ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0063

Combining Attention Mechanisms with RPN Networks for Target Tracking

Shitong Cao[1,a], Yulian Jiang[1,*]

[1]College of Electrical Engineering, Southwest Minzu University, Chendu, China.

[a]caoshitong@stu.swun.edu.cn

*Corresponding author

Abstract

In this paper, a twin network target tracking algorithm that combines spatial and channel attention mechanisms is proposed based on the fully convolutional twin network tracking algorithm (SiamFC). The attention mechanism takes into account the correlation between different positions in each layer and the correlation between different layers in the whole feature map. By combining the two attention mechanisms, not only can the global information be grasped, but also the target features can be highlighted, which can achieve a greater score for the subsequent inter-correlation operation and improve the accuracy and success rate of tracking. The RPN network structure was adopted to divide the network structure of SiamFC into two parts, classification and regression, to filter the target twice and improve the chance of finding the target accurately. Experiments on the OTB2015 and VOT2018 datasets yielded significant results.

Keywords

Target Tracking; Twin Network; Attention Mechanism; RPN Network.

1. Introduction

The development of computer vision has become a popular topic, and most of the vision deals with moving people and objects. scale shifts, fast motion, etc. A robust target tracking algorithm must be able to efficiently extract positive sample information for better subsequent tracking [1,2].

Prior to 2010, classical target tracking algorithms included Kalman Filter, Meanshift [3], Particle Filter [4], subspace learning, optical flow algorithms, and sparse representation methods. Correlation Filter (CF) based on target tracking algorithm achieved excellent performance and fast running speed. Afterwards, Henriques et al. improved the multi-channel function and kernel method based on CSK, resulting in the Kernelized Correlation Filter (KCF) tracking algorithm [5].

2. Related work

Deep learning methods can extract the deep semantic information of images and find the intrinsic connections among them. SiamFC [6] is a visual tracking algorithm based on a fully convolutional twin network, i.e., the image of the first frame and the current detection pin image go through the same feature extraction network to get the feature map; then the two feature maps obtained do inter-correlation operations to find the position with the greatest similarity can be mapped. Then the two feature maps are intercorrelated, and the position with the greatest similarity is mapped back to the original image, thus finding the tracking target. Since then, twin tracking algorithms based on SiamFC have been developed: SiamVGG [7] replaces the Alexnet feature extraction network in SiamFC with a VGG-16 network to obtain a more robust feature representation with a deeper network; Dsiam [8]

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0063

proposes a dynamic Siamese network to learn the appearance of the target online and suppress background information. SiamTri [9] introduces a triadic loss function into SiamFC to enhance the representational power of the model by guiding the loss function to better identify the difference between positive and negative samples; SiamRPN is a combination of twin networks for feature extraction and SiamRPN [10] combines the twin networks for feature extraction with the candidate region generation network, which consists of two branches, classification and regression, to discriminate the target and find the specific location of the target respectively.

The above twin network based on target tracking shows that although there are many improvements, the general idea is still to extract features from the first frame and the current frame. The first frame extracts more effective information to ensure accurate tracking, while the first frame is derived from a simple cropping operation of the image. The feature extraction of the first frame will extract the useless information indiscriminately, which is not only ineffective, but will also cause interference and affect the robustness of the subsequent correlation operation.

To address these problems, this paper proposes to integrate the spatial attention mechanism and the channel attention mechanism into the feature extraction network of the first frame, and to find the importance of between the pixel values at different positions in each layer of the feature map through this spatial attention mechanism, and the importance of between different layers of the feature map through the channel attention mechanism. In SiamFC only the first frame and the detection frame are inter-correlated after feature extraction to find the target. This one-time operation can easily cause misclassification. The introduction of the RPN network can turn one step into two, the first step finds the possible targets, and the second step performs another inter-correlation operation on the possible targets, thus finally finding the location with the maximum target response.

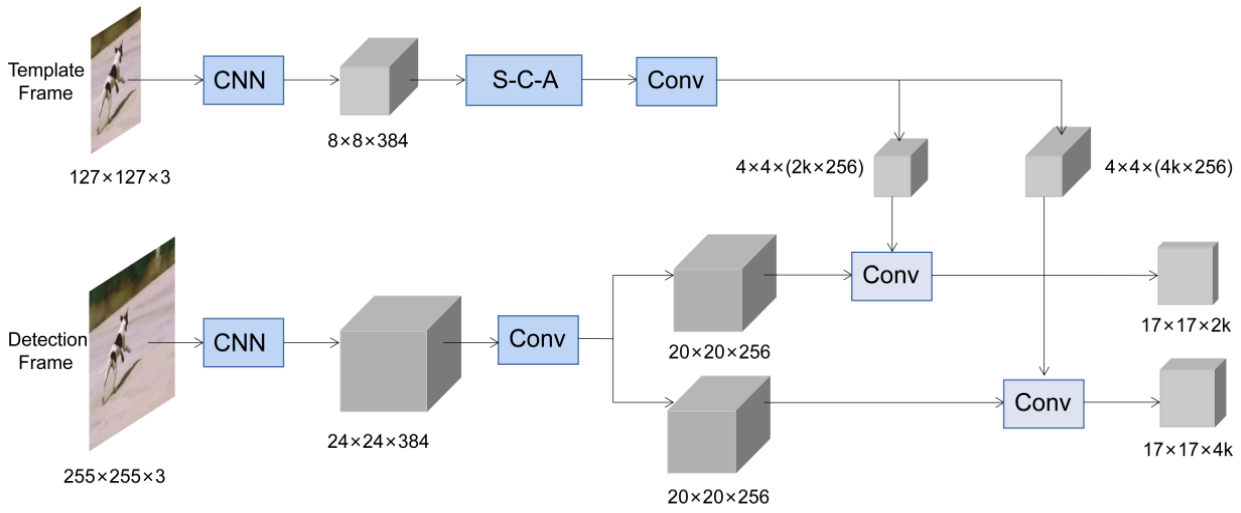


Figure 1. Overall network architecture

3. Overview of the algorithm

3.1 Overall network model

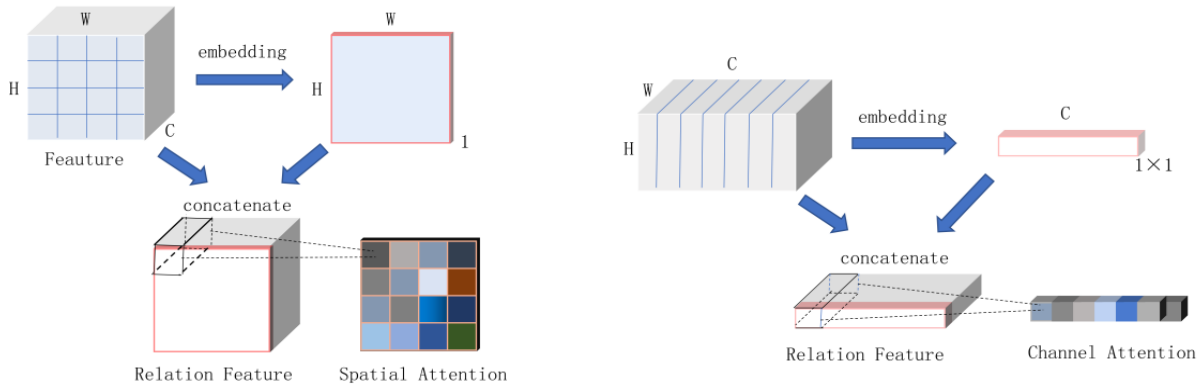
The overall framework of the algorithm is shown in Figure 1. The overall network model is constructed based on a twin network, and the model as a whole has two major branches, which do feature extraction for the current frame and the detection frame respectively. The first branch contains S-C-A (Spatial Attention and Channel Attention) [11], which is an operation that combines spatial and channel attention mechanisms. As can be seen in the figure below, the first frame is obtained by framing a 127 × 127 rectangle, which is not square, so that framing the target must also introduce

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0063

background information. In this paper, the spatial and channel attention mechanisms are added to the model design to find the information contained in the target in the feature map, while suppressing the background information around the target. After obtaining the processed feature map, the model goes through an RPN network containing a classification branch and a regression branch. The classification branch is designed to find regions that may contain the target, and to perform a correlation operation between the filtered possible regions and the feature map obtained in the first frame. The regression branch is designed to better determine the exact location of the target information and map the target information on the feature map back to the original map.

3.2 Spatial and channel attention mechanisms

The attention mechanism is essentially the addition of a set of weights by fusing deep and shallow information so that the fused and compressed information has "experience" and is more likely to be useful. As in figure (a)(b), the feature map is obtained by convolutional pooling, and the attention mechanism is to overlay the weight parameters on different positions of the feature map to bring out the target information. The updated feature map is, where denotes the weight values.



(a) Spatial attention mechanisms (b) Channel attention mechanism

Figure 2. Structure of the spatial and channel attention mechanism model

As in figure (a), the feature map is obtained with height H, width W and number of channels C. To obtain an attention map of size H × W × 1 that covers every point in the feature map space, the embedding operation is: which contains the use of 1 × 1 convolution, batch normalization (BN) and the Relu activation function [12]. concatenate is the splicing operation, i.e., finally, the sum is obtained after the convolution and batch normalization operations. Finally different colours in Spatial Attention indicate different weighting parameters.

As shown in figure (b), the channel attention mechanism is based on the same principle as the spatial attention mechanism, which is the attention information formed by integrating the spatial information, as shown in figure (b). The same embedding and concatenate operations are used to obtain the channel attention weight, and by adding this weight information to different channels, the channel attention mechanism is formed, so that the information in the important channels ,the information in the important channels can be brought into play.

3.3 Network structure

The feature extraction network is shown in Figure 3. The overall network is based on the Alexnet network, with a simple network structure and a lightweight model. Batch normalization is added to it, thus constraining the learning of the model. Spatial and attention mechanisms are added to the fourth convolutional layer, which has experience with the learning of the model and can guide subsequent learning to extract key information.

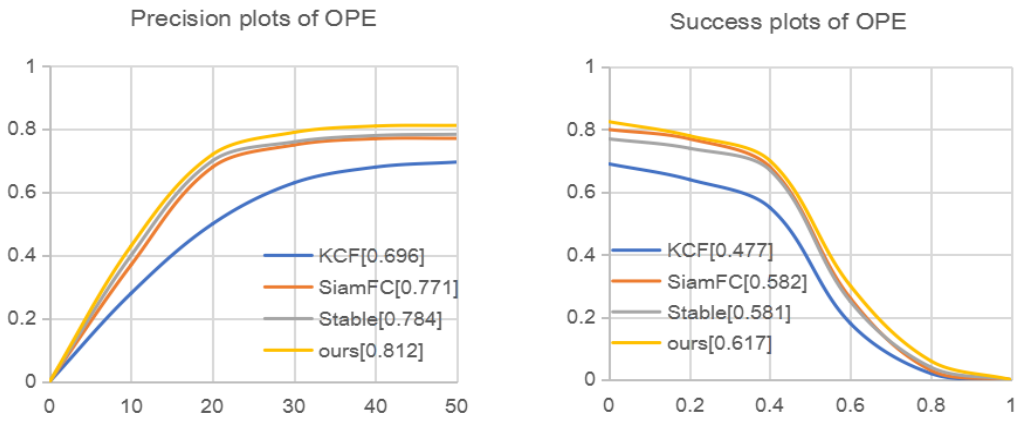
ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0063

Table 1. Network model parameters configuration

Layer	Support	Stride	for exemplar	for search	chans
data			127×127	255×255	×3
conv1	11×11	2	59×59	123×123	×96
bn + Relu			59×59	123×123	×96
pool1	3×3	2	29×29	61×61	×96
conv2	5×5	1	25×25	57×57	×256
bn + Relu			25×25	57×57	×256
pool2	3×3	2	12×12	28×28	×256
conv3	3×3	1	10×10	26×26	×192
bn + Relu			10×10	26×26	×192
conv4	3×3	1	8×8	24×24	×192
bn +Relu			8×8	24×24	×192
S-C-A			8×8	24×24	×192
conv5	3×3	1	6×6	22×22	×128
bn + Relu			6×6	22×22	×128
conv6	3×3	1	4×4	20×20	×256

4. Experimental comparison

In this paper, experiments were done on the OTB100 and VOT2017 datasets, both of which are authoritative public datasets for single-objective training tests. The OTB100 dataset contains 100 video sequences and VOT2018 contains 60 video sequences. Both video datasets contain attributes such as scale variation, fast motion, illumination variation, occlusion, deformation, motion blur, background blur, and low pixel. The tracking model designed in this paper is compared with KCF, ECOhc, SiamFC, and cfnet. For the OTB100 dataset, two evaluation criteria, centre error and area overlap, are selected; for the VOT2018 dataset, accuracy (Accuracy, A), average expected overlap rate (Expect Average Overlap Rate EAO), and tracking speed (Frames Per Second, fps). After experiments, this paper proved to be more competitive in terms of tracking performance.



Location error threshold

(a) Accuracy curve on OTB100

Overlap thresholds

(b) Success rate curve on OTB100

Figure 3. Tracking accuracy and success curves of different models on the OTB100 dataset.

Table 2. Experimental results of different models on the VOT2018 dataset.

Trackers	Accuracy	EAO	Speed (fps)
SiamFC	0.494	0.187	31.9
Stable	0.530	0.169	13.5
ours	0.546	0.278	18.6

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0063

After experimenting with the results, the proposed algorithm was tested on the OTB100 dataset and yielded a 4% improvement in accuracy and a 3.5% improvement in success rate compared to the SiamFC algorithm. Testing on the VOT2018 dataset yielded a 5.2% improvement in accuracy and a 9.1% improvement in the expected average overlap rate compared to SiamFC. Also, it is equally competitive with other algorithms.

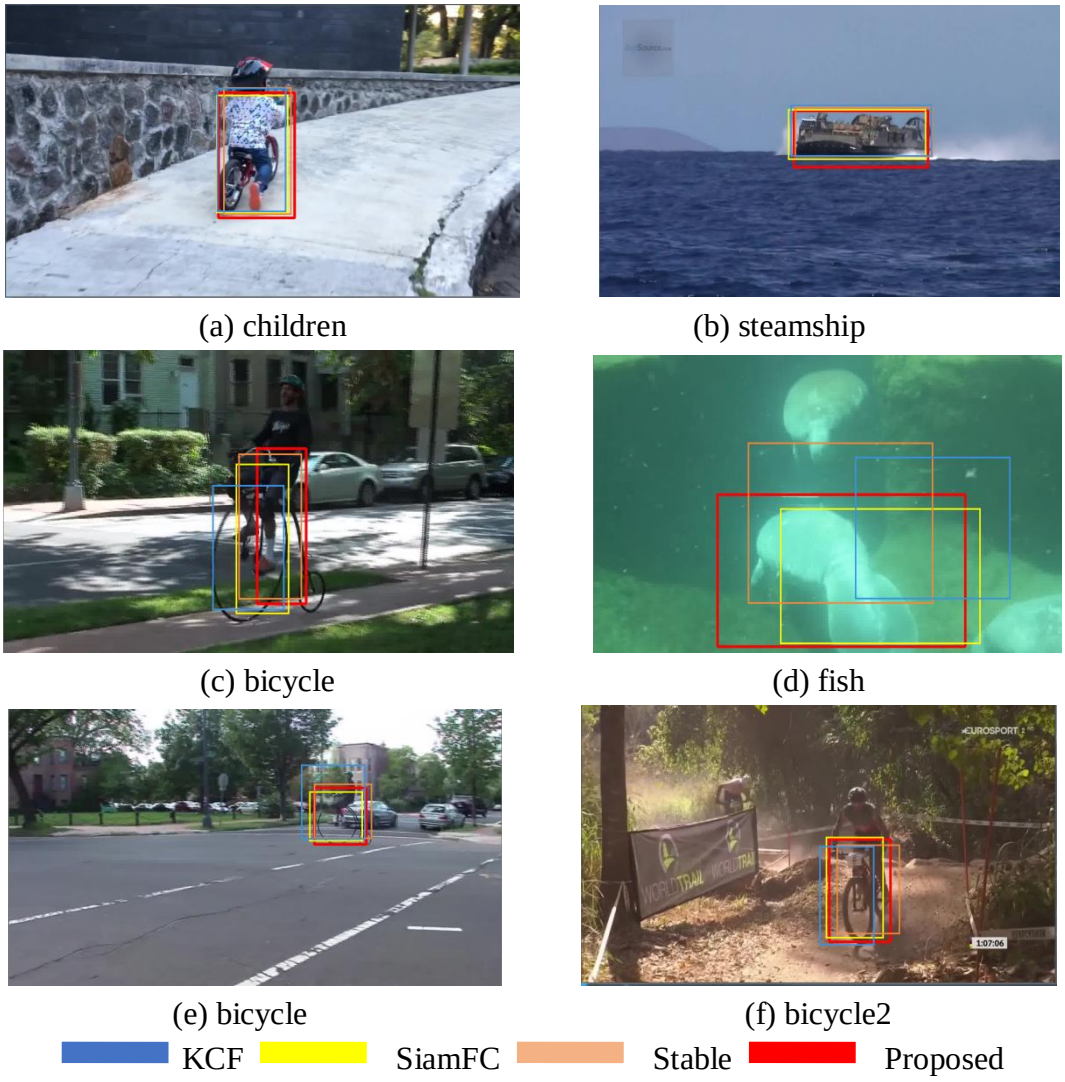


Figure 4. Qualitative results of each tracking algorithm on OTB100 dataset

In (a) children and (b) steamship, the target is clear and moves slowly, and all four algorithms track well at this point; in (c) bicycle and (d) fish, the target is in a complex background, the target is similar to the background, KCF has difficulty tracking the target, SiamFC and Stable drift, and the algorithms in this paper, which fuse attention mechanisms with RPN. In (e) bicycle and (f) bicycle2, the target is moving fast and scale shift occurs, and the algorithm in this paper is relatively more accurate in tracking.

5. Conclusion

This paper designs a target tracking algorithm based on a fully convolutional twin network that incorporates spatial and channel correlation attention mechanisms to achieve focused extraction of target feature information, suppress background information, reduce the chance of model error judgement, and combine with RPN networks to enhance the model's judgement and search for targets, thus effectively improving the accuracy of model tracking. The algorithm has been proven to have

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0063

higher tracking accuracy and robustness than SiamFC algorithm. In the future, we will continue to explore the design and optimisation of the network model to improve the accuracy and speed of tracking.

Acknowledgments

The work of this paper is supported by the Southwest Minzu University Graduate Innovative Research Project (Master Program CX2020SZ100). A special acknowledgment should be given to Southwest Minzu University for its experimental conditions and technical support.

References

[1] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in Proceedings - IEEE INFOCOM, 2000, vol. 2, pp. 775–784. doi: 10.1109/infcom. 2000.832252.

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," IEEE Transactions on Signal Processing, vol. 50, no. 2, pp. 174–188, Feb. 2002, doi: 10.1109/78.978374.

[3] L. Dawei, H. Qingming, J. Shuqiang, Y. Hongxun, and G. Wen, "Mean-shift blob tracking with adaptive feature selection and scale adaptation," in Proceedings - International Conference on Image Processing, ICIP, 2006, vol. 3. doi: 10.1109/ICIP.2007.4379323.

[4] C. Chang and R. Ansari, "Kernel particle filter for visual tracking," IEEE Signal Processing Letters, vol. 12, no. 3, pp. 242–245, Mar. 2005, doi: 10.1109/LSP.2004.842254.

[5] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 583–596, 2015, doi: 10.1109/TPAMI.2014.2345390.

[6] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9914 LNCS, pp. 850–865, 2016, doi: 10.1007/978-3-319-48881-3_56.

[7] Y. Li and X. Zhang, "SiamVGG: Visual Tracking using Deeper Siamese Networks," Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.02804.

[8] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning Dynamic Siamese Network for Visual Object Tracking."

[9] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11217 LNCS, pp. 472–488, 2018, doi: 10.1007/978-3-030-01261-8_28.

[10] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8971–8980, 2018, doi: 10.1109/CVPR. 2018.00935.

[11] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-Aware Global Attention for Person Re-identification," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.02998.

[12] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," Feb. 2015, [Online]. Available: http://arxiv.org/abs/1502.03167.