Object Tracking Algorithm Based on Channel-interconnection-spatial Attention Mechanism and Siamese Region Proposal Network

Junchang Zhang*
School of electronics and information, Northwestern Polytechnical University, Xi'an, People's Republic of China
zhangjc@nwpu.edu.cn

Siqi Lei
School of electronics and information, Northwestern Polytechnical University, Xi'an, People's Republic of China
siqi_lei@mail.nwpu.edu.cn

ABSTRACT

The target tracking algorithm based on the Siamese network has become one of the most mainstream and best tracking algorithms because of the balance of accuracy and speed. However, target tracking algorithms based on the Siamese network are affected by factors such as occlusion, illumination changes, motion changes, size changes and other factors in natural scenes, making designing a robust tracking algorithm a challenging task. In order to improve the feature extraction and discrimination capabilities of the algorithm in complex scenes, a tracking algorithm combining channel-interconnection-spatial attention mechanism was proposed. First a Siamese tracking framework with a deep convolutional network ResNet-50 as the backbone network was built to enhance feature extraction capabilities, then the channel-interconnection-spatial attention module was integrated to enhance the adaptability and discrimination capabilities of the model, then the multi-layer response maps were weighted and fused to make results more accurate, and finally the largescale datasets were used to train the network, and tracking tests on the benchmark OTB-2015 and VOT2016 and VOT2018 were completed. The experimental results show that the proposed algorithm is more robust and better adapt to complex scenes such as target appearance changes, similar distractors, and occlusion than the current mainstream.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**;

KEYWORDS

Siamese networks, Object tracking, Region proposal network, Channel attention, Spatial attention

ACM Reference Format:
Junchang Zhang* and Siqi Lei. 2021. Object Tracking Algorithm Based on Channel-interconnection-spatial Attention Mechanism and Siamese Region Proposal Network. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487075.3487120

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487120

1 INTRODUCTION

As one of the research directions of computer vision, object tracking has a wide range of applications in video surveillance, intelligent transportation, military guidance, aerospace and other fields. However, due to the targets' deformation, rotation, motion blur, and the influence of external application scene lighting changes, background interference, occlusion and other factors, it's still a challenge to establish an efficient and robust target tracking algorithm.

Currently, the mainstream methods in the object tracking field mainly fall into two categories, trackers based on correlation filter and trackers based on deep learning. The traditional correlation filter algorithm makes use of the characteristics of cyclic mutual, and carries out the calculation in the Fourier domain like MOSSE. With the rapid development of hardware technology and the advent of the era of big data, deep learning methods have gradually been recognized by people and have penetrated into many fields of computer vision, and target tracking is one of them. Outstanding feature expression capabilities of deep convolutional networks are used to improve the accuracy of the algorithm, such as MDNet [1], GFS-DCF [2].C-COT [3] learn the convolution operator in the continuous space domain to achieve accurate sub-pixel positioning. ECO [4] improve feature through factorization. In order to take advantage of the end-to-end advantages, researchers have begun to introduce the Siamese framework to train a dedicated end-to-end tracking network. SINT [5] pioneered the introduction of the Siamese network to transform the object tracking task into a similarity learning problem and learn a matching function. SiamRPN [6] is based on SiamFC [7] ,using regional proposal network in Faster R-CNN. In this way, bounding box regression can be used to replace multi-scale detection to obtain the bounding box with the maximum response. DasiamRPN [8] generated training positive sample pairs by augmenting static images and proposed disturbance recognition model and a local-to-global strategy for long-term tracking.

Although the end-to-end tracking algorithm depended on the Siamese network has achieved excellent results in general datasets, there are still some problems. The discrete training network only learns the universal characteristics of the object. If similar interferences appear around the object during online tracking, it cannot be judged. The object tracked online is probably not incorporated in the offline training dataset, so the result of resemblance measurement may not be reliable. We bring forward an target tracking algorithm, inheriting the characteristics based on Siamese network architecture. ResNet-50 is used as the backbone network to make the model stronger more expressive. In addition, we probe attention mechanisms, consisting of channel attention and position attention,

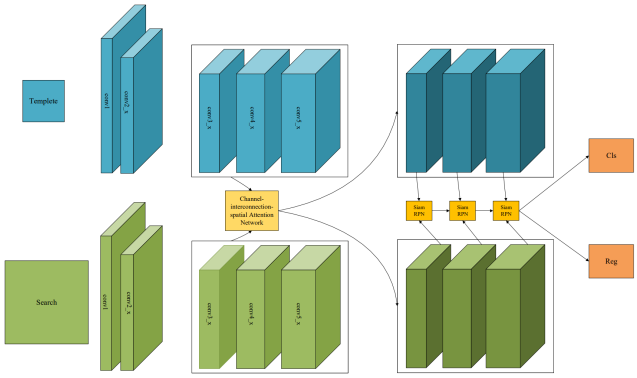

Figure 1: Framework of Algorithm.

so that offline trained model has better adaptability and discriminative ability when tracking online. Fusion of response score maps can make us obtain more accurate position. The network is trained offline in the large-scale data. We validate our algorithm performance in the target tracking general dataset. Compared with the current mainstream algorithms, our algorithm can achieve good tracking accuracy under the premise of ensuring real-time performance.

2 PROPOSED METHOD

The framework of algorithm is shown in Figure 1. It mainly includes Backbone network, Channel-interconnection-spatial attention module and Siamese RPN module. Among them, the Siamese deep network is based on RestNet-50 and integrates into CISAM. It engenders attentional features and then will be sent into RPN module to forecast a region using the classification branch and the regression branch. The tracking method is shown in Table 1

2.1 Backbone Network

He et al. [9] proposed the ResNet structure, which solved the problem of gradient disappearance in deep networks by adding identity mapping to the network, so that the network performance can be further improved when the network depth increases. The number of ResNet layers can be selected from 18, 34, 50, 101, 152, etc. Considering the feature representation capabilities and computational size and other factors, we selected the ResNet-50, shown in Figure 2

2.2 Channel-interconnection-spatial Attention Mechanism

It's demonstrated in Figure 3 that we input two sets of features calculated from the Siamese network. The CISA module incorporates self attention module and mutual attention module. We mark target features and search for images of Z and X, and object shapes for C × h × w and C × H × W.

The convolutional layer extracts features through sliding windows, so that the weights of all spatial positions of the image are the same. However, the features of the target area are more important than the features of nearby areas, so the pixels of the target area should be given larger the weight of. At the same time, the convolutional neural network generates the feature representation of the target through the semantic sub-features of different levels and spatial locations. These semantic sub-features are distributed in the feature vectors of different layers in the form of groups, and are easily interfered by similar features or backgrounds.

2.2.1 Self-attention. On the one hand, the channel attention mechanism [10] is used to assign different weights to the channels. On the other hand, the spatial attention mechanism is used to group the output characteristics of the channel attention network, and to select the semantic attributes of different channels and spatial positions, which can strengthen the response of the target area and suppress the response of the background area can also suppress the response of useless channels to greatly improve the robustness of tracking. Suppose the inputs are $\in R^{C\times H\times W}$, we separately employ two 1×1 convolution layers on them for emerging $Q \in R^{C'\times H\times W}$ and $K \in R^{C'\times H\times W}$ which are then reshaped to $\bar{Q}, \bar{K} \in R^{C'\times N}, N = H \times W, C' = \frac{1}{8}C$. a spatial attention map $A_S^S \in R^{N\times N}$ can be produced as,

$$A_S^S = softmax_{col}\left(\bar{Q}^T\bar{K}\right) \in R^{N\times N} \tag{1}$$

At the same time, the features X are through convolution layer and reshaped to obtain value feature$\bar{V} \in R^{C\times N}$. Then they are multiplied with the attention map and added to the reshaped features

Table 1: The Proposed Method

The proposed method
Input: pre-training, Initial frame object information
Step 1: Use the initial frame object information for target initialization
Cycle Step2-5 for frame=2, 3, . . . , N in the graphics sequence
Step 2: Use the Siamese backbone network Resnet-50 to obtain the t-th frame feature map
Step 3: Input feature map of conv3-5 into CISAM module, and use the formula (2), (4)to strengthen the feature expression
Step 4: Feed the output features of the conv3, conv4, and conv5 into three Siamese region proposal networks and obtain classification and regression
Step 5: Assign a certain weight to classification and regression by the three Siamese region proposal networks, and add them to get the final classification and regression
Output: The position and size of the current frame

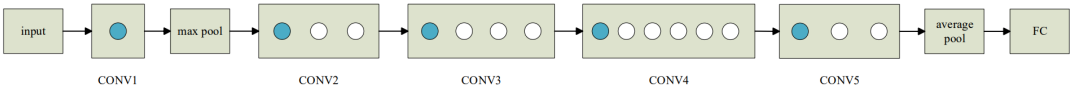

Figure 2: ResNet-50 Structure.

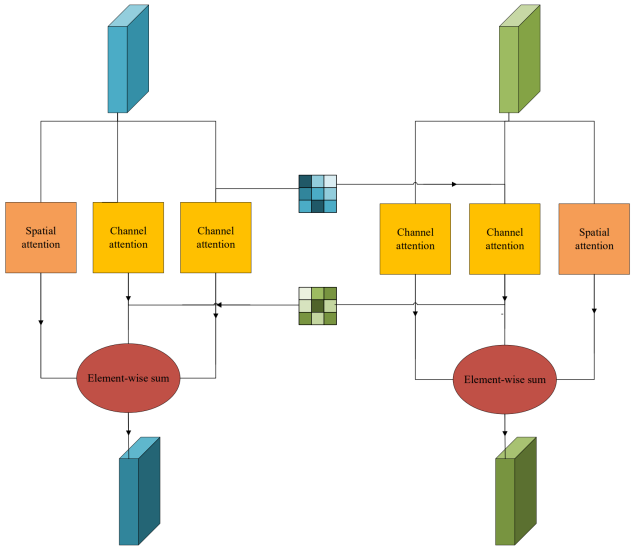

Figure 3: Channel-interconnection-spatial Attention Module.

$\bar{X} \in R^{C\times N}$ with a residual connection as,

$$\bar{X}_S^S = \alpha \bar{V} A_S^S + \bar{X} \in R^{C\times N} \tag{2}$$

2.2.2 Mutual-attention. The mutual attention sub-module is dedicated to changing the current situation of the lack of communication between the two branches of the Siamese network in the process of computing features. Generally speaking, the characteristics of the template branch and the search branch will not interact until they are interrelated. However, in the process of extracting features, for each branch, the information of the other branch is crucial. Especially for object tracking tasks, it's common that multiple similar objects appear close to each other at the same time, or even block each other. If the two branches of the Siamese network carry out effective information interaction in the process of calculating features, it will help each to capture more useful information. The mutual attention sub-module first calculates the attention feature map based on each branch's own information, and then transmits this feature map to another branch. The branch that receives the feature map enhances the features it has extracted based on this feature map and finally achieves more effective feature extraction. Suppose $Z \in R^{C\times h\times w}$ and $X \in R^{C\times H\times W}$ are template features and search features, respectively. Z are first reshaped to $\bar{Z} \in R^{C\times n}$, $n = h \times w$. The channel mutual-attention of the target branch,

$$A^C = softmax_{row}\left(\bar{Z}Z^T\right) \in R^{C\times C} \tag{3}$$

Then we can obtain the mutual-attention as,

$$\bar{X}^C = \gamma A^C \bar{X} + \bar{X} \in R^{C\times N} \tag{4}$$

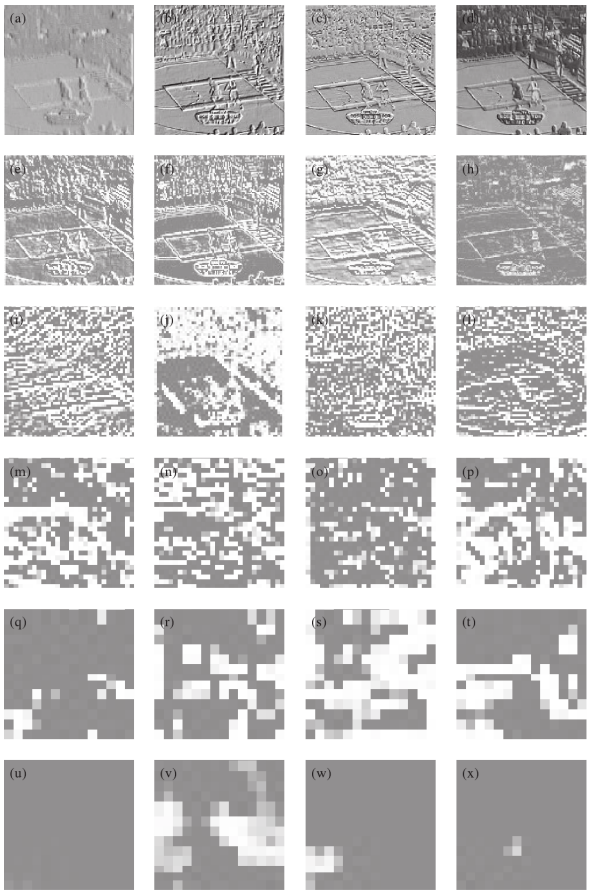

Figure 4: Feature Maps.

Finally, the attentional features can be produced by the combination of the self-attentional features X^S and the mutual-attentional features X^C.
Visualize part of the feature maps of different layers and channels extracted from it, as shown in Figure 4

2.3 Siamese Region Proposal Network

Target tracking requires not only shallow features such as color and shape, but also deep features rich in semantic information. Shallow features are conducive to the positioning of target tracking, but shallow features cannot effectively improve the positioning accuracy of regression branches due to lack of semantic information, while deep features rich in semantic information can effectively improve tracking performance in attributes such as motion changes. The multi-layer aggregation method uses three Siamese RPN blocks [11] and feeds the output features of the conv3, conv4, and conv5

of the target branch, and the output features of the conv3, conv4, and conv5 of the search branch into three Siamese region proposal networks. The output features first go through the 1×1 convolution operation, and then go through the deep cross-correlation operation, and then go through two convolution operations and finally calculate the classification score and the border regression score and perform a weighted summation.

3 TRAINING NETWORK

3.1 Training Data

The data in the training set is selected from ImageNet VID, COCO and Youtube-bb, combined with a specific ratio. Randomly select two frames from the same video sequence and combine them into a pair of template images (127×127) and search images (255×255) as the input of the Siamese network. Our experiment is implemented using PyTorch under the condition of NVIDIA GeForce RTX 2080Ti GPU.

To ensure that all input images are square, we adjust them according to the formula as,

$$s\,(w + 2p) \times s\,(h + 2p) = A$$

where (w, h) is the original target frame size, where p is the context margin, and where s is the scale transformation factor. $A = 127^2$.

3.2 Training Loss

We train the model in an end-to-end manner. The training loss is from Siamese RPN:

$$L = L_{rpn-cls} + \lambda L_{rpn-reg}$$

where $L_{rpn-cls}$ indicates classification loss and $L_{rpn-reg}$ represents regression loss in Siamese RPN.

3.3 Training Process

The learning rate during training is set to $10^{-3} \sim 10^{-6}$. The whole training process contains more than 100 stages, and each stage consists of 6,000 pairs of samples. We calculate the average loss value of 8 pairs of samples each time.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

4.1.1 OTB-2015 [12]. The OTB15 includes a total of 100 video sequences. In this dataset, there are a total of 11 challenge factors: scale transformation, occlusion, illumination change, deformation, low resolution, motion blur, similarity Interference, fast movement, beyond the field of view, rotation in the plane, rotation out of the plane. The evaluation criteria of the OTB15 dataset are tracking accuracy and success rate. Accuracy represents the Euclidean distance between the center of the predicted target frame and the center of the real frame. The greater the accuracy, the smaller the distance. The success rate represents the overlap rate between the predicted target frame and the real frame. The higher the success rate, the greater the overlap.

4.1.2 VOT2016 [13] &VOT2018 [14]. VOT2016 and VOT2018 have become one of the most mainstream datasets in the field of target tracking. The evaluation indicators include the average expected overlap rate (EAO), accuracy (Accuracy) and robustness (Robustness). The larger the average expected overlap rate and accuracy, the better the performance. The smaller the robustness value, the better the performance.

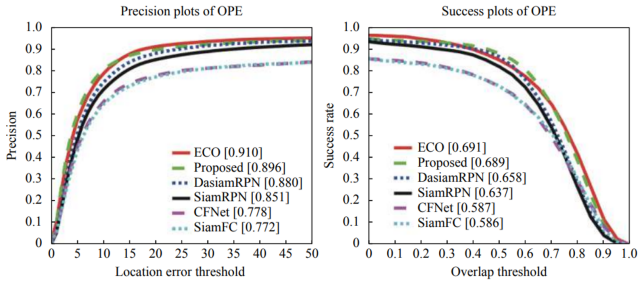

Figure 5: Success and Precision Plots on OTB2015.

Table 2: Results on VOT2016 and VOT2018

Tracker	VOT2016			VOT2018		
	A↑	R↓	EAO↑	A↑	R↓	EAO↑
SiamFC [7]	0.53	0.46	0.235	0.50	0.59	0.188
CFNet [15]	0,55	0.38	0.316	0.52	0.46	0.203
SiamRPN [6]	0.56	0.26	0.344	-	-	-
DaSiamRPN [8]	0.61	0.22	0.411	0.59	0.28	0.383
ECO [3]	0.55	0.20	0.375	0.28	0.28	0.276
SiamMask [16]	0.62	0.21	0.436	0.60	0.25	0.406
SiamMask_E [17]	0.65	0.21	0.452	0.63	0.25	0.427
Ours	0.63	0.18	0.437	0.62	0.21	0.412

4.2 Experimental Results

We compare our tracker with some state-of-the-art tracking trackers including SiamFC, CFNet, SiamRPN, DaSiamRPN, ECO, SiamMask and SiamMask_E. As shown in Figure 5, the comparison of the success rate and accuracy of the tracking results of different trackers on the OTB2015. We achieve a precision of 0.689 and an AUC of 0.896 which surpass that of SiamFC [7] by 10.3% and 12.4% respectively. It proves that the proposed algorithm extracts deeper network features based on the Siamese network framework and incorporates the attention module, so that the network can extract more adaptable features and improve the overall accuracy and robustness of the algorithm.

As shown in Table 2, the comparison of accuracy, robustness and expected of the tracking results of different trackers on VOT2016 and VOT2018. Our tracker performs well with 0.63 accuracy, 0.18 robustness and 0.437 EAO on VOT2016. Compared with recent SiamRPN [3] and DASiamRPN [4], our algorithm increase by 9.3% and 2.6% on EAO respectively. Our method ranks second in the EAO score on the VOT2016 and VOT2018 data sets, second only to SiamMask_E, illustrating the proposed Siamese attention and Region Proposal Network module work well.

4.3 Speed Analysis

On OTB-2015, VOT2016 and VOT2018, the average tracking speed under NVIDIA GeForce RTX 2080Ti GPU can reach 37 FPS, which can effectively track the object.

5 CONCLUSION

We have presented Channel-interconnection-spatial Attention Module consisting of both channel attention and spatial attention for visual object tracking. The new Siamese attention mechanism can strongly enhance object features and improve the robustness against occlusion, lighting changes, motion changes, size changes, and camera movement. Additionally, Siamese region proposal network is used to increase the accuracy. Through the experimental simulation and comparative analysis of multiple algorithms and the proposed algorithm, the results show that the proposed algorithm in this paper is better than most algorithms, and the performance is improved under multiple challenge attributes.

ACKNOWLEDGMENTS

First, I would like to thank my supervisor, Professor Zhang. Under his careful guidance, I have learned how to research, how to grasp the key points, and have a deeper understanding of the research direction. My professional knowledge and programming ability have been greatly improved. In the process of completing the paper, he provided me with a lot of constructive suggestions. I would also like to thank all my teachers who have helped me to develop the fundamental and essential academic competence. Last but not least, I' d like to thank all my friends, especially my three lovely roommates, for their encouragement and support.

REFERENCES

[1] Hyeonseob Nam and Bohyung Han (2016). Learning multi-domain convolutional neural networks for visual tracking. In CVPR.
[2] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler (2019). Joint group feature selection and discriminative filter learning for robust visual object tracking. In ICCV.
[3] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg (2016). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In ECCV.
[4] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg (2017). Eco: Efficient convolution operators for tracking. In CVPR.
[5] Tao Ran,Gavves E, Smeulders A W M (2016). Siamese instance search for tracking[C]//2016 IEEE Conference on Computer Vision Pattern Recognition, 1420-1429.
[6] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu (2018). High performance visual tracking with siamese region proposal network. In CVPR.
[7] Bertinetto L, Valmadre J, Henriques J F, *et al.* (2016). Fully-Convolutional Siamese Networks for Object Tracking[M]//Hua G, Jegou H. Computer Version ECCV 2016 Workshops. Cham:Springer, 9914: 850–865.
[8] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu (2018). Distractor-aware siamese networks for visual object tracking. In ECCV.
[9] He K M, Zhang X Y, Ren S Q *et al.* (2016) Deep residual learning for image recognition [C]|| CVP R 2016. Las Vegas: IEEE Computer Society, 770-778.
[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu (2019). Dual attention network for scene segmentation. In CVPR.
[11] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan (2019). Siamrpn++: Evolution of siamese visual tracking with very deep networks. In CVPR.
[12] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang (2013). Online object tracking: A benchmark. In CVPR.
[13] M Kristan, A Leonardis, J Matas, M Felsberg, R Pflugfelder, L Cehovin, T Voj ´ır, G Hager, A Luke ¨ zic, G Fern andez, *et al.* (2016). ´The visual object tracking vot2016 challenge results. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9914:777–823, 2016.
[14] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, *et al.* (2018). The sixth visual object tracking vot2018 challenge results. In ECCV.
[15] Valmadre J, Bertinetto L, Henriques J F, *et al.* (2017). End-to-end representation learning for correlation filter based tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 5000–5008.
[16] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr (2019). Fast online object tracking and segmentation: A unifying approach. In CVPR.
[17] Bao Xin Chen, John K. Tsotsos (2019). Fast Visual Object Tracking with Rotated Bounding Boxes. In ICCV.