Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 146–153

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

SiaMemory: Target Tracking

Li Bo Chang[a,c*], Shang Bing Zhang[a], Mohamed Sedky[b], Hui Min Du[c] , Shi Yu Wang[a]

a Department of Computer Science and Engineering, NWPU University, Xi'an, China
b Staffordshire University, Stoke-on-Trent, UK
c School of Electronic Engineering, XUPT University, Xi'an, China

Abstract

This paper proposes, develops and evaluates a novel object-tracking algorithm that outperforms start-of-the-art method in terms of its robustness. The proposed method compromises Siamese networks, Recurrent Convolutional Neural Networks (RCNNs) and Long Short Term Memory (LSTM) and performs short-term target tracking in real-time. As Siamese networks only generates the current frame tracking target based on the previous frame of image information, it is less effective in handling target's appearance and disappearance, rapid movement, or deformation. Hence, our method a novel tracking method that integrates improved full-convolutional Siamese networks based on all-CNN, RCNN and LSTM. In order to improve the training efficiency of the deep learning network, a strategy of segmented training based on transfer learning is proposed. For some test video sequences that background clutters, deformation, motion blur, fast motion and out of view, our method achieves the best tracking performance. Using 41 videos from the Object Tracking Benchmark (OTB) dataset and considering the area under the curve for the precision and success rate, our method outperforms the second best by 18.5% and 14.9% respectively.

Keywords: object tracking; deep learning, Siamese networks, recurrent convolutional neural network;

* * Corresponding author. Tel.: +(86) 15929926373
E-mail : changlibo@xupt.edu.cn

10.1016/j.procs.2019.06.022

1. Introduction

In computer vision, because partial occlusions, target deformations, motion blur, background clutters, and object deformation, scale changes and illumination variations, visual tracking is a challenging task [1]. Some researchers improve the robustness of their tracking algorithm by applying machine learning [2], such as TLD (Tracking-Learning-Detection) [3] and KCF (kernelized correlation filters) [4]. Because CNNs (convolutional neural networks) have strong capabilities of learning feature representations, other researchers to apply CNNs to address challenges faced by tracking using traditional machine learning[5]. However, approaches that are solely based on CNNs have the following problems: because of no effective use of temporal continuity, these algorithms cannot well handle obstructed target scenarios; these algorithms that are based on deep-learning models in general have a huge number of parameters, to train such a huge model, i.e. lots of data is required, however, In contrast, in the case of target tracking, much fever training samples are available, hence render performance.

In response to the above issues, this paper proposes the following two improvements: 1. An optimized tracking algorithm based on CNNs by RNNs, that improves the robustness for object occlusions; 2. A strategy for segmented training based on transfer learning: different loss functions are used for different parts of the network to improve the training efficiency of the deep learning network. Some result are shown in Fig 1.The paper is organized as follows. discusses related work in tracking are discussesed in Section 2.We described the tracking framework in section 3. The employ system frameworks are briefly presented in section 3.1, the improved Siamese network structure and LSTM algorithm for object tracking presented section 3.2 and section 3.2 respectively, while we discussed the loss function in section 3.3. Section 4 contains the experimental results. Finally, we provided conclusions in section 5.

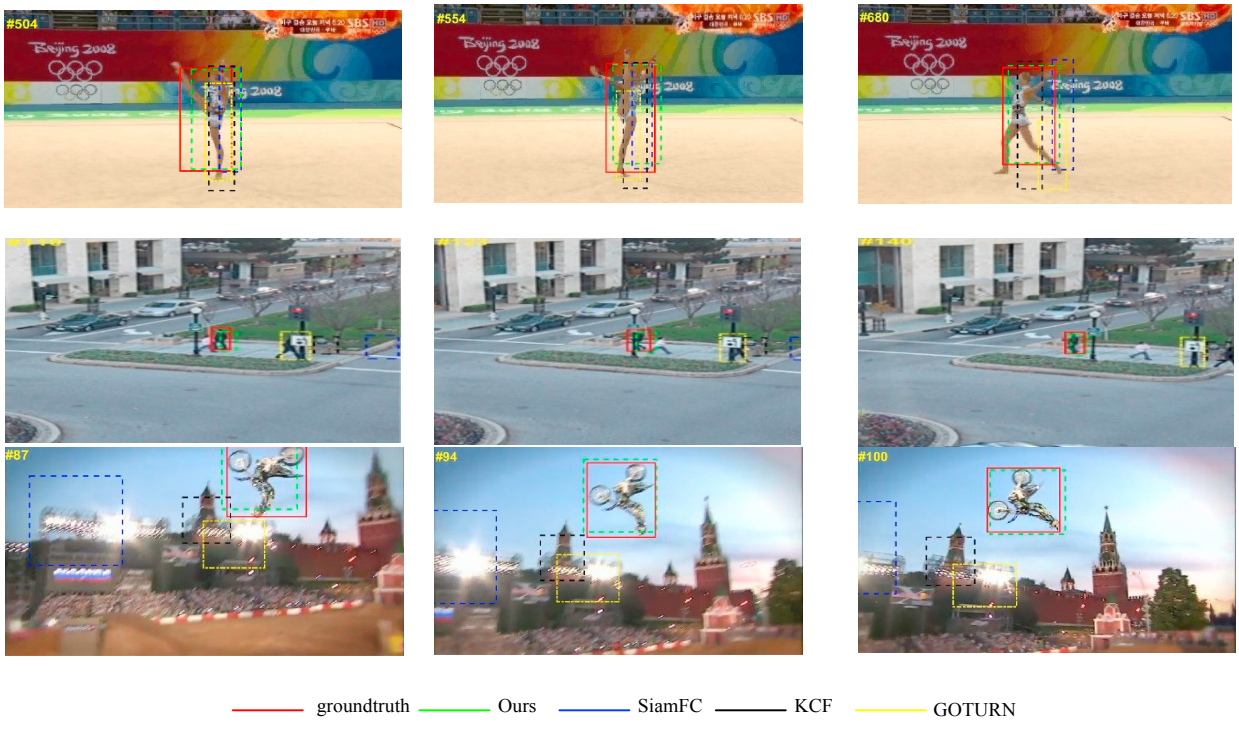


Fig 1: Compare whit KCF, GOTURN and SiamFC, SiaMemory tracker successfully tracks the object under deformations, occlusion or similar, occlusions and similar target challenges in unseen frames

2. Elated Work

For a systematic review and comparison, we refer the reader to a recent benchmark and a tracking challenge report [6]. We briefly explain tracking algorithm that are based on traditional machine learning ,CNNs and RNNs.

2.1. Tracking using traditional machine learning

Many advanced machine learning algorithms have been applied for object tracking, the most comprehensive performance among them is KCF that base on correlation filters and TLD that is a framework designed for long-term object tracking. KCF is a real-time algorithm, but loses the target when dealing with all kinds of complex scenes and struggles to track the appearance disappearance and reproduction of a target. TLD's principle is to use the tracked result as a training sample of the detector to adapt to the change of the appearance of the target, and at the same time use the detector to correct the tracking result and perform the re-detection after the target disappears. It incorporates machine learning based on the combination of trackers and detectors, which solves the problem of insufficient training samples to some extent [7]. However, these approaches is less robust compared with based on DNN algorithms, and it has high complexity, so cannot meet real-time performance.

2.2. Tracking using CNNs

The existing CNN tracker is based on the main exploration pre-trained object recognition network and is built on the discriminant or regression models[8-10]. Compared with traditional algorithms (e.g. KCF), this kind of algorithms can achieve better results, but it has poor real-time performance because of the high computational complexity in addition the performance of such algorithms is not stable. In order to improve the certainty of target tracking, Held D et al [11] proposed a Siamese network structure, this algorithm is based on deep learning can achieve real-time target performance, however its search range is limited due to the poor robustness wait exhibit. Subsequently, algorithms proposed in [12-15] improved such algorithms and improved the robustness of the algorithm by combine them with regression learning methods and correlation filtering. However, these algorithms have converted the tracking problem into a target detection problem, without considering the temporal continuity of the video signal. Therefore, the accuracy of such algorithms is poor when the target is hidden in the scene, additionally, the target cannot be re-located automatically when the target is lost.

2.3. Tracking using RNNs

Recently, some studies have tried to train RNNs for object tracking problems [16-18]. However, the tracing video often represents an irregular sequence signal some videos include various interference factors e.g. camera shake, therefore, these methods have not yet demonstrated competitive results on modern benchmarks.

Ning G et al[19] combines CNNs with RNNs to improve the accuracy of the tracking algorithm, however, the CNNs networks is completed using YOLO network which for target recognition, the target tag is required to be selected by ground truth, therefore this algorithm cannot be used under real conditions. We will unite CNNs that based on full-convolutional Siamese networks(such as [14]) and recurrent neural networks to improve the accuracy and robustness of the target tracking algorithm, especially for target hiding canaries.

3. Method

3.1. Method overview

In Fig2, we illustrated the overview of the tracking procedure. The entire method consists of three parts: 1. Fully-convolutional Siamese collecting rich and powerful visual features, and preliminary positional inference; 2. LSTM

model, deep space, suitable for sequence processing; 3. CNNs collects rich and powerful visual features of the entire video.

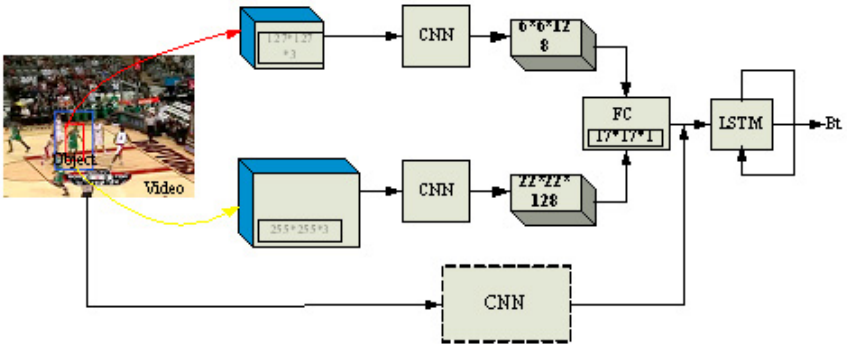


Fig 2: Simplified overview of ours method

3.2. Fully-convolutional siamese networks

Firstly, we use the CNNs to extract features of images and target, and then focusing on the current target, the twice the area as the search region. The goal of this CNNs is to return to the location of the target object in the search area, such as equation 1.

$$g_\rho(x',z') = G\left(f_\rho(x'), f_\rho(z')\right) \quad (1)$$

Where x' and z' are search region image and object image respectively, f_ρ is CNNs with learnable parameters ρ, the goal is to achieve a maximum value of the response map in the target location. Our starting point is a network similar to that of [14], and then we made the following two improvements: 1.Because max pooling has very high nonlinearity, it is difficult to learn through common nonlinear functions, we replace the max-pooling layer with convolution layers, such as conv_p1 and conv_p2 in table 1. 2. For online visual tracking, the lack of training samples becomes even more severe. Thus, directly we use of high dimensional signals to train fully connected networks may be over-fitting, which will degrade the tracker and gradually leads to tracking drift. In order to address the above issue, we use the 1*1 convolution method to reduce the dimension based on the network proposed by SiamFC, such as conv6 in table 1.

Table 1: channels number of each convolutional layer

Layers	Supports	Channels maps	Stride	For exemplar	For search	channels
				127*127	255*255	*3
conv 1	11*11	96*3	2	59*59	123*123	*96
conv_p1	3*3		2	29*29	61*61	*96
conv2	5*5	256*48	1	25*25	57*57	*256
conv_p2	3*3		2	12*12	28*28	*256
conv 3	3*3	384*256	1	10*10	26*26	*192
conv 4	3*3	384*192	1	8*8	24*24	*192
conv 5	3*3	256*192	1	6*6	22*22	*128
conv 6	1*1	128*64	1	6*6	22*22	*64

3.3. LSTM (long short term memory) of object tracking

We add RNNs on the CNNs to improve the tracking accuracy and solve the problem of short-term target disappearance. When the tracked target cannot only be determined by the video frame information or target is hidden, The algorithm can determine the target position according to the time domain information, or reposition the target when the target disappears and appears again because of the RNN network introduces time domain information. We use CNNs network to extract image information to reduce the amount of data and guide RNNs(LSTM) using a tracker that is based on CNNs. Specific realizations are presented in Equations 2-4.

$$\begin{Bmatrix} i_t \\ f_t \\ o_t \\ g_t \end{Bmatrix} = \begin{Bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{Bmatrix} T \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \tag{2}$$

$$c_t = f_t \text{ e } c_{t-1} + i_t \text{ e } g_t \tag{3}$$

$$h_t = o_t \text{ e } \tanh(c_t) \tag{4}$$

The LSTM can be described by Equations 2 to 4, where T is an affine transformation matrix with the parameters to be learned it, ft, ot, ht and ct are the input get, forget gate, output gate, hidden state and memory (cell state) respectively, gt selects the input to update the memory, and denote the sigmoid activation function and element-wise product respectively. In order to improve the real-time performance of the system, we have used one LSTM layer; the xt composed by the CNNs extracted video features and predicted target position.

$$x_t = \{fea_{cnns}, loc_{cnns}\} \qquad fea_{cnns} \in ¡^{1024} \qquad loc_{cnns} \in ¡^{4} \tag{5}$$

Among Equation 5, is a one-dimensional array that is extracted by the CNNs, is a one-dimensional array of four real numbers, which is the target position predicted by the CNNs. x_t consists of the result of the CNN-based short-term target tracker and features representing the current video global information, where fea_{cnns} is a 1024-dimensional vector representing the video global information, and loc_{cnns} is a 4-dimensional vector representing the short-term target tracker's tracking result.

3.4. Loss function design

Assuming that the background is represented by mutually independent and random signals, we use the Bernoulli reconstruction distribution. For the fully convolutional Siamese networks, we use the cross-entropy function as Equation 6, and we define the loss of the score graph as the average of the individual losses as shown in Equation 7.

$$l(y,v) = -y * \left(\log\left(v - (1-y) * \log(1-v)\right)\right) \tag{6}$$

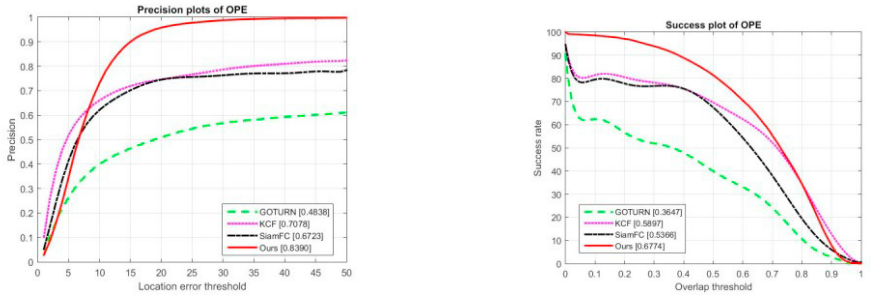
$$L(y,v) = \frac{1}{|D|} \sum_{u \in D} l\left(y[u], v[u]\right) \tag{7}$$

We input the feature vector (e.g. from SqueezeNet [20] and so on) and X_t to LSTM. In our objective module we use the Mean Squared Error (MSE), however, the parameter quantity of the LSTM is large and the target tracking is a small sample problem. When there is background noise or similar target in the training video, or when the training video is not enough, the obtained model may have an over-fitting problem. Regularization has been used for more than ten years before the advent of deep learning. It is mainly used to solve the over-fitting problem in machine learning, and has been fully theoretically proved from the perspective of optimization theory. We add regularization terms based on the least squares method to solve the problem that the model is easy to over-fitting, such as Equation 8.

$$L = \min \frac{1}{2} \sum_{i=1}^{n} \left(\left\| y_i - w x_i \right\|^2 \right) + \frac{\lambda}{2} \left\| w \right\|^2 \tag{8}$$

4. Experimental Results

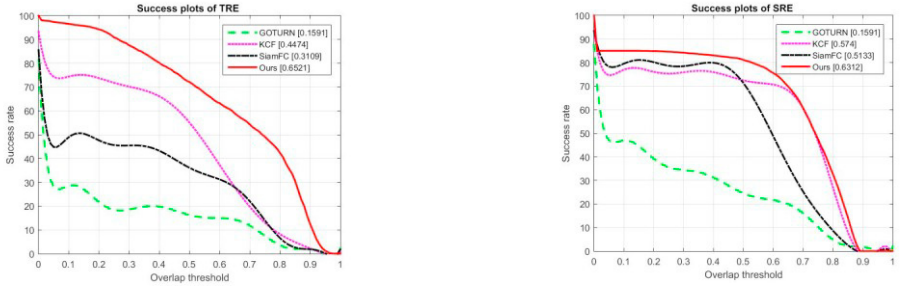
We compare our tracker with state-of-the-art trackers on the benchmark datasets [20-23] for performance evaluation. Our feature extraction network scans the whole video by applying SqueezeNet [20] that is pre-trained using imagenet [24]. We initialize the unchanged convolutional layer of the improved full-convolutional Siamese networks using the pre-trained weights of SiamFC[14], and finally we retain the improved full-convolutional Siamese networks and LSTM model with video in OTB [22] and VOT [23]. We follow the standard evaluation metrics shown in [22].



(a) (b)

Fig 3: plots of precision and success in 41 sequences using OPE on the OTB-2013 and OTB-2015.

Fig. 3 shows this evaluation results. we reported the each tracker's AUC (the area under curve) and distance precision scores in the Fig legend. Among all the trackers, our tracker performs favorably on both the distance precision and overlap success rate.



(a) (b)

Fig 4: Success Plots for TRE (temporal robustness evaluation) and SRE (spatial robustness evaluation) on the OTB

Fig. 3(a) shows the precision plots over the 41 videos in the OTB dataset. The AUC score of our tracker is improved by 18.5% compared to the KCF method. Among the compared deep feature trackers, GOTURN and SiamFC provide the results with AUC scores of 0.4838 and 0.6723 respectively. The AUC score of our algorithm is improved by 73% compared to GOTURN, and by 24.8% compared to SiamFC. If the location error threshold is 20, the accuracy of our trackers reaches 85%.Fig. 3(b) shows the success plot over the 41 videos in the OTB dataset. The AUC score of our trackers is improved by 14.9% compared KCF method that uses hand-crafted features. Among the compared deep feature trackers, GOTURN and SiamFC provide the results with AUC scores of 0.3647 and 0.5366 respectively. The AUC score of our algorithm is improved by 85.7% compared to GOTURN, and is improved by 26.2% compared to SiamFC.

In Fig.4, we shown the SRE and TRE results for robustness evaluation. As shown in Fig 3(a), the spatial robustness of our algorithm is similar to that of SiamFC and KCF. However, as we are considering the temporal continuity of the video, the temporal robustness of our tracker is improved compared to other tracker that is based on CNNs (e.g. SiamFC).

In Fig.5, we further analyzed the performance of the tracker under different types of videos. (e.g. illumination variation, occlusion or deformation) annotated in the benchmark. The results indicate that our tracker is effective in handling background clutters, deformation, motion blur, fast motion and out of view. It is mainly because the RNNs can capture historical information of the target. When the target appearance undergoes obvious changes or occlusion, existing trackers which are based solely on CNNs (e.g.GOTURN, SiamFC) cannot locate the target object accurately, specially lost targets. However, our tracker is not effective in handling low resolution. This is mainly because the global information from the Squeezenet may introduce interference information with low resolution videos. For others scenes (e.g. illumination variation, in and out plan rotation), the performance of ours tracker is similar to the SiamFC.

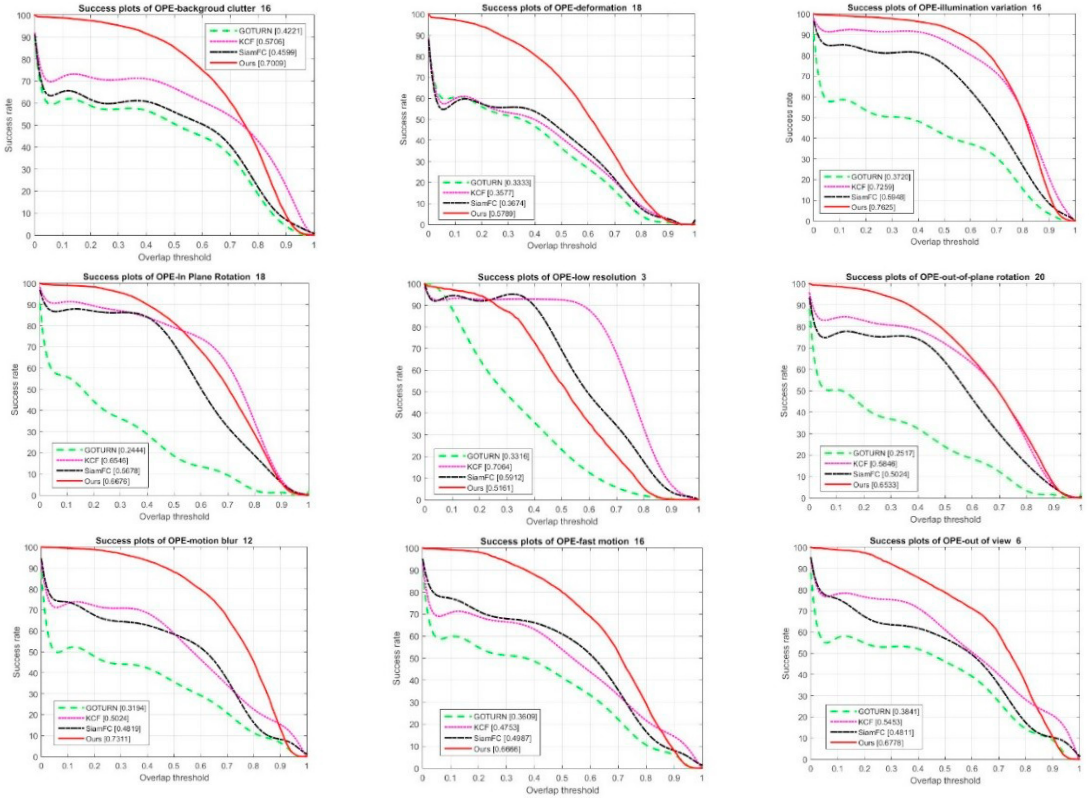


Fig 5: The success plot over nine tracking challenges, including background clutter, deformation, illumination variation, in-plan rotation, low resolution, out-of-plane rotation, motion blur, fast motion and out of view.

5. Conclusion

In this paper, we have successfully developed a new method for visual object tracking. The tracker we proposed can effectively solve the problems of main occlusion, background clutter, deformation, motion blur, fast motion and insufficient vision. Our extensive experimental results and performance comparisons with the most advanced tracking methods of challenging benchmark tracking data sets show that our trackers are more accurate. Using 41 videos from the Object Tracking Benchmark (OTB) dataset and considering the area under the curve for the precision and success rate, our method outperforms the second best by 18.5% and 14.9% respectively.

6. Acknowledgements

This work is supported in part by the Xi'an Science and Technology Plan Project (201805040YD18CG24(5)).

References

1. Yilmaz A. Object tracking: A survey[J]. Acm Computing Surveys, 2006, 38(4):13.
2.Smeulders A W M, Chu D M, Cucchiara R, et al. Visual Tracking: An Experimental Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 36(7):1442-1468.
3. Kalal Z, Mikolajczyk K, Matas J. Tracking-Learning-Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(7):1409.
4. Henriques J F, Rui C, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(3):583-596.
5. Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013:809-817.
6. Matej Kristan, Aleš Leonardis, Jiři Matas, et al. The Visual Object Tracking VOT2016 Challenge Results[J]. COMPUTER VISION - ECCV 2016 WORKSHOPS, PT II, 2016, 8926:191-217.
7. Sun C, Zhu S, Liu J. Fusing Kalman filter with TLD algorithm for target tracking[C]// Control Conference. Technical Committee on Control Theory, Chinese Association of Automation, 2015:3736-3741.
8. Danelljan M, Hager G, Khan F S, et al. Convolutional Features for Correlation Filter Based Visual Tracking[C]// IEEE International Conference on Computer Vision Workshop. IEEE Computer Society, 2015:621-629.
9. Ma C, Huang J B, Yang X, et al. Hierarchical Convolutional Features for Visual Tracking[C]// IEEE International Conference on Computer Vision. IEEE, 2015:3074-3082.
10. Wang L, Ouyang W, Wang X, et al. Visual Tracking with Fully Convolutional Networks[C]// IEEE International Conference on Computer Vision. IEEE, 2016:3119-3127.
11. Held D, Thrun S, Savarese S. Learning to Track at 100 FPS with Deep Regression Networks[J]. 2016:749-765.
12. Chen K, Tao W. Once for All: a Two-flow Convolutional Neural Network for Visual Tracking[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2016, PP(99):1-1.
13.Tao R, Gavves E, Smeulders A W M. Siamese Instance Search for Tracking[C]// Computer Vision and Pattern Recognition. IEEE, 2016:1420-1429.
14.Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[J]. 2016:850-865.
15.Valmadre J, Bertinetto L, Henriques J F, et al. End-to-end representation learning for Correlation Filter based tracking[J]. 2017.
16.Gan Q, Guo Q, Zhang Z, et al. First Step toward Model-Free, Anonymous Object Tracking with Recurrent Neural Networks[J]. Computer Science, 2015.
17. Kahou S E, Michalski V, Memisevic R, et al. RATM: Recurrent Attentive Tracking Model[C]// Computer Vision and Pattern Recognition Workshops. IEEE, 2017:1613-1622.
18.Hong Z, Chen Z, Wang C, et al. MUlti-Store Tracker (MUSTer): A cognitive psychology inspired approach to object tracking[C]// Computer Vision and Pattern Recognition. IEEE, 2015:749-758.
19. Ning G, Zhang Z, Huang C, et al. Spatially supervised recurrent convolutional neural networks for visual object tracking[C]// IEEE International Symposium on Circuits and Systems. IEEE, 2017:1-4.
20.Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[J]. 2016.
21.M. Kristan and et al. The visual object tracking vot2016 challenge results. In ECCV Workshops, 2016
22. Wu Y, Lim J, Yang M H. Object Tracking Benchmark[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1834-1848.
23. Felsberg M, Kristan M, Matas J, et al. The Thermal Infrared Visual Object Tracking VOT-TIR2016 Challenge Results[J]. 2016(ICCVW):639-651.
24. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2014, 115(3):211-252.