	-	-
UPDT	-	-	-	0.53	0.18	0.378
Staple	0.54	0.38	0.295	0.52	0.69	0.169
SRDCF	0.54	0.42	0.250	0.49	0.97	0.120
CSRDCF	0.51	0.24	0.338	0.49	0.36	0.256
C-COT	0.54	0.24	0.331	0.49	0.32	0.267
ECO-HC	0.54	0.30	0.322	0.49	0.44	0.238
ECO	0.55	0.20	0.375	0.48	0.27	0.280
SiamFC	0.53	0.46	0.235	0.50	0.59	0.188
DensSiam	0.56	0.33	0.331	0.54	0.35	0.250
SiamRPN	0.56	0.26	0.344	0.49	0.46	0.244
SA-Siam	0.54	-	0.291	0.50	0.46	0.236
Ours	0.55	0.35	0.261	0.51	0.51	0.221

Table 2. Tracking results of different trackers on VOT2016 and VOT2017

Table 2 shows the comparison results of our method and tracker DSST (Danelljan et al, 2014), MDNet (Nam, Han, 2016), UPDT (Bhat et al, 2018), MEEM (Zhang et al, 2014), Staple (Bertinetto et al, 2016b), SRDCF (Danelljan et al, 2016b), CSRDCF

(Lukezic et al, 2017), C-COT (Danelljan et al, 2016a), ECO-HC (Danelljan et al, 2017), ECO (Danelljan et al, 2017), SiamFC (Bertinetto et al, 2016a), DensSiam (Abdelpakey et al, 2018), SiamRPN (Li et al, 2018), SA-Siam (He et al, 2018). In the VOT2016 dataset, the Accuracy of the algorithm in this paper is 0.55, ranking second, 1% less than the DensSiam and SiamRPN algorithms, the same as the ECO algorithm, 2% higher than the SiamFC algorithm. The Robustness value is 0.35, the robustness relatively poor, ranked lower, but more robust than the SiamFC algorithm. The EAO is 0.261, ranking in the middle of the 14 algorithms. In the VOT2017 dataset, the Accuracy of the algorithm in this paper is 0.51, ranking fourth, 3% less than the DensSiam algorithm, and 1% higher than the SiamFC algorithm. The Robustness value is 0.51, ranking relatively low, but its value is 8% lower than the SiamFC algorithm, and the robustness is better. The EAO is 0.221, ranking ninth.

5. CONCLUSION

Visual object tracking is to estimate the position of the target in the image sequence. It is a research hotspot in recent years and has been applied in many practical applications, such as automated monitoring, intelligent transportation, and robot positioning and navigation. Although research on object tracking has made great progress in recent years, object tracking is still a challenging task. In the process of object tracking, factors such as occlusion, background interference, target scale changes, and ambient light changes may affect the tracking results, and may even cause tracking failure. Therefore, in order to meet different practical application requirements, it is of great practical significance to study visual object tracking algorithms with higher accuracy and better effects. At present, with the popularity of deep learning methods, there are more and more object tracking algorithms based on deep learning, but there are still many problems to be solved in these methods.

Based on the above problems, through reading a large number of related papers at home and abroad, this paper learns the related theories based on the siamese network object tracking algorithm and introduces the attention mechanism to solve the background interference problem in siamese network object tracking. The proposed siamese network object tracking algorithm combined with attention mechanism not only improves the performance of the tracking algorithm, but also further solves the complex background problem in object tracking. The experiment carried out object tracking test on multiple datasets, and compared with multiple object tracking algorithms. Experimental results show that the algorithm in this paper can better solve the complex background problem in object tracking, and has certain advantages compared with other algorithms.

REFERENCES

Abdelpakey, M. H., Shehata, M. S., Mohamed, M. M., 2018. DensSiam: End-to-End Densely-Siamese Network with Self-Attention Model for Object Tracking. *arXiv preprint arXiv:1809.02714*.

Allen, J., Jin, J., Xu, Y., 2004. Object Tracking Using CamShift Algorithm and Multiple Quantized Feature Spaces. *Workshop on Visual Information Processing*.

Bertinetto, L., Valmadre, J., Henriques, J. F., et al, 2016a. Fully-convolutional Siamese Networks for Object Tracking. *European Conference on Computer Vision(ECCV)*.

Bertinetto, L., Valmadre, J., Golodetz, S., et al, 2016b. Staple: Complementary Learners for Real-Time Tracking. *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bhat, G., Johnander, J., Danelljan, M., et al, 2018. Unveiling the Power of Deep Tracking. *European Conference on Computer Vision(ECCV)*.

Bolme, D. S., Beveridge, J. R., Draper, B. A., et al, 2010. Visual Object Tracking Using Adaptive Correlation Filters. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Cheng, Y., 1995. Mean Shift, Mode Seeking, and Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8): 790-799.

Comaniciu, D., Ramesh, V., Meer, P., 2003. Kernel-based Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564-575.

Danelljan, M., Häger, G., Khan, F., et al, 2014. Accurate Scale Estimation for Robust Visual Tracking. *British Machine Vision Conference (BMVC)*.

Danelljan, M., Robinson, A., Khan, F. S., et al, 2016a. Beyond Correlation Filters: Learning Continuous Convolution Operators, for Visual Tracking. *European Conference on Computer Vision(ECCV)*.

Danelljan, M., Hger, G., Khan, F. S., et al, 2016b. Learning Spatially Regularized Correlation Filters for Visual Tracking. *IEEE International Conference on Computer Vision(ICCV)*.

Danelljan, M., Bhat, G., Khan, F. S., et al, 2017. ECO: Efficient Convolution Operators for Tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danelljan, M., Bhat, G., Khan, F., et al, 2019. ATOM: Accurate Tracking by Overlap Maximization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Djuric, P. M., Kotecha, J. H., Zhang, J., et al, 2003. Particle Filtering. *Signal Processing Magazine IEEE*, 20(5):19-38.

Dowson, N., Bowden, R., 2007. Mutual Information for Lucas-Kanade Tracking (Milk): An Inverse Compositional Formulation. *IEEE transactions on pattern analysis and machine intelligence*, 30(1):180-185.

Granström, K., Baum. M., 2017. Extended Object Tracking: Introduction, Overview and Applications. *Journal of Advances in Information Fusion*, 12(2).

Hare, S., Saffari, A., Torr, P. H., 2016. Struck: Structured Output Tracking with Kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10): 2096-2109.

He, A., Luo, C., Tian, Xi., et al, 2018. A Twofold Siamese Network for Real-Time Object Tracking. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Held, D., Thrun, S., Savarese, S., 2016. Learning to Track at 100 FPS with Deep Regression Networks. *European Conference on Computer Vision(ECCV)*.

Henriques, J. F., Caseiro, R., Martins, P., et al, 2012. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. *Proceedings of the 12th European conference on Computer Vision(ECCV)*.

Henriques, J. F., Caseiro, R., Martins, P., et al, 2015. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583-596.

Huang, L., Zhao, X., Huang, K., 2018. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *arXiv preprint arXiv:1810.11981*.

Kalal, Z., Mikolajczyk, K., Matas, J., 2011. Tracking-Learning-Detection. *IEEE Transactions on Software Engineering*, 34(7):1409-1422.

Kristan, M., Leonardis, A., Matas, J., et al, 2016. The Visual Object Tracking VOT2016 challenge results. *European Conference on Computer Vision(ECCV)*.

Kristan, M., Leonardis, A., Matas, J., et al, 2017. The Visual Object Tracking VOT2017 Challenge Results. *The IEEE International Conference on Computer Vision (ICCV)*, 1949-1972.

Lecun, Y., Bottou, L., 1998. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278-2324.

Li, B., Wu, W., Zhu, Z., et al, 2018. High Performance Visual Tracking with Siamese Region Proposal Network. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Lukezic, A., Vojir, T., Zajc, L. C., 2017. Discriminative Correlation Filter with Channel and Spatial Reliability. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Nam, H., Han, B., 2016. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Russakovsky, O., Deng, J., Su, H., 2014. ImageNet Large Scale Visual Recognition Challenge. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Tao, R., Gavves, E., Smeulders, A. W. M., 2016. Siamese Instance Search for Tracking. *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Wang, L., Ouyang, W., Wang, X., et al, 2016. Visual Tracking with Fully Convolutional Networks. *IEEE International Conference on Computer Vision (ICCV)*.

Wang, N., Yeung, D. Y., 2013. Learning a Deep Compact Image Representation for Visual Tracking. *Neural Information Processing Systems(NIPS)*, 809-817.

Wu, Y., Lim, J., Yang, M. H., 2015. Object Tracking Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848.

Zhang, J., Ma, S., Sclaroff, S., 2014. Meem: Robust tracking via multiple experts using entropy minimization. *European Conference on Computer Vision(ECCV)*, 188-203.