Tracking of small infrared targets based on convolutional neural network

Qiang Fan, Gangbo Sun and Erbo Zou

Huaizhong Institute of Electro-Optics, Wuhan National Laboratory for Optoelectronics, Wuhan 430223, China.

E-mail: fanq.2008@163.com

Abstract. As an important application of computer vision, visual tracking has been studied more and more. However, there are still many problems to be solved for infrared target detection, especially for weak infrared target detection and tracking. In addition, convolutional neural networks are widely used in visual applications such as target detection, recognition, and tracking. Therefore, in this paper, an infrared weak target detection and tracking method is designed. Based on the criteria of visual saliency, the real target can estimate by calculating the Euclidean distance between the target and its neighbours. The detected position of the target is sent to a fully convolutional siamese network DeepCF for tracking. As we know, the factors such as lighting changes, scale changes and occlusion cause the tracking failure. Therefore, a tracking confidence $T$ is proposed to judge whether the target is lost. The motion trajectory of the target is continuously estimated if it is lost. The experimental results show that the above method can effectively detect and track infrared weak targets.

1. Introduction

Infrared target detection and tracking has been widely used in video surveillance, situation awareness, autonomous vehicle navigation, military operations, aerospace remote sensing and so on. It is an important issue related to many practical applications. Compared with visible light imaging, the infrared imaging of the target has the characteristics of low resolution and fuzzy contour [1-2]. Thus, the detection and tracking of the target is still a challenging problem due to various interference factors such as partial occlusion, fast and sudden motion, illumination changes, and changes in viewpoint and attitude. In particular, most of the infrared image application scenes are complex backgrounds. When the distance is long, most of the targets are weak point targets, which are easy to be submerged in the background [3-4]. The existing detection and tracking for infrared targets mostly use the traditional image processing methods such as optical flow method, correlation filtering, centroid method, etc. Firstly, the background noise is suppressed and the input image is segmented and extracted by threshold value. Then the precise detection and tracking box are determined by means of morphological filtering. However, the threshold method is sensitive to scene and environment transformation, which makes the detection ability greatly reduced in complex scene or background changes.

In recent years, convolutional neural network (CNN) is used in visual target tracking increasingly [5-6]. But most of them have pool layer, which is not good for tracking small and weak targets. Yet, a fast and accurate detection and tracking method for small infrared target is raised.
2. Methods

Figure 1 is the flow chart of the algorithm based on salient feature extraction and convolution neural network. Detection is performed on the first input frame and tracking is performed on the subsequent frames. In the initial frame, Gauss filter is used to remove the discrete noise points. After removing the mean value of the filtered image, the Euclidean distance between the pixel and the surrounding neighbourhood is calculated. As the result, saliency map of the target is achieved. Through the dynamic threshold, the saliency map is segmented. The subsequent input frame is tracked based on the position obtained from the initial frame. Meanwhile, tracking confidence $T$ is computed during tracking to determine whether the tracking is lost. The detail is discussed as follow.

![DeepCF flowchart](image)

**Figure 1.** DeepCF flowchart. Frame pairs are feed into the network to extract feature and employ correlation filter to compute score map. X represents exemplar frame which include target object and z represents search frame.

2.1. Object detection

In general, the resolution of infrared image is lower than that of visible image and the edge of target is fuzzy. So, the target is difficult to distinguish when the distance is long. In extreme cases, the captured object area is approximately to a point. In fact, the objects can be picked up accurately by human. According to the principle of human visual attention mechanism, the objects of concern are often significantly different from the surrounding background or have a large contrast. Furthermore, in infrared images, the brightness of the target is generally higher than that of the background. Here, the saliency detection method is used to extract the infrared target from the input image. The steps of target detection are as follows:

1) The significance analysis of visible light image first changes the image from RGB color channel to lab color channel. The purpose is to highlight the color characteristics of the target in the image. Due to the lack of color information, the information of R, G and B channel are the same. Indeed, it is directly converted to grayscale image on the basis of the original image for processing.
(2) The noise of infrared image will affect the accurate marking of the target. The linear weighting of the whole image with Gaussian filter can effectively overcome a large number of random noise on the image and reduce the false alarm rate.

(3) Calculating the mean value of the filtered image. Next, removing the mean value from the original image. The result is recorded as $I_{\text{gaussian}}$.

(4) The Euclidean distance between each pixel of the original image and the processed result is obtained, and the significant image is obtained.

$$I_s = ||I(x,y) - I_{\text{gaussian}}(x,y)||$$

(5) Get the maximum value in the significant graph. The saliency graph is normalized.

(6) At the end, based on the saliency map processed in the above steps, the dynamic threshold can be obtained by using the OTSU. The final target area can be detected.

2.2. Object tracking

The application of CNN makes the trained network can learn the multi-dimensional characteristics of the target. Compared with the traditional method, it has higher accuracy, stronger generalization ability and better robustness. According to this, we design a siamese neural network named DeepCF, which is an end to end lightweight network. The network structure is shown in figure 2. Specifically, the tracker regards the correlation filtering as a special layer added in the siamese network. Further, the network output is defined a response map as the target position. By analyzing the closed form of mathematical solution, it is easy to get a network that is most suitable for tracking learning characteristics.

![Figure 2. Fully convolutional siamese network architecture.](image)

In order to achieve real-time tracking, a correlation filter is added after a network in the full convolution asymmetric siamese network, and the two branches of the network are correlated. In the equation (2), a cross-correlation filtering module is used to calculate a standard correlation filter template by solving a ridge regression problem in the frequency domain from a training feature map.

$$\varepsilon = ||\sum_{i=1}^{n} w^i \ast \phi(x)^i - y ||^2 + \lambda \sum_{i=1}^{n} ||w^i||^2$$

Where $w^i$ refers to the channel $l$ of filter $w$, $\ast$ means circular correlation and the constant $\lambda \geq 0$ is regularization coefficient. For the detection process, we crop a search patch and obtain the features $\phi(z)$ in the new frame, the translation can be estimated by searching the maximum value of correlation response map $g$:

$$g = F^{-1}(\sum_{i=1}^{n} \hat{w}^i \ast \phi^i(z))$$

Where the $w^i$:

$$\hat{w}^i = \frac{\phi^i(x) \ast \hat{y}^*}{\sum_{i=1}^{n} \phi^i(x) \ast (\hat{\phi}^i(x))^* + \lambda}$$

Here, $\hat{y}^*$ denotes the discrete Fourier transform $F(y)$, $y^*$ represents the complex conjugate of a complex number y and $\odot$ denotes Hadamard product.
2.3. Tracking confidence parameter T

At present, in order to improve the tracking speed, most of the trackers do not adopt the real-time update strategy. Yet, the result of the DeepCF tracker is update each frame. The update result as a template for the next frame. However, in the process of tracking, it is inevitable to that the target occlusion and other factors will lead to obvious changes in the response map of correlation filter. The filter is very sensitive to the reliability of the sample. Once there is a deviation, it is easy to transmit the bias to the next frame. With the continuous accumulation of the bias, the target will eventually lost. Therefore, a more reliable template update strategy is needed to increase the reliability of samples.

We know that the expectation of the tracking result is the Gaussian response centered on the target, which means that the real position of the object in the response map is peaked, while the non-object area is flat. If the peak is lager, the more reliable the sample. Conversely, the difference between the peak and the surrounding value is not obvious, the sample is less reliable. Here, tracking confidence T is designed as:

$$T = \frac{|P_{\text{max}} - P_{\text{min}}|}{m \cdot n \sum_{i=0}^{n} (P_{\text{max}} - P_{\text{min}})^2}$$  

Where \(m\) and \(n\) is the position of the pixel, \(P_{\text{max}}\) and \(P_{\text{min}}\) is the maximum and minimum value of the response map respectively.

2.4. Anti-occlusion algorithm

Once the target is occluded which determined by parameter T, it indicates that the tracking result of the current frame is not authentic. The target position of the current frame is predicted based on the tracking result of the previous M frame. In this paper, the curve fitting method is used to predict the target position. The purpose of this method is to facilitate the embedded hardware implementation. When the target is occluded and deviated from the predicted position, a larger image search area is selected. To better approximate the trend of real target motion and remove the influence of interference, the coordinates \(x(t)\) and \(y(t)\) of the target are established by curve fitting. The first 20 frames are choose to fit by using quadratic function. The target position prediction is as follows.

$$\hat{x}(t) = a_x + a_t t + a_{t^2}$$

$$\hat{y}(t) = b_x + b_t t + b_{t^2}$$

Using Least Square Method to Operate, the formula above can be computed:

$$\begin{bmatrix} a_x \\ a_t \\ a_{t^2} \end{bmatrix} = A^{-1} \begin{bmatrix} \sum_{i=1}^{m} x(i) \\ \sum_{i=1}^{m} x(i) t \\ \sum_{i=1}^{m} x(i) t^2 \end{bmatrix}$$

$$\begin{bmatrix} b_x \\ b_t \\ b_{t^2} \end{bmatrix} = A^{-1} \begin{bmatrix} \sum_{i=1}^{m} y(i) \\ \sum_{i=1}^{m} y(i) t \\ \sum_{i=1}^{m} y(i) t^2 \end{bmatrix}$$

Here, matrix \(A\) is:

$$A = \begin{bmatrix} m & \sum_{i=1}^{m} t & \sum_{i=1}^{m} t^2 \\ \sum_{i=1}^{m} t & \sum_{i=1}^{m} t^2 & \sum_{i=1}^{m} t^3 \\ \sum_{i=1}^{m} t^2 & \sum_{i=1}^{m} t^3 & \sum_{i=1}^{m} t^4 \end{bmatrix}$$

3. Implementation details

For each video, we choose each pair of frames and fed the cropped pair of target patches of 3×padding size to the network. The cropped inputs are resized to 128×128. Stochastic gradient descent (SGD) with momentum of 0.85 was applied to train the network from the scratch. The weight decay \(\gamma\) was set to 0.0005, the learning rate was set to 1e-5. The model is trained for 40 epoch with a mini-batch size of 128. The proposed method is implemented in MATLAB with MatConvNet. All
experiments are conducted on a workstation with Intel i7 8700K at 3.2GHz and a single NVIDIA GeForce GTX 1080 Ti GPU.

4. Results
The OTB dataset is a standard benchmark for visual tracking [7-8]. The extended version contains 100 challenging targets with 11 annotated attributes including scenarios such as illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutters (BC), and low resolution (LR). We evaluate the proposed method with comparisons to some state-of-the-art trackers. These trackers are including MDNet, DeepSRDCF, SRDCF, SiameseFC, HDT, HCFT, CNN-SVM, LCT, MEEM and DSST. All trackers with the ground-truth object state in the first frame and report the average precision and success rate of all results. The success plots show the percentage of frames in which the overlap score exceeds a threshold. The precision plots show the percentage of frames where the center location error is within a threshold. Figure 3 and figure 4 presents the precision and success plot on OTB50 and OTB100, respectively. The results show that the proposed method performs a third grade among these trackers.

Figure 3. Precision and success plots on the OTB50 benchmark.

Figure 4. Precision and success plots on the OTB100 benchmark.
Figure 5. Snapshots of the results generated by the proposed method.

Figure 5 illustrates some typical infrared objects such as UAV, bird and bus. Obviously, the proposed method can effectively tracking weak infrared target like UAV and bird. Meanwhile, it can overcome the occlusion problem. In contrast, our method achieves a good balance between tracking accuracy and tracking speed.

5. Conclusion
In this paper, a fast and accurate tracker DeepCF for dim infrared target is proposed. This tracker can effectively suppress the surrounding and highlight the contour edge of the real target with less computation. A full convolutional siamese network is used to extract the features of different levels of the target and a correlation layer is employed as the criterion to generate a response map. The estimate position of the target can achieve from the response map readily. During the process of tracking, the confidence parameter $T$ is computed continuously to judge whether the target is occluded or lost. Since the target is occluded or lost, the anti-occluded algorithm is utilized to update the trajectory. Compared with the current engineering methods, this method can meet the real-time requirement and facility to deploy on the embedded system.

Acknowledgments
This work was financial supported by grants from the National Major Special Program of Scientific Instrument & Equipment Development of China (No. 2013YQ290489)

References
[1] Smeulders, Arnold WM, et al.2014 IEEE transactions on pattern analysis and machine intelligence 36 pp 1442-68
[2] Yang, Hanxuan, et al. 2011 Neurocomputing 74 3823-31
[3] Li, Peixia, et al. 2018 Pattern Recognition 76 323-38
[4] Liu Q, Lu X, He Z, et al. 2017 Knowledge-Based Systems 134 189-98
[5] Bertinetto, Luca, et al. 2016 Fully-convolutional siamese networks for object tracking European conference on computer vision Springer Cham
[6] Fan Q, Lei B, Tan H. 2019 A robust anti-occlusion object tracking method Eleventh International Conference on Digital Image Processing (ICDIP 2019) International Society for Optics and Photonics 11179 111793K

[7] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013 Online object tracking: A benchmark in Proceedings of the IEEE conference on computer vision and pattern recognition

[8] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2015 Object tracking benchmark IEEE Transactions on Pattern Analysis and Machine Intelligence 37 1834–48