Survey of Object Tracking Algorithm Based on Siamese Network

Survey of Object Tracking Algorithm Based on Siamese Network

Mengle Zuo[1], Xuyang Zhu[1], Yongchao Chen[1], Junyang Yu[1,*]

[1]School of Software, Henan University, Kaifeng, China

[*]Email: jyyu@henu.edu.cn

Abstract. The network model using deep learning have been widely used in the sphereof visual object tracking for the past few years. The Siamese network can utilize the model based on deep learning to achieve a balance between the tracking accuracy and speed in the visual object tracking. This work mainly introduces the development process of the visual target tracking field and traditional target tracking algorithms. It focuses on the Siamese network structure and the improved the Siamese algorithm, and compares tracking results of related algorithms. Aiming at the deficiencies of existing Siamese object tracking algorithms, the future development trend is prospected.

1. Introduction

The object tracking is a important research direction of computer vision, and it has many applications scenarios, such as self-service robots, intelligent surveillance, autonomous driving and drone tracking. The object tracking uses the object location and scale provided by the first frame of the video sequence to accurately mark the location and scale of the object in later frames. For solving problems of object occlusion, similar background interference and object deformation in the tracking process, many excellent algorithms have been proposed.

The development of object tracking can be be illustrated in three stages. The first was around 2000. This period is mainly the application of classic algorithms and machine learning in target tracking. These algorithms have the characteristics of low computational complexity, fast running speed, and low requirements for hardware resources, but the algorithm's robustness and accuracy are relatively low. The second development stage of object tracking is from 2010 to 2016. With the MOSSE [1] tracker proposed, a large number of researchers began to pay attention to the tracker based on correlation filtering. Because of its fast speed and high accuracy, it has a good tracking effect in various evaluation data sets. The third development stage of target tracking is 2016 to present. The deep learning algorithm represented by the Siamese network has continuously improved the robustness and accuracy of the algorithm with more and more abundant data sets, proving its strong end-to-end learning ability in object tracking.

This paper divides object tracking algorithms into two categories: namely traditional algorithms and algorithms based on Siamese networks. Traditional algorithms include correlation filtering algorithms and correlation filtering algorithms with deep networks. The Siamese algorithm generally include both template

branch and search branch. It converts the tracking problem into a similarity matching, and can achieve a balance of tracking accuracy and speed.

2. Traditional object tracking algorithm

2.1. Correlation filtering tracking algorithm

The tracking algorithm utilizing correlation filtering can effectively update the weight of the filter to achieve online tracking by using the characteristics of the loop. The MOSSE tracker uses the target area of the first frame to generate related filters to adapt to the problem of illumination, lens zoom, object pose transformation and non-rigid deformation, and runs at a speed of 669 FPS. The CSK [2] tracker uses Fourier changes to complete the detection process, and utilizes the circulant matrix to handle the problem of dense sampling. The KCF [3] algorithm is an improvement of the CSK algorithm. It uses a circulant matrix to collect positive and negative samples and uses Fourier matrix transformation to improve the running speed. After that, the DSST [4] tracker uses the trained correlation filter to monitor the scale change, creating a method that combines translation and scale filtering.

2.2. Correlation filtering algorithm based on deep features

The correlation filtering algorithms based on depth features improve the robustness of the network model by using depth features. The combination of correlation filtering algorithm and depth feature effectively improves the accuracy of the algorithm. The C-COT [5] tracker points out that the DCF formula only focuses on single-resolution feature mapping, and introduces a new method for training continuous convolution filters on the basis of DCF. Afterwards, the ECO [6] tracker improved C-COT and solved the problem of slow speed.

After that, a series of methods based on deep network models were proposed, MDNet [7], TCNN [8], for instance. They use the powerful feature extraction capabilities of deep learning, so that the trained model does not need to be updated during the tracking process. Although these algorithms have achieved some good results, they cannot balance tracking accuracy and running efficiency. The Siamese tracking algorithm adapts the deep learning model as their feature extraction network, utilizes the cross-correlation similarity matching for the target template and the search region as well as training with more datasets. Siamese trackers can not only ensure thetracking speed, but also ensure that the tracking precision is stronger than that of correlation filter trackers.

3. The basic structure of the Siamese tracker

Usually Siamese tracker have a template branch and a search branch. The input of template branch takes the target picture in the first frame in the tracking sequence, and the input of search region branch takes the target picture of the subsequent frames in the tracking sequence. As shown in figure 1, the Siamese network structure of the two branches is the same and the parameters are shared. The Siamese tracking structure tries to find a set of parameters so that the similarity measure is the smallest among the same categories and the largest among different categories.

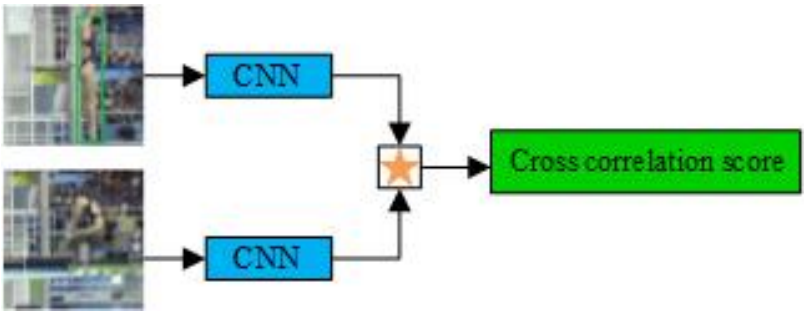

Figure 1. Basic construction of Siamese tracker

With the emergence of SINT [9] and SiamFC [10], the Siamese tracker has attracted great attention. The Siamese tracker learns a matching function and uses the depth features of the target region in the first frame to perform similarity matching with the search region features in subsequent frames. Since the network model is unchanged, even if the target is occluded or temporarily lost, it may be found in the follow-up tracking process.

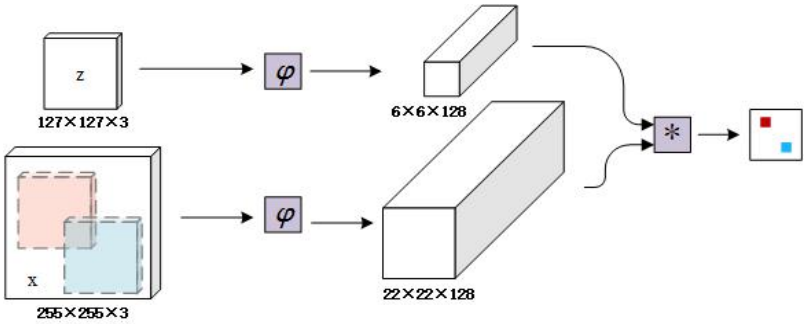

Figure 2. The network framework of SiamFC

The most representative Siamese structure tracker is SiamFC. The network of SiamFC is shown in figure 2, which has a template branch z and a search branch x. The target area of the first frame is the input z, and the search branch is the current frame picture. The networks of the two branches are the same and the parameters are shared. SiamFC uses deep features for similarity matching. The Siamese tracker performs the same transformation, and calculates the correlation according to the following formula.

$$f(z,x) = g(\varphi(z),\varphi(x)) \tag{1}$$

Where $\varphi(z)$ represents the depth feature of the template branch input image z, $\varphi(x)$ represents the depth feature of the search branch input image x, and the function g is the convolution operation on $\varphi(x)$ with $\varphi(z)$ as the kernel.

The depth features of the terget template and search rgion are subjected to a deep Similarity matching operation to get a cross-correlation response map. The more similar target template, the larger the value corresponding to the position in the response map.

4. Improved algorithm based on Siamese network

4.1. SiamRPN

After the SiamFC tracker, researchers proposed SiamRPN tracker based on SiamFC tracker. For the problem that SiamFC has no scale estimation, SiamRPN introduces the RPN in Faster R-CNN [11]. SiamRPN tracker detects the target scale with regression of bounding box to obtain the optimal bounding box. The network of SiamRPN is shown in figure 3. The network adapts end-to-end training, and treats tracking as part of detection task.

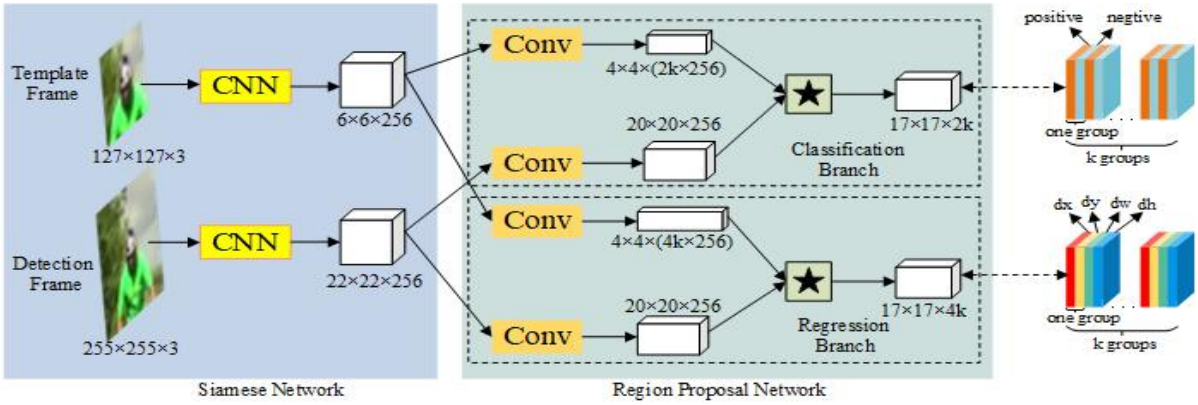

Figure 3. The network framework of SiamRPN

The framework consists both a Siamese network and a regional suggestion network, which are used for feature extraction and suggestion generation respectively. Among them, the Siamese network uses an unfilled full roll machine network. The regional proposal network contains two branches, namely the classification branch and the regression branch, they respectively used to prospect classification of target and background and regression of bounding box. The performance of improved SiamRPN in the VOT2016 dataset is shown in Table 1 and it runs at 160 FPS. The Accuracy rate of SiamRPN was the highest, 3.0% higher than that of C-COT in the VOT2016 dataset, and ECO-HC has the lowest EAO that is 0.322.

Table 1. Comparisons in the VOT2016 dataset

Tracker	EAO	Accuracy	Failure	EFO
ECO-HC	0.322	0.53	1.08	15.13
C-COT	0.331	0.53	0.85	0.507
SiamRPN	0.3441	0.56	1.08	23.3

4.2. SiamRPN++

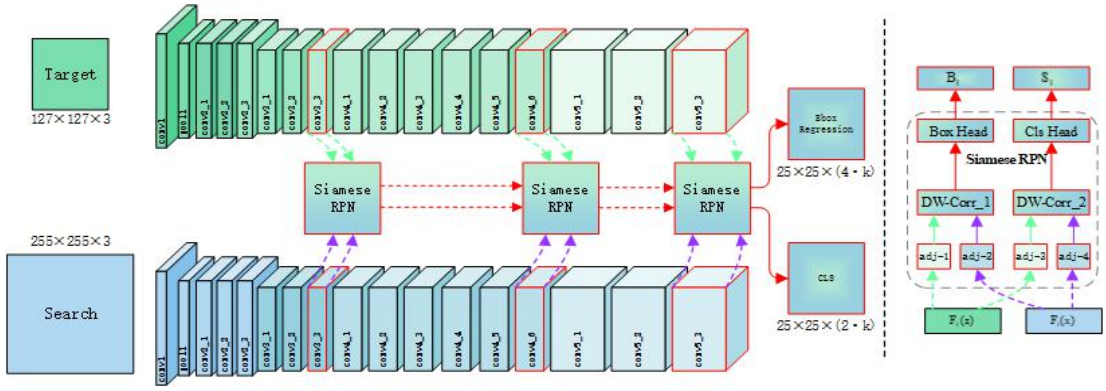

Figure 4. The network framework of SiamRPN++

SiamRPN++ [12] is a deep network model using multi-layer feature fusion obtained by improving the backbone network on the basis of SiamRPN. In order to make uniform sampling move around the image center, SiamRPN++ adopts a spatial sensing sampling strategy, and overcomes the problem of assigning greater weight to the center of the image during the feature learning process of the deep network, creating a backbone network for ResNet [13] Condition. The network structure of SiamRPN++ is shown in figure 4. For effectively using the different image features obtained by the different layers of ResNet convolution, SiamRPN++ uses a multi-level cascade method to complete the layer-by-layer aggregation of features, and achieves the improvement of target tracking performance. The performance of improved SiamRPN++ in

the VOT2018 dataset is shown in Table 2. Compared with SiamRPN, SiamRPN++ improves by 1.4% and 2.6% respectively for Accuracy and AO in VOT2018 dataset. But SiamRPN has the highest robustness that is 0.276.

Table 2. Comparisons in the VOT2018 dataset

Tracker	EAO	Accuracy	Robustness	AO
SA_Siam	0.337	0.566	0.258	0.429
SiamRPN	0.383	0.586	0.276	0.472
SiamRPN++	0.414	0.600	0.234	0.498

4.3. SiamCAR

For avoiding hyper-parameter adjustments in the regional proposal network, reduce the degree of human intervention. The SiamCAR [14] tracker solves the target tracking problem end-to-end in a pixel-by-pixel manner. SiamCAR decomposes the tracking task into both the classification of the pixel category and the regression of the bounding box of the object at the pixel. The network framework of SiamCAR is shown in figure 5.

Table 3. Comparisons in the GOT-10K dataset

Tracker	AO	$SR_{0.5}$	$SR_{0.75}$	FPS
SiamRPN_R18	0.483	0.581	0.270	97
SiamRPN++	0.517	0.616	0.325	50
SiamCAR	0.569	0.670	0.415	52

The SiamCAR tracker aims to extract a response map containing rich category information and semantic information, and directly predict the object location and the information of bounding box by using a unique response map, so it contains a Siamese subnetwork and a classification and regression subnetwork. The performance of improved SiamCAR in the GOT-10K dataset is shown in Table 3. SiamCAR trackers solve tracking problems from a pixel-level feature perspective, reducing the need for human intervention, enabling them to achieve higher accuracy and robustness to stay the ahead of performance. The average overlap rate of SiamCAR was 5.2% higher than that of SiamRPN++ in the GOT-10K dataset.

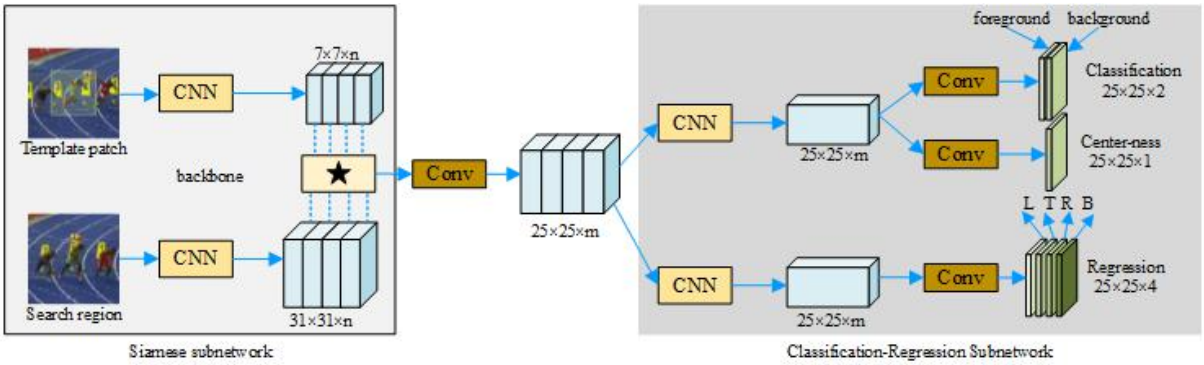

Figure 5. The network framework of SiamCAR

5. Future Prospect of Tracking Method Based on Siamese Network

Although the Siamese tracker can maintain the balance of tracking accuracy and efficiency, the Siamese structure still has some shortcomings. First, because the existing Siamese tracker does not have a model update mechanism, the tracker cannot obtain contextual information that is conducive to target positioning. Secondly, the Siamese tracker only uses the target template feature for training, ignoring the background

information that may solve the problem of similar target interference. Finally, for objects that do not appear in the training set, the learned similarity measurement function is not necessarily robust.

Regarding the shortcomings of the Siamese structure, the researchers solved some of the existing problems by introducing different mechanisms and achieved good results. However, in order to adapt to changes in goals, the Siamese structure should rely on an online update mechanism in the future, and ensuring high running efficiency. Because the update of the template will significantly increase the amount of calculation and reduce the tracking speed. At this time, studying how to adapt the model to target changes and speed up the tracking is an important topic in the future.

6. Conclusion

This work introduces generally the development process of the target tracking field and introduces the Siamese network framework and some improved Siamese network algorithms in detail. The Siamese trackers are based on the deep learning model, transforms the tracking problem into a similarity matching problem, and effectively balances the tracking accuracy and efficiency of the algorithm. As one of the important research directions for target tracking in the future, the Siamese tracker can achieve more accurate tracking by using end-to-end learning methods and using a large number of data sets for training.

Acknowledgments

This work was financially supported by the key Technologies R & D Program of Henan Province, China under Grant 212102210078.

References

[1] Bolme D S, Beveridge J R, Draper B A, et al 2010 Visual object tracking using adaptive correlation filters *2010 IEEE computer society conference on computer vision and pattern recognition* (IEEE) pp 2544-2550

[2] Henriques J F, Caseiro R, Martins P, et al 2012 Exploiting the circulant structure of tracking-by-detection with kernels *European conference on computer vision* (Springer, Berlin, Heidelberg) pp 702-715

[3] Henriques J F, Caseiro R, Martins P, et al 2014 High-speed tracking with kernelized correlation filters *IEEE transactions on pattern analysis and machine intelligence* **37(3)** pp 583-596

[4] Danelljan M, Häger G, Khan F, et al. 2014 Accurate scale estimation for robust visual tracking *British Machine Vision Conference* pp 1-5

[5] Danelljan M, Robinson A, Khan F S, et al 2016 Beyond correlation filters: Learning continuous convolution operators for visual tracking *European conference on computer vision* (Springer, Cham) pp 472-488

[6] Danelljan M, Bhat G, Shahbaz Khan F, et al 2017 Eco: Efficient convolution operators for tracking *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 6638-6646

[7] Nam H and Han B 2016 Learning multi-domain convolutional neural networks for visual tracking *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 4293-4302

[8] Nam H, Baek M and Han B 2016 Modeling and propagating cnns in a tree structure for visual tracking arXiv preprint arXiv:1608.07242

[9] Tao R, Gavves E and Smeulders A W M 2016 Siamese instance search for tracking *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 1420-1429

[10] Bertinetto L, Valmadre J, Henriques J F, et al. 2016 Fully-convolutional siamese networks for object tracking *European conference on computer vision* (Springer, Cham) pp 850-865

[11] Faster R 2015 Towards real-time object detection with region proposal networks *Advances in neural*

information processing systems p 9199

[12] Li B, Wu W, Wang Q, et al. 2019 Siamrpn++: Evolution of siamese visual tracking with very deep networks *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 4282-4291

[13] He K, Zhang X, Ren S, et al. 2016 Deep residual learning for image recognition *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 770-778

[14] Guo D, Wang J, Cui Y, et al. 2020 SiamCAR: Siamese fully convolutional classification and regression for visual tracking *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp 6269-6277