2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Online Learning of Spatial-Temporal Convolution Response for Robust Real-Time Tracking

Jinglin Zhou*,Rong Wang†,Jianwei Ding‡
School of Information Technology and Network Security
Peoples Public Security University of China
Beijing, China
Email: *henlouser@163.com,†dbdxwangrong@163.com,‡jwding@foxmail.com

***Abstract*—The challenges of generic visual tracking have attracted great attentions. However, it is still difficult for most of the existing trackers to track objects accurately on real-time occasion. We propose a framework which integrate a verifying mechanism and a correcting mechanism to improve the accuracy of real-time tracking. Under online learning, both target location and sample model update in parallel. Validations are carried out in every frame according to spatial-temporal convolution response. Furthermore, a correcting mechanism would be activated when the current tracking results considered to be unreliable. Synchronously, an online target model updating strategy is constructed to filter the contributive samples, which makes the sample model update confidently. The proposed tracker is evaluated on four popular benchmarks, achieving a state-of-the-art performance while runs at real-time speed.**

I. Introduction

Tracking is a problem full of challenge as the different statement of target and the variable environment would introduce numbers of interference. For example, a tracker might be good at tracking a fast moving target but failed to adapt to illumination variation. A tracker might have an advantage in scale variation but have difficulties in heavy occlusion. A significant amount of effort have been taken into generic visual tracking [1], [2], [3], nevertheless, there is still a long way to travel to tackle these issues in real-time vision applications [4].

Inspired by the great achievements in visual recognition, the convolutional neural network has attracted great attentions on visual tracking these years. Deep convolution feathers are exacted to model the samples, demonstrating the powerfulness of its representations ability. Based on the convolutional neural network, the newly proposed trackers MDNet[5] and TCNN[6] train models with annotated videos, reaching the state-of-the-art performance in existing tracking benchmarks. However, due to the limitation of huge computation and the lack of widely adaptive to different sequence, it is hard to common convolutional neural network based algorithm achieve high accuracy real-time tracking problems.

Recently, the correlation filters based algorithms are widely popular with its efficient computation and outstanding performance in visual tracking. The correlation filters transform the features into Fourier domain with Fast Fourier Transform, achieving a high-speed frame in tracking. The early correlation filters based algorithm MOSSE [7] proposed by Bolme et al. exceeding 600 frames per second with a 47.5% average accuracy rate. Since then, the internal structure of the correlation filter methods were further optimized with the performance being better and better. KCF [8] proposed by Henriques J F et al. introduced cyclic shifts and Kernel trick on the basis of MOSSE, realizing a real-time tracker with the average accuracy rate increased to more than seventy percent. Furthermore, to improve scale adaptive and edge effect of target, the translation filter and the scale filter are integrated to the architecture.

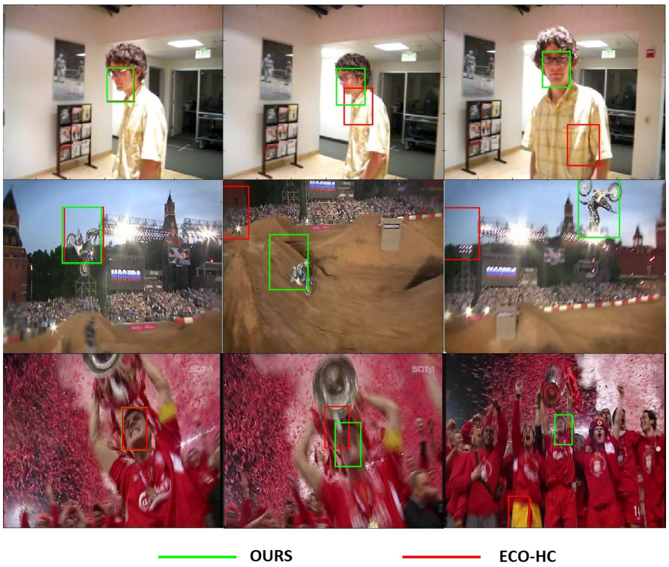


Fig. 1. A comparison of our approach with the baseline tracker ECO on three sequences from OTB2015 (from up to bottom: David, MotorRolling, Soccor). In all three cases, ECO-HC suffers from drifting caused by amplified error, along with the correct mechanism is activated by verifying with our method.

As most of the correlation filter based algorithms enjoy a high possibility of drifting under disturbance, most existing tracking algorithm above devote themselves to optimizing target model or designing advanced feathers to improve the tracking performance. We are engaged in considering the problem in another aspect, introducing a verifying and correcting mechanism to reduce the influence of interference. The idea of verifying is radically new. A target model is constructed in the classical tracker TLD [9] to evaluate the tracking components

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

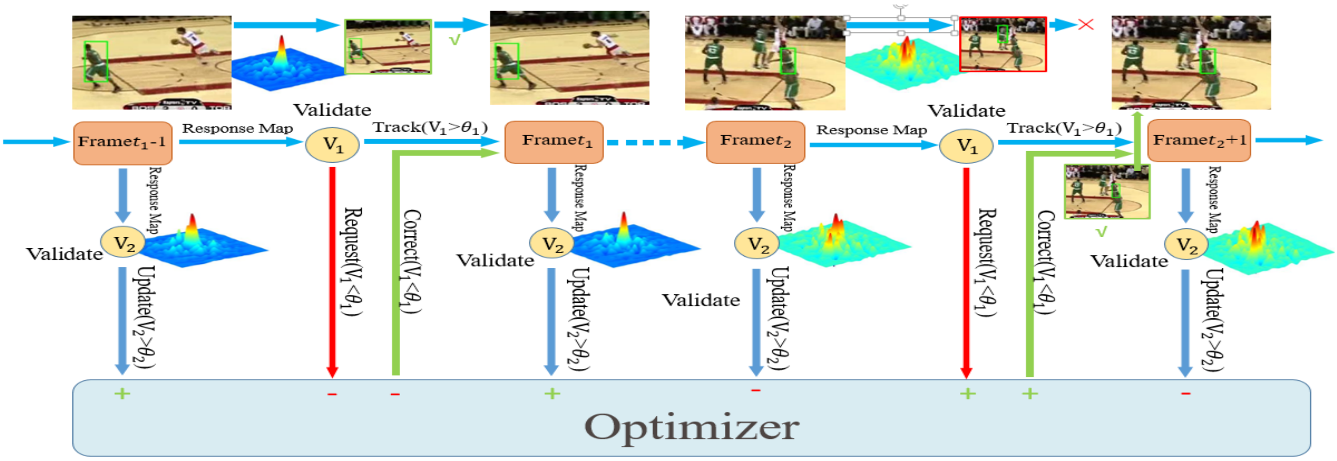


Fig. 2. The pipeline of our framework in which tracking with validations while the optimizer would be activated if satisfies certain conditions.

and update the tracker when it was considered to be necessary. The closest works to our method is LMCF [10] and PATV [11] in recent year, embedding a confidence to verify the tracking results, followed by a correct mechanism to update either target location or the sample model of target. Unlike in previous algorithm, we construct a framework to correct both target location and sample model in a high-level confidence, aiming at both high accuracy and real-time tracking. A measure is constructed to verify whether the tracking results are confidence or not and correct mechanism would be activate when drifting happened. And the sample model would update under supervision. The algorithm proposed in this paper is evaluated on four challenge benchmarks: OTB2013, OTB2015, VOT2016 and VOT2017. Our algorithm achieves the state-of-the-art accuracy while runs at speed in excess of 35 frames per second on a single CPU.

II. Our Method

A. Overview

In this section, we propose a framework organized by two components, tracking and optimization, as its pipeline is shown in figure 2. Both two components are validated by convolution response map, for real-time and high accuracy tracking. On the tracking component, multi-dimensional feathers are transformed by the Fast Fourier Transform (FFT) and a fast correlation tracking is made between two adjacent frames.

The optimizing component consist of the sample model and the memory matrix. The sample model is built with the sample region and updated when the tracking results are considered to be reliable. The memory matrix stores the tracking results and validations in every frame. With the correcting request inputting, the optimizer detect the target in the new frame by responding a correlation the last reliable tracking results and the new frame.

B. Baseline Tracker

We employ a discriminative correlation filters based tracker Efficient Convolution Operator Tracker (ECO) as our baseline [12]. There are two kinds of ECO tracker proposed by authors, one is based on deep convolutional features and the other one is only based on hand-crafted features (HOG and Color Names) named ECO-HC, which is one of the best real time trackers in VOT2017 challenge and we would employ in this paper.

The first advantage of ECO is the transformation of different resolution feature map to a continuous spatial domain. The transformation introduces a continuous function representing response, enabling precise target position. The second advantage of ECO is the efficient computation. This provided by the selection of the most contributive filters and the reduction of sample model updating frequency.

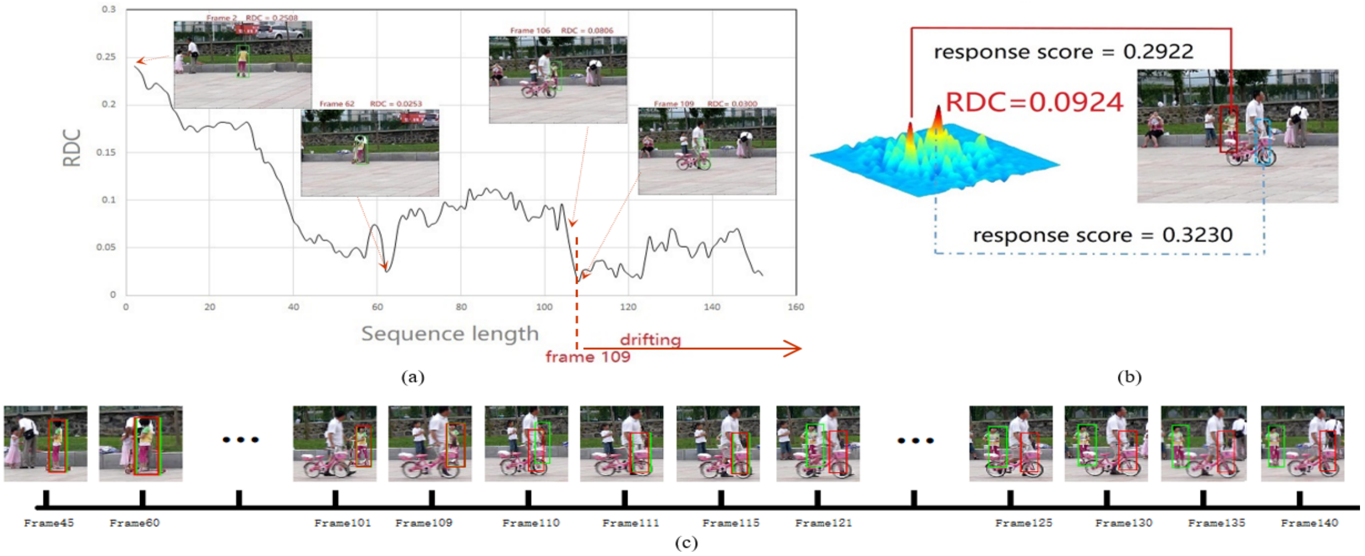
Here, we briefly describe the ECO formulation, which has a strong relationship with our verifying measures. To ensure the real-time performance, we only use the hand-crafted features including Hog and Color Names. Assuming that the different resolution feature maps have transferred to a continuous domain $t \in [0, T)$, and we denote the interpolated feature map as J. Moreover, a formulation is introduced for training convolution filters f to represent the convolution response between sample and target. Besides, a dimensional reduction is carried out to select the most contributive features with a matrix P. Above all, we could measure the convolution response with scores

$$S_f = Pf * J \tag{1}$$

Each score denotes the relation between the sample region and one region in the frame. What is more, a highest score represents the most similar region with the target in the current frame.

C. Evaluation Criterion Embedding

Most correlation filter based trackers locate object by communicating a vector between the target position of last frame and maximum correlation response position of the current frame, without considering whether the tracking result in the last frame is reliable or not. Nevertheless, the maximum



(a)

(b)

(c)

Fig. 3. An example of the relationship between the RDC and interference. Images are from the sequence Girl in VOT2017 and the green boundingboxes are the tracking results of ECO-HC tracking algorithm. As is shown in (a), the RDC would reduce under disturbance such as similar object or occlusion. For example, there is a boy next to the girl in the frame 62 with a reduction of RDC. From frame 106 to frame 109, a man with a pink bike pass in front of the girl, along with the RDC on the decline. And since a complete occlusion in frame 109, the bounding box begin to drifting. As is show in (b), the maximum response region might not be the target position when interference occurred such as occlusion. The image is the 124th frame in the sequence Girl. And (c) represents a comparison on ECO-HC and our method, where red boundingboxes represent ECO-HC and the green ones represent ours.

response region might not be the target position always, especially when some distractions exist, as is shown in the figure 3 (b). Drifting would happen frequently if tracking continues without corrective measures taken as shown in the second row of the figure 3 and 4.

Intuitively, as the convolution response scores would be influenced by interference, we try to measure the distribution of response scores to validate whether there exist interference or not. In general, the response in convolution response map agree with the Gaussian distribution.

$$S_f \sim N(\mu, \sigma^2) \tag{2}$$

Where μ denotes the expectation of response scores and σ denotes the standard deviation. Usually there is a response much stronger than the others are in each frame, which is considered to correspond to target location. Instead, there would be two or more large response in the convolution response map if there exist disturbance. Hence, the Gaussian distribution would be gentler under disturbance than that in normal, which means the value of σ would be small.

On the basis of above, we employ a novel criterion from correlation response map called it Response Distribution Criterion. The *Response Distribution Criterion* is defined as

$$RDC = \sqrt{\sum_{i=1}^{\tau}(S_f(i) - \mu)^2} \tag{3}$$

Here, S_f denotes the response scores as expression (1), and $S_f(i)$ denotes the response score which its numerical value ranking i in the response map. And the τ denotes a threshold that we only choose the top τ response scores to analysis. The μ denotes the expectation of the top τ response scores. The RDC represents the dispersion of response scores in current frame. Intuitively, with only the serval most confusing region would achieve the maximum correlation response score, we choose the top τ response scores instead of all response scores to highlight the difference. A high view of RDC indicates the selected maximal correlation response scores are spread out over a wider range of value, which means the maximal correlation response is significant and the noises is small. On the contrary, a low value of *RDC* indicates that the response scores tend to be close to each other and the response map is undulate. In other words, the possibility of existing interference is great, as is shown in figure 3 (a).

The *RDC* validates the distribution of response in tracking processes, along with the tracking results and the validations being stored. When the *RDC* is higher than a predefined threshold θ_1, the tracking result is considered reliable.

D. Tracking by Online Optimization

In most of existing algorithm, locations of target were computed frame by frame on the basis of the tracking result in the nearest frame. This way would amplify the interference introduced from one frame, for the error would be accumulated frame by frame, as shown in the second row of figure 4.

As discussed in subsection C, we could verify whether the results is reliable by response scores and *RDC*. Here we employ *RDC* to evaluate whether there exist interference. With the evaluation of existing tracking results, we propose an optimization strategy to correct the low-confidence results. Suppose that the result in frame t is considered reliable with *RDC*, we would locate the target in frame $t+1$ by computing

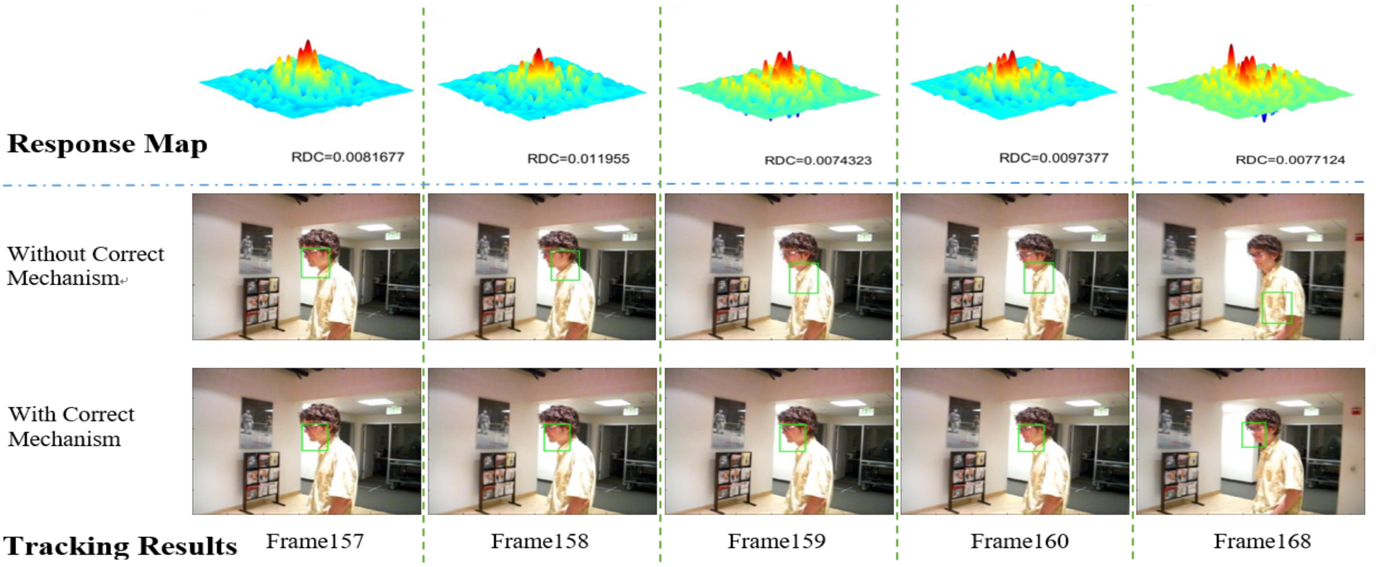


Fig. 4. The second column are the shots of sequence David from OTB2015 (from left to right and top to bottom: frame157, frame158, frame159, frame160, frame168), where the green bounding boxes represents the tracking results of ECO-HC which happened to be shifting. The fist column are the convolution response map one to one match with the frame, where the highest peak indicates the maximum response region which tracking results lactated. As is shown in the response map of frame158, there are three region enjoys a high and similar response scores and the highest one is not the perfect region. With the unreliable results of frame158, errors is amplified frame to frame and the drifting happened. The third column are the tracking reasluts after correcting.

a vector between two maximum response in frame *t* and frame $t + 1$. And along with an evaluation taken to judge whether the result is reliable. If the tracking result considered reliable, tracking would gone on similar to above. If the tracking result considered unreliable, the tracking in frame $t+2$ would on the basis of the nearest reliable result in frame t instead of frame $t + 1$. Follow up steps are analogous in turn.

An integer n is defined to record the number of frames from the last reliable result to current frame. The correcting mechanism would be activated when the location of target in the current frame t is estimated from the last confident frame $t - n$. Comparison on tracking with and without correct mechanism are shown in figure 4.

E. High-Confidence Sample Model Updating

Our tracker inherit the updating sample model strategy of the baseline tracker and make some further improvements with convolution response map. To avoid the over fitting model, the ECO tracker screens out a training set consists of representation for different aspects of appearance. This strategy have achieve perfect performing. On the basis of this strategy, a further screening is setting up a lowest limitation by *RDC*, to confirm the reliability of sample model updating. Considering the property of the moving object, the threshold θ_2 is lower than θ_1, and the model would also update if the response scores are relatively stable with the low-level confidence in a certain time.

III. Experiments

We extensively evaluate our tracker on the four most popular challenging benchmark datasets, including OTB2013 [13], OTB2015 [14], VOT2016 [15] and VOT2017 [16]. For OTB2013 and OTB2015, the comparison is carried out among 8 state-of-the-art trackers, including ECO-HC, DeepSRDCF [17], SRDCF [17], Staple [18], LCT [19], DCFnet [20], MEEM [21] and KCF [8]. For VOT2016 and VOT2017, we compare our approach with the top 10 trackers in the challenge. Both convolutional neural network based trackers and correlation filters based trackers are taken into account.

We employ the Distance Precision Rate (DPR) and the Overlap Success Rate (OSR) to evaluate trackers on OTB2013 and OTB2015. For VOT2016 and VOT2017 datasets, the performance of trackers is measured in terms of Accuracy (A) and Expect Average Overlap rate (EAO).

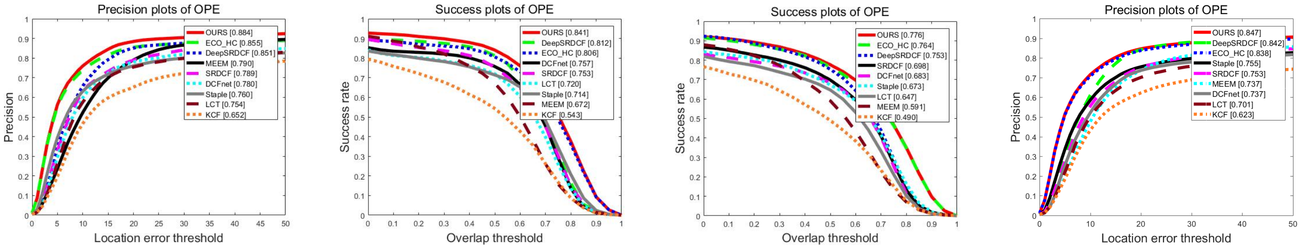
A. Implementation Details

Our tracker is implemented by Matlab 2016a and validated on a machine equipped with an Intel Core i7 running at 2.50 GHz. The thresholds of RDC θ_1 and θ_2 are set to 0.05 and 0.03, the is set to 5 and the threshold of response score is set to 0.19.

B. OTB-2015 Dataset and OTB-2013 Dataset

The OTB2013 and OTB2015 are online object tracking benchmark including 50 sequence and 100 sequence. Each sequence is annotated by 11 different attributes, which consists of illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane and out-of-plan rotation, out of view, background clutter and low resolution.

As is shown in figure 5, we report the results in one-pass evaluation (OPE) using DPR and OSR. Overall, our approach performs favorably compared with all the other state-of-the-art trackers on both OTB2013 and OTB2015. Table presents a state-of-the-art comparison in average Overlap Success Rate



(a) Comparisons on OTB2013 (b) Comparisons on OTB2013 (c) Comparisons on OTB2015 (d) Comparisons on OTB2015

Fig. 5. Comparisons with state-of-the Cart tracking algorithms on OTB2013 and OTB2015 with the DPR and OSR.

TABLE I
COMPARISONS ON OTB2013 AND OTB2015 USING DPR AT A THRESHOLD OF 20 PIXELS AND OSR AT AN OVERLAP THRESHOLD OF 0.5.

		OURS	ECO-HC	Staple	DCFnet	SRDCF	DeepSRDCF	LCT	MEEM	KCF
OTB2013	DPR	0.884	0.855	0.760	0.780	0.789	0.851	0.754	0.790	0.652
	OSR	0.841	0.800	0.719	0.752	0.746	0.803	0.723	0.669	0.551
OTB2015	DPR	0.847	0.838	0.755	0.737	0.753	0.842	0.701	0.737	0.623
	OSR	0.776	0.764	0.673	0.683	0.698	0.753	0.647	0.591	0.490

TABLE II
FOR OTB2015, AVERAGE PRECISION AND SUCCESS RATE OF OURS AND THE BASELINE TRACKER ECO ON DIFFERENT ATTRIBUTES: ILLUMINATION VARIATION (IV), OUT-OF-PLANE ROTATION (OPR), SCALE VARIATION (SV), OCCLUSION (OCC), DEFORMATION (DEF), FAST MOTION (FM), IN-PLAN ROTATION (IPR), OUT-OF-VIEW (OV), BACKGROUND CLUTTERED (BC), AND LOW RESOLUTION (LR).

	Attribute	IV	OPR	SV	OCC	DEF	MB	FM	IPR	OV	BC	LR	Overall
DPR	OURS	0.819	0.827	0.806	0.804	0.827	0.804	0.804	0.791	0.747	0.840	0.854	0.847
	ECO-HC	0.770	0.803	0.796	0.798	0.784	0.765	0.796	0.760	0.774	0.820	0.866	0.838
OSR	OURS	0.777	0.736	0.719	0.735	0.751	0.686	0.755	0.749	0.672	0.789	0.623	0.776
	ECO-HC	0.735	0.709	0.703	0.729	0.704	0.654	0.728	0.738	0.674	0.771	0.633	0.764

on both datasets. It shows that our approach outperforms other state-of-the-art trackers in both DPR and OSR. On OTB2015, our tracker achieves a DPR of 84.7% and an OSR of 77.6%, performing better than our baseline tracker compared with its DPR of 83.8% and 76.4% which ranking the second in our experiments. Although the DCFnet and DeepSRDCF employ the deep feathers, our approach performs better than it. Besides, benefiting from the efficiency of correlation filter, our approach runs at a speed of 35 frames with the real-time standard.

We further presents the performance of our approach and the baseline tracker ECO-HC under different attributes in OTB2015, as is shown in Table . In terms of DPR and OSR, our approach obtain better results under 9 out of 11 attributes. Compared with ECO-HC, our approach perform more robust owing to the correct mechanism under occlusion, illumination variation, deformation and so on. However, the approaches still have difficulties in low resolution and out-of-view, showing there is room for distinguishing normal and abnormal situations.

Compared with these trackers, our approach track the target more reliably. The correct mechanism would modify the target location even though a short drift occurred. However, with the strict strategy of sample model updating, it would miss some useful sample and could update the model timely, leading more deviation than ECO-HC tracker in the situation of low resolution and out-of-view.

C. VOT2016 Dataset and VOT2017 Dataset

Next, we validate our approach on the challenging dataset Visual Object Tracking (VOT). The VOT challenges hold every year with an updated dataset and annotations, enjoying a great reputation in the field of tracking. For VOT2016, three primary measures are used to analyze performance: accuracy (A), robustness (R) and expected average overlap (EAO). We employ two of them except for R, as our contribution mainly concentrate on the correct after interference occurred. Table and figure 6 presents a comparison to the top 8 participants in VOT2016. Our approach ranking second after the CCOT which tracking with deep features, while runs at a speed of 35 frames per second nearly ten times than it. A new challenge is added in VOT2017 called real-time challenge. Trackers are required to produce the output faster than 25 frames per seconds in the VOT2017 real-time challenge, aimed to promoting the importance of real-time applications. The comparison on the VOT2017 real-time experiment is shown in figure 7. Our approach ranks third among them.

TABLE III
COMPARISON WITH THE TOP 10 TRACKERS ON THE VOT2016 CHALLENGE. THE PROPOSED METHOD PROVIDES THE STATE-OF-THE ART ACCRACY AND ROBUSTNESS.

	CCOT	TCNN	SSAT	MLDF	Staple	DCF	EBT	SRBT	OURS
EAO	0.331	0.325	0.321	0.311	0.295	0.293	0.291	0.290	0.328
Accuracy	0.52	0.54	0.57	0.48	0.54	0.53	0.44	0.50	0.55

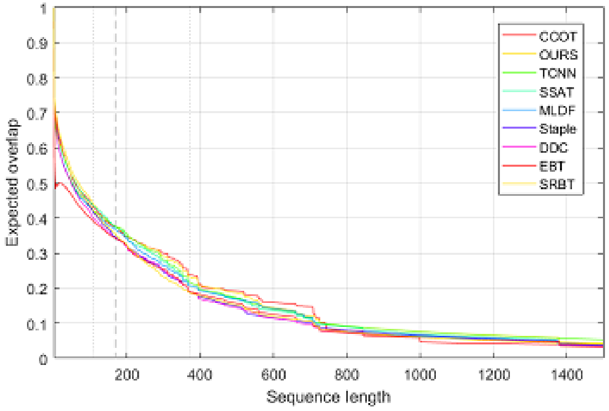


Fig. 6. Comparisons to the top 8 trackers on VOT2016 with the EAO curve.

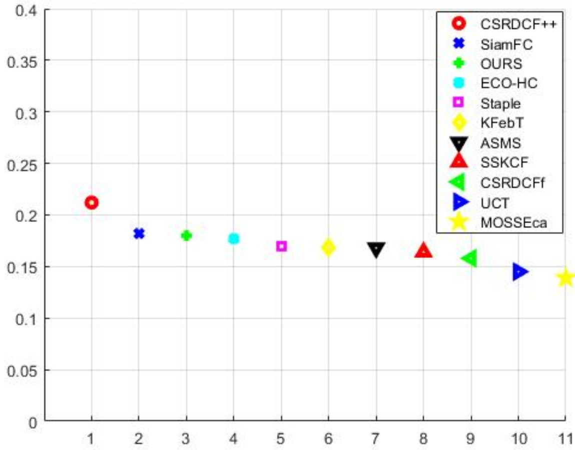


Fig. 7. The EAO plot of our approach and the top 10 trackers in challenge for the VOT2017 realtime experiment.

IV. Conclusion

In this paper, we propose a novel real-time object tracking method based on online learning. A strategy is proposed to verify the current tracking results according to spatial-temporal convolution response and consider it as a condition of learning. Furthermore, we construct an online target model updating strategy and introduce a correcting mechanism, to make the sample model updates reliably and to rectify the object location respectively. The encouraging results are demonstrated in experiments performed on four popular datasets, achieving a state-of-the-art performance both on accuracy and on speed. Further work would involve the improvement of the verification performance by introducing an online update threshold.

Acknowledgment

This work is supported by National Key Research and Development Plan under Grant No.2016YFC0801005. This work is supported by the National Natural Science Foundation of China under Grant No. 61503388.

References

[1] Galoogahi, et al. Learning Background-Aware Correlation Filters for Visual Tracking, in ICCV, 2017.
[2] Danelljan Martin, et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking, in ECCV, 2016.
[3] Mueller, et al. Context-Aware Correlation Filter Tracking, in CVPR, 2017.
[4] A. W. Smeulders, et al. Visual tracking: An experimental survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7,pp. 1442-1468, 2014
[5] Nam, Hyeonseob and B. Han. Learning Multi-domain Convolutional Neural Networks for Visual Tracking, in CVPR, 2016.
[6] Kang Kai, et al. Object Detection from Video Tubelets with Convolutional Neural Networks, in CVPR, 2016.
[7] Bolme.D.S, et al. Visual object tracking using adaptive correlation filters, in CVPR, 2010.
[8] [8] Henriques, et al. High-Speed Tracking with Kernel-ized Correlation Filter, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3,pp. 583-596, 2015.
[9] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 7,pp. 1049, 2012
[10] Wang Mengmeng, Y. Liu, and Z. Huang. Large Margin Object Tracking with Circulant Feature Maps, in CVPR, 2017.
[11] Fan, Heng, and H. Ling. Parallel Tracking and Verifying: A Framework for Real-Time and High Accuracy Visual Tracking, in ICCV, 2017.
[12] Danelljan Martin, et al. ECO: Efficient Convolution Operators for Tracking, in CVPR, 2017.
[13] Y. Wu, J. Lim, and M.-H. Yang, Online Object tracking benchmark, in CVPR, 2013.
[14] Y. Wu, J. Lim, and M.-H. Yang, Object tracking benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9,pp. 1834-1848, 2015.
[15] M. Kristan, et.al. The visual object tracking vot2016 challenge results, in ICCV workshop, 2016.
[16] M. Kristan, et.al. The visual object tracking vot2017 challenge results, in ICCV workshop, 2017.
[17] M. Danelljan, et al. Learning spatially regularized correlation filters for visual tracking, in ICCV, 2015.
[18] L. Bertinetto, et al. Staple: Complementary learners for real-time tracking, in CVPR, 2016.
[19] C. Ma, et al. Yang. Long-term correlation tracking, in CVPR, 2015
[20] Wang Qiang, et al. DCFNet: Discriminant Correlation Filters Network for Visual Tracking, ArXiv, 2017.
[21] J. Zhang, S. Ma, and S. Sclaroff. Meem: robust tracking via multiple experts using entropy minimization, in ECCV, 2014.