Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Overview of Algorithms for Visual Single Object Tracking

Kaidi Shi *

School of North China Electric Power University, Baoding 071000, China.

*Corresponding author Email: ncepu_skd@163.com

Abstract: Single object tracking is one of the research directions in the field of computer vision. Its basic task is to give the target to be tracked in the first frame, and obtain the position information of the target in each frame in the subsequent frames, so as to complete the further research on the motion behavior and law of the target. Due to the high complexity and strong interference of the tracking scene, the diversity of target apparent changes and the mutual fusion of kinds of information, the tracker needs to balance the performance measurement indicators such as robustness, accuracy and real-time. This paper will review the development of target tracking algorithm, especially the development of Tracker Based on deep learning and siamese-based trackers, and explore the existing challenges and prospects the future research directions worthy of attention, so as to provide reference for the future research work in this field.
Keywords: Computer vision, Single object tracking, Deep learning, Siamese-based trackers.

---

1. INTRODUCTION

The research of object tracking can be traced back to the use of optical flow method by Lucas[1] and others in 1981. After that, with the proposal of various high-quality data sets and the development of machine learning and deep learning, from mean shift, correlation filtering and other algorithms to today's object tracking algorithm based on neural network, more and more excellent methods have emerged in the field of target tracking.

The development of object tracking can be roughly divided into three stages: the first stage focuses on the application of classical algorithms and machine learning in target tracking around 2000. This kind of algorithm has low computational complexity and fast running speed, but its robustness and accuracy are relatively low. In the second stage, with the proposal of Mosse algorithm[2], correlation filtering method has become one of the research hotspots, and with the characteristics of fast speed and high accuracy, it has a good ranking in each evaluation data set. During this period, deep learning methods also began to make achievements in the field of image processing, and some excellent algorithms such as MDNet[3] appeared. The last stage is from 2016 to now. With the enrichment of datasets, the in-depth learning method represented by twin networks has continuously improved the robustness and accuracy of the algorithm, and its performance has been among the best in all major data sets and competitions, showing a strong ability. Especially in recent years, attention mechanism

has shown excellent results in visual tasks. The depth tracker based on transformer has achieved excellent performance in recent years, and has received extensive attention.

2. DEVELOPMENT OF OBJECT TRACKING ALGORITHM

2.1 Traditional object tracking algorithm.

Traditional object tracking algorithms usually use manual features to model the target, and then train robust discriminative or generative models to achieve target tracking. Typical methods include optical flow method, mean shift algorithm, sparse representation and correlation filtering algorithm.

The concept of optical flow was first proposed by Gibson in 1950. It is the instantaneous speed of the pixel motion of a space moving object on the observation imaging plane. It is a method to find the corresponding relationship between the previous frame and the current frame by using the changes of pixels in the time domain in the image sequence and the correlation between adjacent frames, so as to calculate the motion information of objects between adjacent frames. It is based on the following assumptions: the brightness between adjacent frames is constant; The frame taking time of adjacent video frames is continuous, and the motion of objects between adjacent frames is relatively small; Maintain spatial consistency, that is, the pixels of the same sub image have the same motion. Optical flow method assigns a velocity vector to each pixel in the image, thus forming a motion vector field. At a certain time, the points on the image correspond to the points on the three-dimensional object one by one, and this correspondence can be calculated by projection. According to the velocity vector characteristics of each pixel, the image can be dynamically analyzed. However, when optical flow method is used to detect moving objects, the amount of calculation is large, which can not guarantee the real-time and practicality.

Mean shift algorithm is a nonparametric method based on density gradient rise. It is often used in object tracking, classification and other scenes in image recognition. Its core idea is to select a center point randomly, then calculate the average value of the distance vector from all points to the center point within a certain range of the center point, calculate the average value to an offset mean value, and then move the center point to the offset mean value position. Through this repeated movement, the center point can gradually approach the best position. This idea is similar to the gradient descent method. By constantly moving in the direction of gradient descent, the local optimal solution or global optimal solution on the gradient can be reached. Its remarkable advantage is that the algorithm is simple and easy to implement, which is very suitable for real-time tracking. However, tracking small targets and fast-moving targets often fails, and it can not recover tracking under all occlusion.

Sparse representation gives a set of over complete dictionaries, linearly represents the input signal with this set of over complete dictionaries, and makes a sparsity constraint on the coefficients of the linear representation, then this process is called sparse representation. The object tracking method based on sparse representation transforms the tracking problem into a sparse approximation problem. For example, L1tracker [4], the pioneer of sparse tracking, believes that candidate samples can be represented sparsely through target templates and trivial templates, and a good candidate sample should have a more sparse coefficient vector. Sparsity can be obtained by solving an L1 regularized least squares optimization problem. Finally, the candidate samples with the minimum reconstruction error with the target template are used as the tracking results. L1tracker uses trivial templates to deal

with occlusion, and uses non negative constraints on sparse coefficients to solve the problem of background speckle. However, in the classical sparse representation algorithm, the modeling and updating method of the target template is inefficient, and the tracking performance is unreliable. There are many subsequent improvements based on L1tracker, such as paper [5, 6].

Correlation filtering was first applied to signal processing to describe the correlation, or similarity, between two signals. Its principle is that the convolution response of two correlated signals is greater than that of uncorrelated signals. In target tracking, the filter trained by the target template is used to filter the video frame, and the maximum value is found as the target position of the current frame in the obtained response graph. Therefore, the process of object tracking can be transformed into the process of correlation filtering the image of the search area to find the target position, that is, to find the maximum position of the filter response graph.

According to this idea, a large number of methods based on correlation filtering have been proposed, such as the earliest Mosse tracking method using the simplest correlation filtering idea. Then there are many related improvements based on Mosse, such as CSK[7] and KCF[8] with the introduction of core method, which have achieved good results, especially the KCF calculated by cyclic matrix. On the basis of KCF, a series of methods have been developed to deal with various challenges. For example, DSST[9] can deal with scale changes, and the reliable patches based correlation filtering method can deal with occlusion, etc. However, all the above methods based on correlation filtering are affected by boundary effect. In order to overcome this problem, SRDCF [10] came into being. SRDCF uses spatial regularization to punish the correlation filter coefficients, and obtains results comparable to deep learning tracking methods.

Early correlation filtering algorithms used manual features to represent targets, such as KCF with HOG[11] features. These methods are fast but not robust. The reason is that manual features are difficult to adapt to various changes of targets. On the other hand, the features extracted by convolution neural network have good anti-interference ability and target representation ability. Therefore, In the field of target tracking, the research of using convolution neural network to extract features combined with correlation filtering method appears.

2.2 Siamese-based trackers.

At present, the object tracking algorithm based on Siamese neural network has become the mainstream method of single target tracking. All this comes from SiamFC[12] in 2016. SiamFC actually regards tracking as a matching problem. Siamese (twin neural network), as the name suggests, is a paired structure. Specifically, the structure has two inputs, one is the template as the benchmark, and the other is the candidate sample to be selected. In the single target tracking task, the template as the benchmark is the object we want to track. Usually, the target object in the first frame of the video sequence is selected, and the candidate sample is the image search area in each subsequent frame. What the twin network needs to do is to find the candidate area in each subsequent frame that is most similar to the template in the first frame, that is, the target in this frame, so that we can track a target.

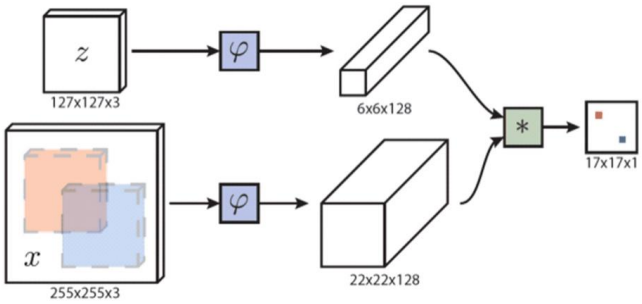

Figure 1. Framework illustration of SiamFC algorithm[12].

But SiamFC discards the background information completely. In other words, SiamFC lacks the use of the background, and it uses a fixed size scale to mark the tracked target, which is dynamic in the video, so SiamFC cannot automatically adjust the scale of the mark box. At the same time, there are excellent methods to solve the scale problem in the field of target detection, so some scholars try to integrate twin network and target detection technology. Literature [13] proposed SiamRPN algorithm in 2018, which treats target tracking as a single sample detection task. SiamRPN draws on the regional recommendation network RPN [14]. After the twin network extracts features, it sends the feature map into the classification branch and regression branch, so that the tracker can return to the target position and shape. While ensuring real-time performance, the anchor box mechanism of RPN effectively reduces the change of target scale.

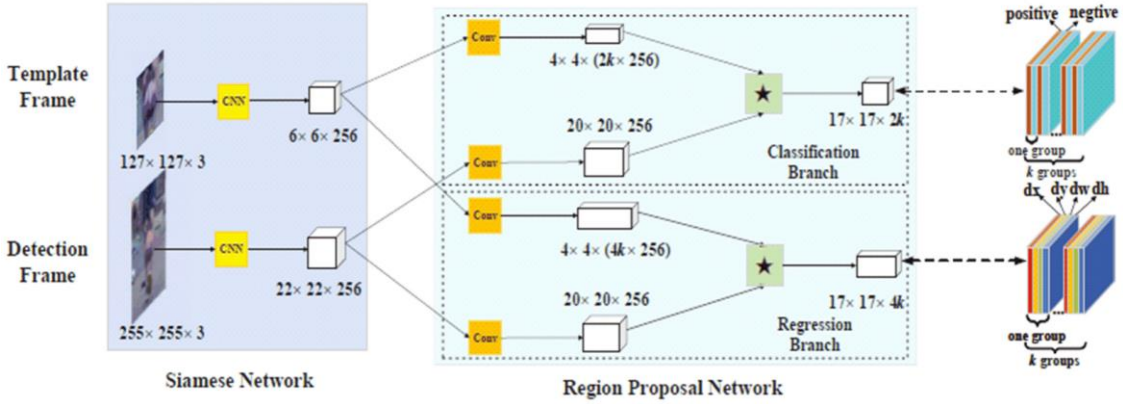

Figure 2. Framework illustration of SiamRPN algorithm[14].

However, SiamRPN has weak generalization performance for the model, and still has a high response when the target is lost. In the training stage, there is a problem of sample imbalance, that is, most samples have no semantic background. This leads the network to learn only the distinction between the background and the foreground, that is, to extract objects from the image. Some scholars have solved this problem from the perspective of data enhancement and proposed DaSiamRPN[15] algorithm. Compared with the original algorithm, DaSiamRPN enhances the discrimination ability of the classifier and improves the generalization performance of the model. In order to further improve the performance of SiamRPN, Li et al. Tried to replace the backbone network of Siamese with a deeper network ResNet[16] from AlexNet[17]. The author found that without padding, as the network depth increases, the feature map will become smaller and smaller, and too many features will be lost. Adding padding will destroy the translation equivariant of convolution and make the position learned by the twin network deviate. Therefore, we try to use the uniformly distributed sampling method to eliminate the above effects, and propose SiamRPN++[18]. At the same time, we add

multi-layer fusion and deep cross-correlation mechanism after the feature extraction network to improve the accuracy and reduce the amount of parameters.

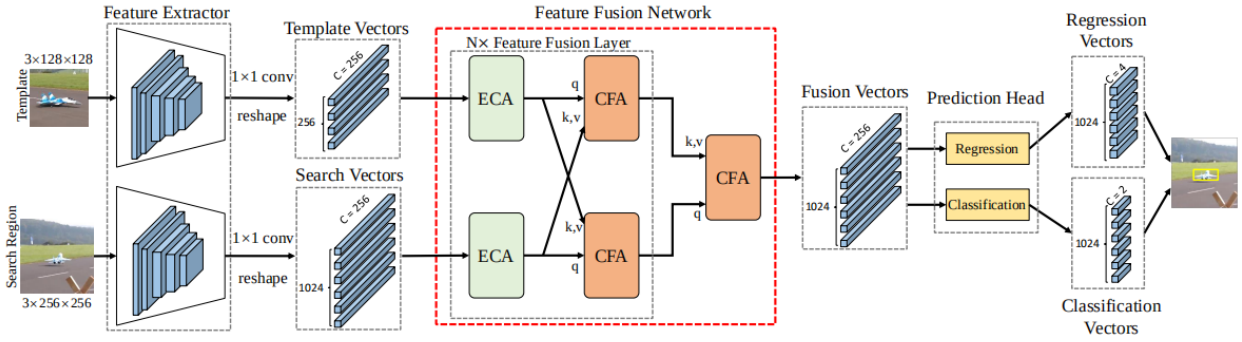

Figure 3.Framework illustration of TransT algorithm[22].

SiamDW[19] proposed by other scholars also solves the problem of Siam series network depth. SiamDW believes that there are two reasons why Siam series networks cannot be deepened. First, with the increase of network depth, the receptive field of the network increases, which reduces the discrimination of the network and the accuracy of regression; Second, padding will introduce spatial bias, because if padding is used, it will have padding for the convolution kernel as a template image, while for the search image, some areas are padding, which the author believes will lead to discontinuity and poor target recognition. Therefore, the author of this paper explores the ways to deepen the Siamese series network from two aspects: one is to adjust the size of the receptive field through stripe; Aiming at the padding problem, the author designs a new residual network structure to delete the elements affected by padding on the feature map.

The above algorithm has good performance and speed, but the fixed target template is easy to cause error accumulation, resulting in the model can not adapt to the drastic changes in the appearance of the target, and its robustness has a bottleneck, so the introduction of online update mechanism is necessary. Valmadre Tried to add a correlation filter layer to the twin network structure to realize the online update function, and proposed the CFNet[20] algorithm. CFNet deduces the forward and back propagation formulas of correlation filtering, so that correlation filtering can be embedded into the pre training network to realize end-to-end learning. DCFNet[21] is different from CFNet in network structure. It replaces cross-correlation operation with correlation filter layer on the basis of SiamFC to realize online update and improve the robustness of the algorithm.

Transformer structure was originally proposed in natural language processing tasks. The core module of transformer is the attention mechanism, which can aggregate the global information to the desired location. Because this structure can make full use of GPU and other hardware devices for parallel computing, it is more suitable for processing long sentences than the traditional RNN method. In recent years, attention mechanism has shown excellent results in visual tasks. The depth tracker based on transformer has achieved excellent performance recently, and has received extensive attention.

Specifically, Chen et al. [22] proposed TransT and designed a transformer feature fusion network with its own ego context augmentation (ECA) and cross feature augmentation (CFA) modules, which can avoid the problems of traditional cross-correlation operations falling into local optimal solutions and losing some semantic information. Yan et al. [23] proposed STARK, which directly regards target tracking as a bounding box prediction problem, and designed a tracker based on transformer,

which globally models the spatio-temporal characteristics between template frames and search frames through self attention and cross attention modules.

These trackers have achieved great performance improvement by trying to establish long-term temporal dependencies to sense context information, showing the transformer's powerful global modeling ability. However, the current exploration of transformer is still in its infancy, and there are still many places to explore, such as reducing computational complexity, rethinking the overall tracking architecture based on transformer, and so on. In addition, the computing cost of running the traditional convolutional backbone network (ResNet) on the platform with limited computing resources is still high, so it may be an interesting direction to study the efficient backbone network of mobile platform, transformer tracker based on lightweight, tracker based on pure transformer architecture, etc. in the future.

3. SUMMARY

Although the twin tracking algorithm has certain advantages over other methods, it is still difficult to meet the needs of tracking tasks for speed and performance in the application of actual scenes, and there is room for further research and development. Combined with the summary and analysis of each algorithm in the article and the comparison results of the experimental part, and considering the current research hotspot in the field of vision, the future further research of twin target tracking algorithm can be considered from the following aspects:

1) Generalization ability. Current trackers rely heavily on a large number of labeled training data, so the relative cost of acquiring data is very high, and the amount of specific scene data is small, the complexity is low, and the diversity is insufficient. So far, improving the generalization ability of the model can be effectively improved with the help of small sample learning and self supervised learning. First, small sample learning can solve the problem of insufficient generalization ability of the model caused by the small number of training samples. Self supervised learning can effectively solve the problem of lack of annotation of training data. By improving the generalization ability of the model with the help of training data, it can be adaptively applied to complex scenes, and can improve the overall development level in the field of target tracking.

2) Tracking refinement. From multi-scale search method to anchor frame based and anchor frame free target tracker, the algorithm model design tends to be refined. The anchor based method can deal with the changes of target scale and aspect ratio, but this method is very sensitive to the number, size and aspect ratio of anchor frames. At the same time, it is simple and effective to classify objects directly and return their borders based on the anchor free method. However, these methods have alignment problems between the predicted frame and convolution features, which limits the performance of these trackers. If we can solve whether the predicted frame and feature are aligned, we believe that the performance of the tracker will be significantly improved.

3) Model architecture. The neural network architecture based on experience design is gradually replaced by the structure based on neural network structure search. In the detection task, many work has focused on the backbone network, feature extraction network and detection head search. In the tracking task, there are also a lot of empirically designed networks, especially the part of template branch and search branch fusion is obtained through a large number of experiments (including

convolution type, number of channels, fusion operation and the number of fusion points). The structure with better performance will be obtained by neural network architecture search, but at present, in the definition of search space, the definition of training indicators Training convergence and other issues still need the unremitting efforts of researchers.

REFERENCES

[1] Lucas B D,Kanade T.An iterative image registration technique with an application to stereo vision.Proc of the 7th Intetnational Joint Conference on Artificial Intelligence.Francisco CA:Motgan Kaufmann Publishers,1981:674-679.

[2] Bolme D S,Beveridge J R ,Draper B A,et al.Visual object tracking using adaptive correlation filters.Proc of IEEE Computer Society Conference on Computer Version and Pattern Recognition.Washington DC:IEEE Computer Society,2010:2544-2550.

[3] Nam H,Han B. Learning multi-domain convolutional neural networks for visual tracking.Proc of IEEE Conference on Computer Version and Pattern Recognition.Washington DC:IEEE Computer Society,2016:4293-4302.

[4] MEI X,Ling H B.Robust visual tracking using L1 minimization//Proc of the IEEE International Conference on Computer Version.Washington,USA:IEEE,2009:1436-1443.

[5] Zhang T,Ghanem B,Liu S,et al.Robust Visual Tracking via Multi-Task Sparse Learning. Proceedings of the 2012 Conference on Computer Version and Pattern Recognition, 2012:2042-2049.

[6] Zhang T,Liu S,Ahuja N,et al.Robust Visual Tracking Via Consistent Low-Rank Sparse Learning[J]International Journal of Computer Vision,2014,111(2):171-190.

[7] Henriques J F,Caseiro R,Martins P,Batista J.Exploiting the circulant structure of tracking-by-detection with kernels. In:Proceedings of Computer Vision.Lecture Notes in Computer Science, vol.7575. Berlin,Heidelberg:Springer,2012.702−715.

[8] Henriques J F,Caseiro R,Martins P,Batista J.High-speed tracking with kernelized correlation filters.IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(3):583−59.

[9] Danelljan M,Hager G,Khan F S,Felsberg M.Accurate scale estimation for robust visual tracking. In: Proceedings British Machine Vision Conference.London,England:BMVA Press, 2014.65.1−65.11.

[10] Danelljan M,Hager G,Khan F S,Felsberg M.Learning spatially regularized correlation filters for visual tracking.In:Proceedings of the 2015 IEEE International Conference on Computer Vision.Santiago, Chile:IEEE,2015.4310−4318

[11] O'Rourke S M, Herskowitz I, O'Shea E K.Yeast go the whole HOG for the hyperosmotic response.Trends in Genetics,2002,18(8):405-412

[12] Bertinetto L,Valmadre J,Henriques JF,et al.Fully convolutional siamese networks for object tracking.In:HuaG,J égou H.eds.European Conference on Computer Vision.Cham:Springer.2016. 850-865.

[13] Li Bo,Yan Junjie,Wu Wei,et al.High network.IEEE Conference on Computer Vision and Pattern Recognition,2018:8971-8980.

[14] Ren S,He K,Girshick R,et al.Faster R-CNN:towards real-time object detection with region proposal net-works.International Conference on Neural Information Processing Systems, 2015:91-99.

[15] Zheng Zhu, Qiang Wang, Bo Li, Wu Wei, Junjie Yan, Weiming Hu.Distractor-aware Siamese Networks for Visual Object Tracking..Proc.of the European Conference on Computer Vision, 2018:101-117.

[16] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks.Conference on Neural Information Processing Systems,2012:1106-1114.

[17] He Kaiming,Zhang Xiangyu,Ren Shaoqing,et al.Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.

[18] Li Bo,Wu Wei,Wang Qiang,et al.SiamRPN++:evolution of siamese visual tracking with very deep net-works.IEEE Conference on Computer Vision and Pattern Recognition,2019:4282- 4291.

[19] Zhang Zhipeng,Peng Houwen.Deeper and wider siamese networks for realtime visual tracking.IEEE Conference on Computer Vision and Pattern Recognition,2019:4591-4600.

[20] VALMADRE J,BERTINETTO L,HENRIQUESJ,et al.End-to-End representation learning for correlation filter based tracking.Proc.of the IEEE Conference on Computer Vision and Pattern Recognition, 2017:5000-5008.

[21] Wang Q,Gao J,Xing J,et al.DCFNet:Discriminant Correlation Filters Network for Visual Tracking. ICIP,2017.

[22] Chen X,Yan B,Zhu J,et al.Transformer Tracking.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2021.
[23] Yan B,Peng H,Fu J,et al. Learning spatio-temporal transformer for visual tracking.2013,arXiv: 17154, 2021.