A Siamese Network Tracking Algorithm Based on Hierarchical Attention Mechanism

Hu Zhang¹,², Dong Hu¹,²,³,⁴,* and Yingcan Qiu¹,⁴

¹ Jiangsu Province’s Key Lab of Image Procession and Image Communications, Nanjing, 210003, China
² Education Ministry’s Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing, 210003, China
³ Education Ministry’s Engineering Research Center of Ubiquitous Network and Health Service, Nanjing, 210003, China
⁴ Nanjing University of Posts and Telecommunications, Nanjing, 210003, China

*Email: hud@njupt.edu.cn

Abstract. A siamese network tracking algorithm based on hierarchical attention mechanism is proposed in this paper. In order to obtain more robust target tracking results, different layer features are fused effectively. In the process of extracting features, attention mechanism is used to recalibrate the feature map, and AdaBoost algorithm is used to weight the target feature map, which improves the reliability of the response map. Besides, the Inception module is also introduced which not only increases the width of the network and the adaptability of the siamese network to the scale, but also reduces the parameters and improves the speed of network training. Experimental results show that this method can effectively solve the impact of background clutter and improve the accuracy of tracking.

1. Introduction

The task of target tracking is to locate the target in the initial frame of a given video sequence, then locate the target in the subsequent video frames and predict the size of the target. Target tracking has a wide range of prospects in many fields, such as human machine interaction, military reconnaissance, unmanned driving and security.

Since 2010, correlation filters have been introduced into the detection-based target framework with good results. Correlation filters use cyclic convolution to train filters to distinguish the appearance of objects and backgrounds. Minimum Output Sum of Squared Error (MOSSE), proposed by Bolme et al. [1], first applied related ideas to target tracking. The correlation filter can make the target tracking real-time because of its fast calculation speed, but the accuracy of its tracking results needs to be improved [2]. With the rise of deep learning, target tracking, as an important research field in computer vision, has also been widely used in deep learning algorithms. When in-depth learning is first applied, only the features obtained from in-depth learning are incorporated into the training of related filters to obtain more robust tracking results [3]. Danelljan et al. [4] proposed the DeepSRDCF method, which uses the features of convolution layer output to replace the traditional feature representation. Because these features are learned, they are more representational than the traditional HOG and CNs features, but at the same time, they also bring about an increase in computational effort. In 2016, Luca Bertinetto et al. [5] proposed the siamese network framework, which is a multi-branch,
weight-sharing network structure, mainly used to calculate image similarity. By extracting depth semantic information using convolution neural network, the similarity between target and template can be better distinguished. Ma et al. [6] proposed a Hierarchical Convolutional Features by analyzing the network characteristics of VGG-19. The deep features reflect the semantic information of the target, and is more robust to the target, while the shallow features reflect the details of the target, and locates the target more accurately.

Traditional Siamese network frameworks use single-layer features, which are not sufficient for accurate tracking in complex environments. Based on this, a Siamese network tracking algorithm based on hierarchical attention mechanism is proposed. Find the importance of each channel by introducing a layer of attention and then recalibrate them to enhance useful features and suppress target-independent features based on their importance. Using the fusion layer combines the deep features with the shallow features, which enables the network to extract more detailed information about the target, and is more robust to complex situations such as target occlusion and illumination changes [7]. Then the feature response maps of different layers of the network are weighted by the AdaBoost algorithm, which makes the larger error get smaller weight and the smaller error get larger weight, which makes the stronger judgment ability in the filter play a key role [8].

This is generally how articles are organized. Section 2 gives a brief review of Siamese neural networks. In the third section, we enter the core content of this article, which is the Siamese network we designed based on the hierarchical attention mechanism. The fourth section discusses the experimental simulation. Finally, the final summary is provided in the fifth section.

2. Siamese Neural Network

A siamese network has two inputs, one is a template and the other is a search area. These two inputs enter two neural networks, map the inputs to a new space, form a new spatial representation, and then learn a similarity function., calculate the similarity between the template and the same size area in the search area, and then find the candidate block with the greatest similarity as our target location on the new picture. A generic template block is an image frame using the first frame. The advantage of full convolution neural network is that the image of the search area does not need to be the same size as the template image, and it can increase the size of the search area image as input. The formula for the similarity function of this network is expressed as:

\[ f(z, x) = \varphi(z) * \varphi(x) + b_1 \] (1)

Where \( \varphi(z) \) is considered a convolution kernel and convolutes on \( \varphi(z) \), \( b_1 \) is the value at each location in the score graph.

We train positive and negative samples of the searched image using discriminant methods, and the target logical loss is defined as:

\[ l(y, v) = \log(1 + \exp(-y * v)) \] (2)

Where \( y \) represents the true value and the range is \((-1, 1)\), \( v \) represents the actual score of the sample-search image, the probability of positive samples is \( (1 + e^{-y})^{-1} \), and the probability of negative samples is \( 1 - (1 + e^{-y})^{-1} \), then the average logical loss at all locations during target training can be obtained by the definition of cross-entropy:

\[ L(y, v) = \frac{1}{D} \sum_{u \in D} l(y[u], v[u]) \] (3)

Where \( D \) represents the final score map and \( u \) represents all locations in the score map.

The convolution parameter \( \theta \) of the training is minimized by a random gradient descent method to obtain an optimal solution:
arg min \theta \frac{E(L(y, f(z, x; \theta)))}{E(L(y, f(z, x; \theta)))} = \frac{E(L(y, f(z, x; \theta)))}{E(L(y, f(z, x; \theta)))} \tag{4}

3. Siamese Network Based on Hierarchical Attention Mechanism

In this part, we make a detailed explanation of the algorithm we propose. This section will be divided into three modules. The first module introduces the application of attention mechanism. The second module will analyze the features of layered convolution. The last module will introduce how to apply attention mechanism and hierarchical features to the Siamese network.

3.1. Attention Mechanism

Recently, some researchers have attempted to apply attention mechanisms to convolution neural networks. Wang et al. [9] proposed residual attention mechanism Network uses an attention module in encoding-decoding mode to make the network not only perform better but also be more robust to noise by redefining the feature graph. Hu et al. [10] introduced a compact module to develop the relationship between channels. They used the feature of the average pooled layer to calculate the attention between channels. This section introduces the channel attention and spatial attention of the feature map, which greatly reduces the computational load of the feature map and speeds up the feature extraction.

The channel attention module focuses on the proportion of each channel. Input feature F aggregates the spatial information of the feature map by maximizing and averaging pooling to form two different feature descriptions, $F_{max}$ and $F_{avg}$, which are the maximum and average pooling features, respectively. These two features generate the final channel attention map through a multilayer sensor network. Multilayer perception network consists of a hidden layer with an activation size set to $\frac{C r \times c}{1}$, of which $r$ is the zoom factor. The final output eigenvector is generated by multiplying and adding pixels by pixel based on the characteristics of the multilayer sensor network. We can express this process with the formula (5):

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_0(F_{avg}) + W_1(F_{max})) \tag{5}$$

Where $\sigma$ represents the activation function, $W_0$ and $W_1$ are the two weight values of the multilayer perception machine, $W_0 \in \mathbb{R}^{C / r \times C}$, $W_1 \in \mathbb{R}^{C \times C / r}$, and the activation function uses the Relu function.

Unlike channel attention, spatial attention focuses primarily on location information. Re-calibrating channel feature a consists of the characteristics of channel attention output, then two different feature descriptions $F_{max}$ and $F_{avg}$ are obtained by maximizing and averaging the pooling operations, and a convolution layer is used to link the two together to generate the final two-dimensional spatial attention map. This calculation process can be represented by a formula (6):

$$M_s(F) = \sigma(f^{7 \times 7}([AvgPool(F); MaxPool(F)])) = \sigma(f^{7 \times 7}([F_{avg}; F_{max}])) \tag{6}$$

Where $\sigma$ denotes an activation function and $F_{avg} \in \mathbb{R}^{1 \times H \times W}$, $F_{max} \in \mathbb{R}^{1 \times H \times W}$, $f^{7 \times 7}$ denotes a filter size convolution operation.

Figure 1. Attention module.
As shown in Figure 1, the input image extracts the features through convolution layer, and then recalibrates the features through the channel attention module and the spatial attention module. The resulting features can represent the location information of the target and better fit the proportion of each channel.

3.2. Characteristic Analysis of Convolution Layer

Figure 2 shows the convolution characteristics of each layer extracted from the convolution neural network of the siamese network. Figure 2(a) represents the input image, 2(b) represents the characteristics of the first layer filter. Figure 2(c-g) signature maps representing the first to fifth layers of input image feature extraction (only 64 are selected for visualization), the details of the first layer signature map are clearer, the edge information of the target task is extracted, similar to the original image (a). The resolution of the feature map (d-g) is getting lower and lower, and the deeper the network, the more difficult it is to see the features on the picture, but the feature activation at the boundary between the target person and the background is significant. A few of the feature maps in the last layer activate the edge information and a few are hardly activated. It can be inferred that the feature maps in the higher layer have a more obvious classification of the target.

![Figure 2. Hierarchical convolution feature map.](image)

The bottom feature map contains more detail information, can see the edge of the object and texture information. As the network depth increases, the feature map contains less detail, and the high-level feature map contains more semantic information. These semantic features are more conducive to target classification and the low-level feature is more conducive to target positioning. In order to improve the accuracy and robustness of target tracking, this section combines high-level and low-level features to feature the target.

To analyze the impact of network extracted features on target tracking, we performed a comparative experiment using features extracted from different layers of the VGG network. Table 1 shows the success and accuracy of target tracking under different layer feature combinations.

| Convolution layer | 3   | 4   | 5   | 3, 5 | 4, 5 | 3, 4, 5 |
|-------------------|-----|-----|-----|------|------|--------|
| Success rate      | 0.563 | 0.591 | 0.603 | 0.633 | 0.612 | 0.619  |
| Accuracy          | 0.746 | 0.772 | 0.795 | 0.826 | 0.807 | 0.812  |

Table 1 compares the tracking results of our algorithm under feature combinations extracted from different layers. We find that the combination of features from the third and fifth layers can achieve better success and accuracy. So our algorithm uses the combination of features from the third and fifth layers. Low-level features can show better target details, which is more advantageous for target positioning than high-level features. Significance reflects more semantic information and is more advantageous for target classification.

3.3. Siamese Network of Hierarchical Attention Mechanism
Figure 3. A siamese network framework based on hierarchical attention mechanisms.

As shown in Figure 3, the target is tracked based on the siamese network framework. The network has two branches, one corresponding to the branch of the sample image, generally the first frame image $Z$, the other to the search image, the current frame image $X$. The two images are trained by the convolutional neural network in an asynchronous manner, and then after a series of operations, the sample image is re-marked with the feature map through the attention module, and then the search image is related to the operation. Finally, a response map between the two is obtained. We also apply the layered features. In the third and fifth layers of the convolution neural network, we fuse the output response map to get the final target response map. A formula can be used to represent the matching function between the two branches:

$$R^n = \sum_i \text{Conv}(\phi^n_i(Z), \phi^n_i(X)) + b$$

(7)

Where $R^n$ represents the result of the response map after correlation between the $n$th characteristic map of sample picture $Z$ and current picture $X$; $\text{Conv}(Z, X)$ represents the convolution operation on $X$ with $Z$ as the convolution core, $\phi^n_i(\cdot)$ represents the $i$th characteristic map of layer $n$, and $b$ represents the number of deviations used to adjust the response map.

In order to improve the generalization ability of target tracking network and increase the network's adaptability to scale, we add an Inception module, which can increase the network's adaptability to target scale. Different branches have different perception fields, and this Inception module can extract the multiscale information of the target, which has better adaptability to the scale of the target.

In siamese network tracking, the quality of response map affects the determination of target location and tracking performance. In order to allocate appropriate weights for the response graph, this paper presents an adaptive weighting method based on AdaBoost algorithm for the response graph, which gives smaller weights to the feature map with larger errors and larger weights to the feature map with smaller errors, so that the stronger judgment ability in the filter can play a key role. We calculate the error between the output response map and the expected Gaussian distribution centered on the target location, and then adapts the weighted response map using the AdaBoost algorithm based on this error. In frame $t-1$, if the target response map location is $(x, y)$, the expected Gaussian distribution is $g^{-1}(x, y)$, then the error function of response graph $R_i^{-1}(i = 3, 5)$ relative to the Gaussian distribution is:
\[ \varepsilon_{t-1} = \text{Mean}(\frac{\text{abs}(R_{t-1}(x, y) - g_{t-1}(x, y))}{R_{t-1}(x, y) + g_{t-1}(x, y)}) \]  

(8)

Where \( \text{abs}(\cdot) \) denotes an absolute value operation, \( \text{Mean}(\cdot) \) denotes an average operation, \( R_{t-1}(x, y) \) denotes the response value of the layer \( i \) feature map location \( (x, y) \) of frame \( t-1 \), \( g_{t-1}(x, y) \) denotes the expected Gaussian distribution of the target location \( (x, y) \), and then calculates the proportion of each response map in the final classifier based on this error function:

\[ \alpha_{t-1} = \frac{1}{2} \log \frac{1 - \varepsilon_{t-1}}{\varepsilon_{t-1}} \]  

(9)

Where \( \log(\cdot) \) is a natural logarithm.

The final response graph is:

\[ R' = \sum_{i=3}^{5} \alpha_{t-1} R_{t-1} \]  

(10)

So the target location \( (x, y) \) in the final response graph at \( T \)-frame time is:

\[ (x, y) = \arg \max_{i, y}(R'(x, y)) \]  

(11)

4. Experiments

To train our algorithm, we use SGD to minimize the logical loss. The training dataset is an ILSVRC 2015-VID dataset, which contains more than 30 basic categories and a total of 4500 target image sequences for training. The initial value is set to a Gaussian distribution and the scale is scaled according to the improved Xavier method. In this paper, the initial learning rate of convolution layer is set to 0.001, the training process includes 50 iterations, each iteration includes 5000 sample pairs, and each 50 epoch learning rate becomes 0.87.

4.1. Comparison Algorithm and Test Sequence

In this experiment, we compare it with the previous work SSC [11] and three other algorithms, SiamFC [5] and ECO [12] and KCF [13]. The performance of this tracking algorithm is evaluated by qualitative and quantitative analysis of these algorithms. In addition, this paper uses a representative sequence from OTB2015 to test, and the information of these sequences is shown in Table 2:

| Sequence name | Frames | Resolving power | Challenge                              |
|---------------|--------|-----------------|----------------------------------------|
| Ironman       | 166    | 720*304         | background clutter, scale change, occlusion |
| Lemming       | 1336   | 640*480         | Shielding, scale change, light change, quick motion |
| Soccer        | 392    | 640*360         | occlusion, background clutter, scale change |

4.2. Qualitative Analysis

In this section, Ironman, Lemming, and Soccer are used as examples to test. The comparison results are shown in Figure 4.
Figure 4. A comparison of the results of siamese network tracking based on attention mechanism.

Figure 4 is a comparison of the attention-based siamese network algorithm with several other algorithms. Taking Ironman, Lemming and Soccer video sequences in OTB2015 as examples, the performance of the algorithm is evaluated.

The sequences in Figure 4 are all affected by background clutter. In Figure 4, the SSC began to lose track of the target position at frame 11, and the KCF algorithm drifted at frame 15. The ECO algorithm in the 45th frame also drifted, and the SiamFC algorithm drifted in the 128th frame, and the algorithm in this paper can accurately track the target position. For the Lemming sequence, the SiamFC algorithm started to lose the target at frame 383, and the ECO algorithm also lost the target at frame 973. KCF tracking the target at frame 1028 was not very accurate, and the algorithm in this article can accurately track the target. In the Soccer sequence, because the background is too cluttered, other algorithms are missing the target. The algorithm in this paper can still track the target accurately. By qualitative analysis of these sequences, the algorithm in this paper has better accuracy and robustness than other algorithms.

4.3 Quantitative Analysis

The algorithm in this paper and four other algorithms are tested for their average accuracy and success under 100 test sequences of OTB2015. Figure 5 shows the average accuracy and success charts of the algorithm presented in this paper on OTB2015.
The accuracy threshold for this paper is set to 20 pixels. From Figure 5 (a), it can be seen that the algorithm in this paper still has 82.6% accuracy at 20 pixel threshold, which is 0.9%, 4.9%, 8.2% and 10.6% higher than ECO, SiamFC, SSC and KCF, respectively. In Figure 5 (b), we set the threshold of overlap accuracy to 0.5, and the success rate in this paper is 63.3%. Compared with ECO, SiamFC, SSC and KCF, the success rates of this paper are improved by 1.1%, 4.7%, 9.6% and 11.9%, respectively. In the case of background clutter, the algorithm in this paper still maintains high accuracy and accuracy. In Figure 5 (c), the accuracy of the algorithm presented in this paper is 2.3%, 9.4%, 12.2%, and 20.3% higher than that of ECO, SiamFC, SSC, and KCF, respectively, under background clutter. In Figure 5 (d), the success rate of the algorithm under background clutter is 1.9%, 9.9%, 10.6% and 16.6% higher than that of ECO, SiamFC, SSC and KCF, respectively.

5. Conclusions
This paper presents a siamese network tracking algorithm based on hierarchical attention mechanism. In order to obtain more robust target tracking results, different layers of features are fused, and the attention mechanism is used to recalibrate the feature map in the process of extracting features, and the AdaBoost algorithm is used to weight the target feature map, which improves the reliability of the response map. Experiments show that our proposed tracker performs satisfactorily on the visual tracker benchmark.

References
[1] Henriques J F, Caseiro R and Martins P 2012 Exploiting the circulant structure of tracking-by-detection with kernels European conference on computer vision pp 702-715.
[2] Wen S, Dong M, Yang Y, Zhou P, Huang T and Chen Y 2019 End-to-end detection-segmentation network for face labeling IEEE Transactions on Emerging Topics in Computational Intelligence doi: 10.1109/TETCI.2019.2947319
[3] Wen S, Liu W, Yang Y, Zhou P, Yan Z, Guo Z, Chen Y and Huang T 2020 Multi-label image classification via feature/label co-projection IEEE Transactions on Systems, Man and Cybernetics: Systems doi: 10.1109/TSMC.2020.2967071
[4] Danelljan M, Hager G and Shahbaz Khan F 2015 Convolutional features for correlation filter based visual tracking Proceedings of the IEEE International Conference on Computer Vision Workshops pp 58-66
[5] Bertinetto L, Valmadre J and Henriques J F 2016 Fully-convolutional siamese networks for object tracking European Conference on Computer Vision pp 850-865
[6] Ma C, Huang J-B and Yang X 2015 Hierarchical convolutional features for visual tracking Proceedings of the IEEE international conference on computer vision pp 3074-3082
[7] Wen S, Liu W, Yang Y, Zeng Z and Huang T 2019 Generating realistic videos from keyframes with concatenated GANs IEEE Transactions on Circuits and Systems for Video Technology 29 2337-248
[8] Wen S, Wei H, Huang T and Zeng Z 2018 Memristive fully convolutional networks: an accurate hardware image-segmentor in deep learning *IEEE Transactions on Emerging Topics in Computational Intelligence* 2 324-334

[9] Wang F, Jiang M and Qian C 2017 Residual attention network for image classification *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 3156-3164.

[10] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 7132-7141

[11] Qiu Y, Hu D and Qiu W 2019 Target tracking based on visual saliency combining super-pixel segmentation and conditional number blocking *Proceedings of the 3rd International Conference on Computer Science and Application Engineering* pp 1-6

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

[13] Henriques J F, Caseiro R and Martins P 2014 High-Speed Tracking with Kernelized Correlation Filters *IEEE transactions on pattern analysis and machine intelligence* pp 583-596