Dynamic Fusion of Convolutional Features based on Spatial and Temporal Attention for Visual Tracking

Dongcheng Zhao[1,2], Yi Zeng[1,2,3,4]
[1]*Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[2]*University of Chinese Academy of Sciences, Beijing, China*
[3]*Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China*
[4]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
{zhaodongcheng2016, yi.zeng}@ia.ac.cn

***Abstract*—Convolutional neural networks (CNN) based trackers have been widely employed in visual object tracking due to their powerful representations. Features from different CNN layers encode different information. Deeper layers contain more semantic information, while the resolution is too coarse to localize the target. Shallower layers carry more detail information but are less robust for appearance variations. In this paper, we propose an algorithm which incorporates the Spatial and Temporal attention to take full advantage of the Hierarchical Convolutional Features for Tracking (STHCFT). We firstly learn correlation filters on each convolutional layer. Based on the spatial attention inspired by the paraventricular thalamus (PVT) in the brain, we choose the most important layer to build the base response, and the others to be the auxiliary responses. In addition, we make full use of the temporal attention to determine the weights of the auxiliary responses. Finally, the target is located by the maximum value of the fused responses. Extensive experimental results on the benchmark OTB-2013 and OTB-2015 have shown the proposed algorithm performs favorably against several state-of-the-art trackers.**

I. Introduction

Visual object tracking is a fundamental and essential cognitive function for human and machine perception, and has various applications such as video surveillance [1], human-computer interaction [2], and human motion analyses [3]. In this paper, we consider single object tracking which continuously localizes a target in a video-sequence given a target bounding box in the first frame. The main difficulty of this problem is how to build a tracker that can tolerate various critical situations, such as scale variation, fast motion, and background clutters, etc.

Recently, discriminant correlation filters (DCF) based trackers [4]–[8] have shown state-of-the-art performance in the visual object tracking benchmark [9], which have attracted extensive attention. The DCF trackers train a regressor by exploiting the properties of circular correlation. By using Fast Fourier Transform (FFT), the trackers can perform fast in the Fourier domain. However, most of the DCF based trackers use hand-craft features, which have limited the performance to some extent. The Convolutional Neural Network (CNN) has achieved outstanding performance on many computer vision tasks, such as image classification [10], image segmentation [11], and face recognition [12]. Different from hand-craft features, the features of CNN are obtained by the model automatically and contain more semantic information. Therefore, it is of great interest to apply convolutional features for DCF-based tracking.

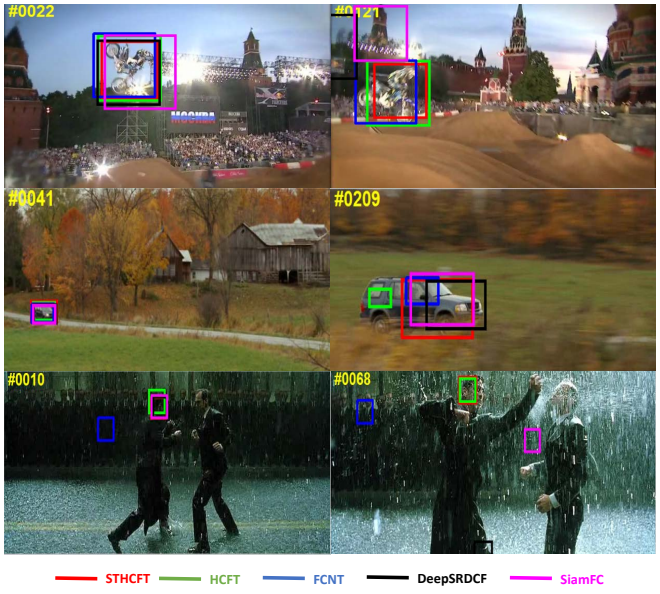

Fig. 1. Comparisons of our STHCFT with four CNN based state-of-the-art trackers in the changing scenario.

A deep convolutional neural network consists of several convolutional layers. The deeper convolutional layers capture more abstract and semantic information and have been successfully employed for image classification. However, losing more details due to the low spatial resolution has made them not so discriminative to the objects with similar appearances in the same category. The shallower layers provide higher spatial resolution, which is crucial for accurate target localization. However, they are less robust to the appearance variations such as deformation and occlusion of objects. So it is of great importance to fuse multiple convolutional layers for visual tracking.

To address the problem, we propose a model to dynamically fuse convolutional features for visual tracking based on spatial

978-1-7281-2009-6/$31.00 ©2019 IEEE paper N-19324.pdf

and temporal attention. As can be seen in Fig.1, our tracker has a better performance compared with the four state-of-the-art CNN based trackers. The contributions of this paper are summarized below:

- We introduce a spatial attention mechanism to dynamically choose one convolutional layer to build the base response map and the others to be the auxiliary response maps.
- We introduce a temporal attention mechanism to determine the weight of each auxiliary response map.
- The hierarchical convolutional response maps are dynamically fused by the spatial and temporal attention (STHCFT), which performs favorably against existing state-of-the-art methods.

II. Related Work

A. DCF based trackers

Since the discriminative correlation filters have achieved great success in video object tracking, many extensions have been made to further improve its performance. [13] proposed an adaptive correlation filter by Minimizing the Output Sum of Squared Error (MOSSE) based on the gray-scale feature, which can be considered as the first one to introduce correlation filters into tracking. To further improve the performance, CSK [14] introduced the kernel trick into the DCF framework; Furthermore, CSK framework was extended with multi-channel feature input based on Gaussian kernel [15] and color feature [16]. [17] proposed a framework to integrate the powerful features including HOG and color-naming together to further boost the performance.

In addition to the features, the unwanted boundary effects were produced due to the basic periodic assumption in DCF based trackers, which will produce some synthetic examples compared to the real sample. This caused degradation on the performance of the standard DCF trackers. To address this problem, LBCF [18] enlarged the search region to allow the training signal to be a larger size than the filter, which would reduce the synthetic samples, and with the Alternating Direction Method of Multipliers (ADMM), the closed-form solution could be obtained. The CACF [21] tracker took the global context into account and incorporated it directly into the learned filter. In SRDCF [19], Dan et al. introduced a spatial regularization component in the DCF tracker to penalize correlation filter coefficients depending on their spatial location. In CSRDCF [20], the channel and spatial reliability concepts were introduced into DCF tracking, which adjusted the filter to support the part of the object suitable for tracking.

Apart from enriching features and relaxing the boundary effects, researchers take other aspects into consideration. For example, scale estimation, [22] proposed a framework by learning discriminative correlation filters based on a scale pyramid representation. Improving the training sample, [23] down-weighted the corrupted samples while increasing the impact of correct ones by estimating the quality of the samples. For ensemble methods, [24] equipped a basic framework with two KCF trackers to cope with the complex surrounding environment and large appearance variations. However, all these methods use handcraft features, which hinder their accuracy and robustness to a certain extent.

B. CNN features based trackers

Inspired by the great success of CNN in object recognition, researchers in the field of object tracking have been studying how to apply CNN in tracking. There are several works to utilize the features of CNN to further improve the performance of DCF based framework. DeepSRDCF [25] replaced the HOG and Color features with shallow CNN features in SRDCF framework. CREST [26] integrated feature extraction and the correlation filter into an end-to-end framework, as well as, in order to further utilize the features of CNN, redisual learning was introduced. FlowTrack [27] took advantage of flow information of consecutive frames with convolutional neural networks to improve the feature representation. These works only take advantage of one layer feature of CNN, while different convolutional layers of CNN encode different levels of information, lower layers provide more precise localization but less robust to the deformation and occlusion of objects and deeper layers encode semantic information which is robust to the appearance variations. Therefore, combining multiple layers for visual tracking is of great importance.

There are several works combining the different layers of CNN for tracking, such as [28]–[33]. Although HCFT [28] adaptively learned correlation filters on each convolutional layer and hierarchically infer the maximum response of each layer to locate targets, the weight of each layer is predefined manually. The FCNT [29] introduced a distractor detection scheme to dynamically choose the result of different layers of CNN for the final tracking result and only one layer to be chosen without considering the dynamic fusion of different layers. The HDT [30] considered the correlation filter based on each convolutional features as a weak tracker and hedged them with a dynamic weight to form a stronger tracker while ignoring the characteristics of each layer. C-COT [31] and ECO [32] employed an implicit interpolation model to efficiently integrate multi-resolution deep feature maps to solve the problem in the continuous spatial domain, where ECO is an improved version of C-COT in performance and speed. Both of them treated the features from each layer equally for different conditions. [33] extended the ECO framework with the weighted convolution responses from each feature block, while the best weight is tried out manually. To sum up, a model to dynamically fuse different convolutional features of CNN has so far been rarely studied.

III. The STHCFT Tracking Model

The pipeline of the proposed model is shown in Fig.2. We will introduce the correlation filter, the spatial and temporal attention mechanisms we used in the work. And we will also give some explanations on related inspirations from Neuroscience.

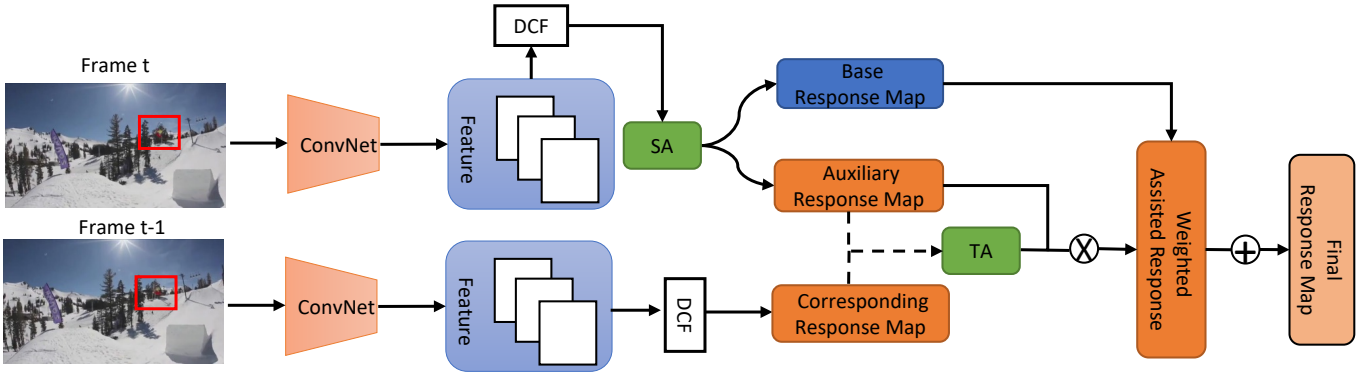

Fig. 2. The architecture of the proposed STHCFT tracker. SA: Spatial Attention. TA: Temporal Attention. The cropped image is sent into the ConvNet. And the correlation filters are learned on each convolutional layer. According to the spatial attention, we can get the base response. The temporal attention determines the weight of each auxiliary response map.

A. Correlation Filters

Firstly, we revisit the conventional correlation filter, which predicts the position of the object by the maximum value of the correlation response map based on the learned discriminative classifier. In this work, we denote $X^k \in R^{W\times H\times D}$, where W, H, and D represent the width, height, and the number of channels of the feature of the k^{th} convolutional layer. We consider all the cyclic shifts $w_{w,h}$, $(w,h) \in \{0,...,W-1\} \times \{0,1,...H-1\}$ as all the training examples for the classifier. The regression targets y follow a Gaussian function $y(w,h) = e^{-\frac{(w-W/2)^2+(h-H/2)^2}{2\delta^2}}$ with value 1 to represent the center target. A goal of training a correlation filter is achieved by minimizing the loss function:

$$W^k = \arg\min_W ||F(Y) - F(X^k) \bullet W||_F^2 + \lambda||W||_F^2 \quad (1)$$

The $\bullet$ is a linear kernel in the Hilbert space:

$$F(X^k) \bullet W = \sum_{d=1}^{D} F(X^k)_{*,*,d} \odot W_{*,*,d} \quad (2)$$

$\lambda \geq 0$ is used to control the impact of the regularization term. $F(.)$ indicates the Fast Fourier Transform (FFT). The subscript d represents the d^{th} channel, the operator $\odot$ represents the point-wise product.

Equation (1) has a simple closed solution, which can be quickly computed in the Fourier domain. The learned filter in the frequency domain on the d^{th} channel can be written as:

$$W^k_{*,*,d} = \frac{F(Y) \odot \overline{F}(X^k)_{*,*,d}}{F(X^k) \bullet \overline{F}(X^k) + \lambda} \quad (3)$$

$W^k_{*,*,d}$ is the filter of the d^{th} channel of the k^{th} layer. $\overline{F}$ represents the complex conjugate of the Fourier transform.

For the detection process, we crop a search patch, and obtain the features in the k^{th} convolutional layer: T^k, and the actual response can be computed as:

$$R^k = F^{-1}(\overline{F}(T^k) \bullet W^k) \quad (4)$$

F^{-1} indicates the inverse transformation of the discrete Fourier transform.

And the location is detected by finding the maximum response score:

$$(x^k, y^k) = \arg\max_{x',y'} R^k(x', y') \quad (5)$$

B. Spatial Attention for Choosing Base Response

Using the correlation filter, we can get the response of each layer. The peak of this response map and the degree of oscillation partly indicate the reliability of this response. Intuitively, a response map, as shown in Fig.3, the higher the peak, the smoother the surrounding, more trustworthy the result is. If a response map has multiple response peaks, then the credibility of the response to distinguish between the target and the background is very low.

Here, we introduce the average peak-to-correlation energy (APCE) used in LMCF [34] to measure the fluctuated degree of response maps and the confidence level of the detected targets. It is defined as:

$$APCE = \frac{|F_{max} - F_{min}|^2}{mean(\sum_{w,h}(F_{w,h} - F_{min})^2)} \quad (6)$$

where F_{max}, F_{min}, $F_{w,h}$ denote the maximum, minimum and the w-th row h-th column elements of R^k. For sharper peaks and smoother responses, the APCE will become larger, indicating that the response is trustworthy to be the base response map.

C. Temporal Attention for Auxiliary Response Weight

In addition to determining base response with APCE, since multiple peaks will occur in the other response maps, the weight determined by the spatial information cannot be trusted. At this time, the temporal attention mechanism is introduced to determine the weight. Considering the continuity between frames, the response distribution of two frames should be similar after moving the maximum value to the same position. The temporal attention enables the response map having major changes down-weighted while increasing the impact of others.

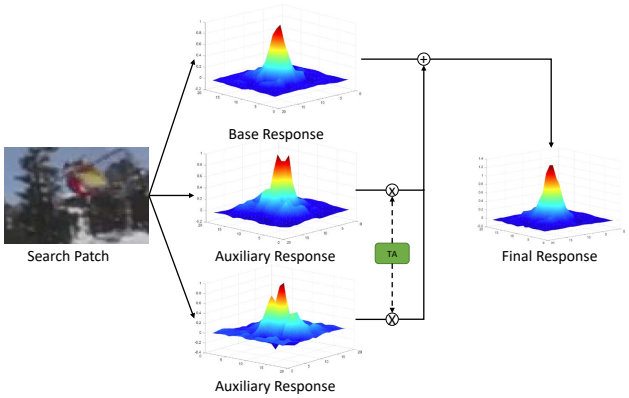

Fig. 3. Illustration of the dynamic fusion of multiple response maps from different convolutional layers in sequence Skiing in OTB-2015. The more oscillating the response map, the worse it is.

Here we use the temporal attention used in [43] which is shown as:

$$TAW^k = \frac{1}{||R_t^k - R_{t-1}^k \oplus \Delta||_2^2 + \eta} \tag{7}$$

where $\oplus$ denotes a shift operation of the confidence map, and Δ means the corresponding shift of maximum value in confidence maps from frame t-1 to t. η is set to prevent the denominator from being 0. After normalization, we can get the final weight shown in (8).

$$TAW^k = \frac{TAW^k}{\sum_{k=1}^{N} TAW^k} \tag{8}$$

D. Dynamic Fusion of Weighted Auxiliary and Base Response

We can get the final response map based on the base response and the weighed auxiliary responses, which is shown in (9) :

$$R^{total} = \sum_i \omega_i R_i^{auxiliary} + R^{base} \tag{9}$$

In Fig.3, we can see that the dynamic fusion of multiple layers will choose a better base response, also the weighted auxiliary responses will help to further improve the base response map for tracking.

E. Model Update and Scale Estimation

For every frame, we use a common linear interpolation with the history model to update the model. We denote the numerator in (6) as A^d and the denominator as B^d. The updating process can be written as (10):

$$\begin{aligned} A_t^d &= (1-\beta)A_{t-1}^d + \beta F(Y) \odot \overline{F}(X^k)_{*,*,d} \\ B_t^d &= (1-\beta)B_{t-1}^d + \beta F(X^k) \bullet \overline{F}(X^k) \\ W_t^d &= \frac{A_t^d}{B_t^d + \lambda} \end{aligned} \tag{10}$$

As for the target scale estimation, we follow the DSST tracker [22].

An overview of the proposed model is summarized in Algorithm 1.

Algorithm 1 The proposed STHCFT tracking algorithm

Input:

Frames $\{I_t\}_1^T$ and the initial bounding box;

Output:

Target locations of each frame $\{P_t\}_2^T$;

1: **repeat**
2: Crop an image from the frame I_t at the last location p_{t-1} and send it into the VGG-19 to get the corresponding features X^k;
3: Using the correlation filter trained with (1), the respective response R^k is obtained;
4: Base response R^{base} is obtained by using (6);
5: The weight ω_k of auxiliary response is obtained by using (7) and (8);
6: Fuse the base response and the weighted auxiliary with (9);
7: Estimate the target location p_t with (5) and the scale of the target as [22];
8: Crop an image patch at p_t and extract the convolutional features to get the new correlation filter
9: Update the correlation filter with (10) and the scale estimation model as [22];
10: **until** End of video sequences

F. Explanation from neuroscience

Judging the importance of information is an advanced brain function, which helps people to better adapt to the changing environment. In addition to the fixed physical properties of the sensory input, such as the color or brightness, the behavioral relevance makes a great contribution to the attention. It is a relative property that depends on past experience, current homeostatic state, and behavioral context. The thalamus is composed of several distinct subnuclei which are different from anatomy and function. Among them, the paraventricular thalamus (PVT) is particularly suitable for integrating information that is applicable to behavioral relevance which allow the brain to access the importance of events to make appropriate choices [35].

Since there are bidirectional connections between the thalamus and many other brain regions, in addition, CNNs are usually used to simulate the transmission of information between different visual cortex, inspired by the function of the PVT, we propose a model to choose the most important layer dynamically as the base response and the others as the auxiliary responses whose weights are determined by the temporal attention.

IV. Experiments

A. Experimental Setups

Implementation Setup. For the convolutional features, here we use imagenet-verydeep-19 (VGG-19) trained on ImageNet, the last three convolutional layers are used to capture the appearance of the target. The spatial resolution will gradually reduce due to the pooling operation. To get the same resolution

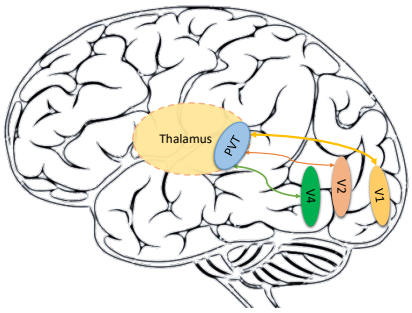

Fig. 4. The PVT area in the brain helps to modify the attention to find the most important feature and determine the weights of the other features.

for further fusion, here we use the bilinear interpolation to resize each feature map to a fixed larger size. Here we use the conv3-4, conv4-4 and conv5-4 layer and resize the conv4-4 and conv5-4 the same size as conv3-4, which is 1/4 size of the input.

The parameters for each convolutional layer are set the same, $\lambda = 0.0001$, $\eta = 0.1$, and $\beta = 0.01$.

Also, to avoid the boundary effects, each feature of the convolutional layer is weighed by a cosine window.

We implement our experiment with a PC with an Intel I7-5820k 3.30GHZ CPU, 16GB RAM, and a Geforce GTX Titan GPU. The Matconvnet is used for the computation forward computation of the VGG-19 to get the feature map.

Benchmarks. We implement the experiment on the OTB-2013 [44] and OTB-2015 [45] benchmark datasets which contain 50 and 100 sequences respectively. They are annotated with 11 attributes which cover various challenging factors, including scale variation (SV), occlusion (OCC), illumination variation (IV), motion blur (MB), deformation (DEF), fast motion (FM), out-of-view (OV), in-plane rotation (IPR) and low resolution (LR).

Evaluation Metrics. We evaluate the proposed method with the one-pass evaluation (OPE) with precision and success plot metrics on OTB-2013. In addition to the accuracy, we also evaluate the spatial and temporal robustness on OTB-2015 with temporal robustness evaluation (TRE) and spatial robustness evaluation (SRE). The precision score measures the rate of the distance between the estimated position and the ground-truth within a certain threshold. The threshold we set here is 20. The success plot measures the overlap ratio between the estimated bounding box and the real bounding box, and the success score is the area under the curve (AUC) of the success plot. The threshold we set here is 0.5.

B. Ablation Studies

Our tracking algorithm is composed of the correlation filter formed by the features of each convolutional layer of the CNN, the spatial attention and the temporal attention mechanism. In this section, we conduct ablation analysis to analyze the effects of spatial attention and temporal attention mechanism.

We implement the experiment analysis on the OTB-2013 dataset. Firstly, both the spatial attention and temporal attention are removed from the STHCFT, we simply add the original response maps from the convolutional layers after bilinear interpolation; Then, the temporal attention mechanism is removed and only the spatial attention mechanism is remained to determine the base response map, and the weight of the rest response map, the weighted fused response map is used for the final tracking; Last, we remove the spatial attention mechanism and only the temporal attention mechanism is used to determine the base response map and the weight of the rest response map, and the weighted responses are added for the final tracking.

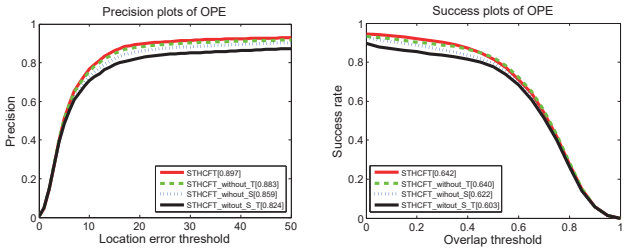

Fig. 5. The precision plot and success plot using OPE on the OTB-2013 dataset. With the integration of spatial attention and temporal attention, the performance of the tracker is improved gradually.

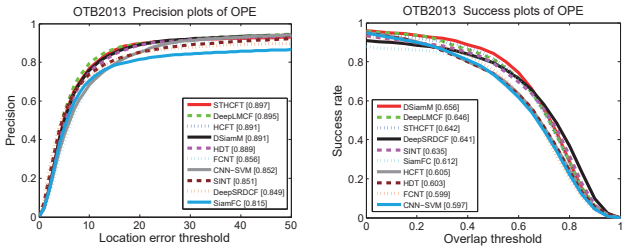

Fig. 6. The precision and success plot on OTB-2013 using OPE on comparing STHCFT with other state-of-the-art CNN based trackers.

Fig.5 shows the quantitative evaluation under AUC and average distance precision scores. We can see that the spatial attention and temporal attention will have a great improvement compared with the simply added response map.

C. Comparison with the state-of-the-art trackers

We conduct quantitative and qualitative evaluations of the benchmark datasets including OTB-2013, OTB-2015. The details are discussed in the following.

Quantitative Evaluation:

OTB-2013: We compare our STHCFT tracker with the state-of-the-art CNN based trackers: DsiamM [46], HCFT [28], HDT [30], DeepSRDCF [25], CNN-SVM [36], SINT [37], SiamFC [38], FCNT [29], DeepLMCF [34].

Fig.6 demonstrates the precision plot and the success plot with the 9 CNN based trackers on 50 sequences. The proposed STHCFT outperforms all the other trackers in terms of precision score. In the success score, the STHCFT is the third place.

In DsiamM, in addition to the conventional features, it introduces the target appearance variation and background suppression. In DeepLMCF, a special update mechanism is

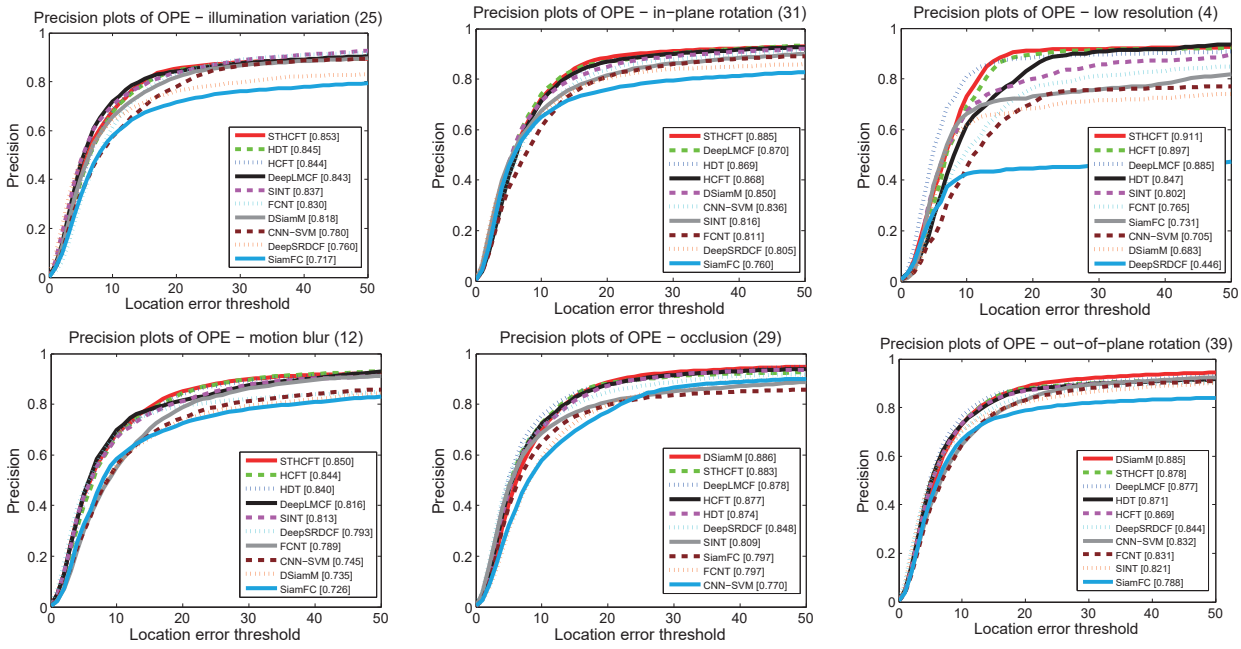

Fig. 7. The precision plot over six tracking challenges, including illumination variation, in-plane rotation, low resolution, motion blur, occlusion, out-of-plane rotation on OTB-2013 using OPE comparing STHCFT with other state-of-the-art CNN based trackers.

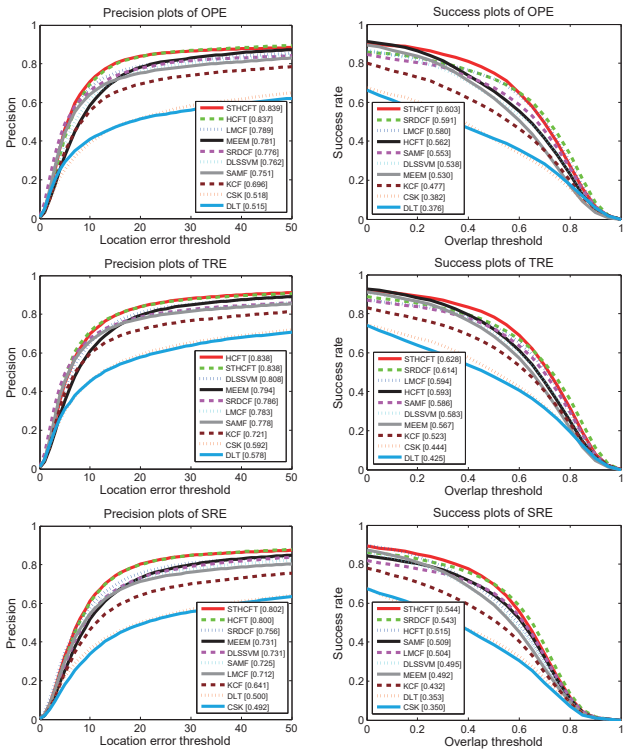

Fig. 8. The precision and success plot on OTB-2015 using OPE, SRE, and TRE on comparing STHCFT with other 9 state-of-the-art trackers.

introduced. So both of them perform better in success plot. Our work can be considered as an improvement of HCFT which makes the weight of each convolutional layers fixed. In Fig.6, we can see both precision plot and success plot, our work has an improvement, especially for success plot, our STHCFT has an improvement of 5%.

Also to facilitate better analysis of the tracking performance, we also show the one pass evaluation on the precision score under different attributes such as illumination variation, in-plane rotation in Fig.7. The results show that our STHCFT is effective in dealing with illumination variation motion blur, the reason is that STHCFT could dynamically fuse different convolutional features for tracking. When the appearance of the object changes greatly, the weight of the deep layers will be increased, when the object is similar to the background, it will choose the lower layer as the base response. In occlusion and out-of-plane rotation, STHCFT does not perform as well as DsiamM, it is because the introduction of a target appearance variation and background suppression in DsiamM, it will alleviate the interference of candidates from background as well as handling the appearance variation.

OTB-2015: We also compare STHCFT tracker on the OTB-2015 benchmark with the 9 state-of-the-art trackers, including DLT [39], CSK [14], LMCF [34], SRDCF [19], HCFT [28], DLLSVM [40], MEEM [41], KCF [15], SAMF [42]. The DLT, HCFT, DLSSVM are CNN-based trackers. The CSK, KCF, SAMF, SRDCF, LMCF are hand-craft features correlation filter based trackers. The MEEM is representative tracking algorithm.

As shown in Fig.8, STHCFT shows the best tracking accuracy and robustness in all one-pass evaluation (OPE), temporal robustness evaluation (TRE) and spatial robustness

Fig. 9. Qualitative results on comparing STHCFT with other trackers in 12 challenging sequences in OTB2015. From left to right and top to down are Basketball, KiteSurf, Bird1, Matrix, CarScale, Shaking, Diving, Skiing, Football, Soccer, Ironman, Freeman4.

evaluation (SRE) using the distance precision rate at 20 pixels, overlap success rate at 0.5.

Qualitative Evaluation: Fig.9 shows some results of the top performance trackers: DLT, HCFT, KCF, SRDCF, STHCFT. KCF and SRDCF are correlation filter trackers based on hand-craft features, DLT uses the conventional features, HCFT and STHCFT use the multi-layer futures while the weight of HCFT for different layers is fixed.

KCF and SRDCF perform well on illumination variation and fast motion (Basketball, Football), however, due to the limitation of the hand-craft features, they fail when the object has a large appearance changed, such as deformation (Matrix), motion blur (Ironman) and occlusion (Soccer). Although the DLT tracker uses the CNN features, its failure to integrate multiple features makes it lack of rich semantic information, which leads to the failure in in-plane-rotation, out-of-plane-rotation (KiteSurf, Skiing). Because of the fusion of multiple CNN features, the HCFT and STHCFT perform well on most of the sequences, however, because of the weights of the HCFT are fixed for different CNN layers, for some chosen challenging sequences such as CarScale, Shaking, it may perform worse than STHCFT. STHCFT tracker integrates the spatial information and temporal information to dynamically choose the base response and the weight of the auxiliary response to further take advantage of the different CNN features to improve the performance of tracking.

V. Conclusion

In this paper, we propose STHCFT to make full use of the different CNN features for tracking. We dynamically choose the most important feature to constitute the base response filter based on the spatial information inspired by the PVT in the brain. In addition, we dynamically determine the weight of other layers based on the temporal information to get the weighted auxiliary response map. Finally, we fuse the base response map and the weighted auxiliary response map for the final tracking. Experiments on the standard benchmark OTB-2013 and OTB-2015 indicate that STHCFT tracker performs favorably against state-of-the-art trackers.

ACKNOWLEDGEMENT

This study is supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDB32070100), the Beijing Municipality of Science and Technology (Grant No. Z181100001518006). the CETC Joint Fund (Grant No. 6141B08010103), and the Major Research Program of Shandong Province 2018CXGC1503.

References

[1] Emami, Ali, et al. “Role of spatiotemporal oriented energy features for robust visual tracking in video surveillance.” Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on. IEEE, 2012.

[2] Rautaray, Siddharth S., and Anupam Agrawal. “Vision based hand gesture recognition for human computer interaction: a survey.” Artificial Intelligence Review 43.1 (2015): 1-54.

[3] Wang, Liang, Weiming Hu, and Tieniu Tan. "Recent developments in human motion analysis." Pattern recognition 36.3 (2003): 585-601.

[4] Bolme, David S., et al. "Visual object tracking using adaptive correlation filters." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[5] Danelljan, Martin, et al. "Learning spatially regularized correlation filters for visual tracking." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[6] Danelljan, Martin, et al. "Accurate scale estimation for robust visual tracking." British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press, 2014.

[7] Zhang, Kaihua, et al. "Fast visual tracking via dense spatio-temporal context learning." European Conference on Computer Vision. Springer, Cham, 2014.

[8] Bertinetto, Luca, et al. "Staple: Complementary learners for real-time tracking." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

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

[10] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." Proceedings of the IEEE international conference on computer vision. 2015.

[11] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2018): 834-848.

[12] Lawrence, Steve, et al. "Face recognition: A convolutional neural-network approach." IEEE transactions on neural networks 8.1 (1997): 98-113.

[13] Bolme, David S., et al. "Visual object tracking using adaptive correlation filters." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[14] Henriques, Joo F., et al. "Exploiting the circulant structure of tracking-by-detection with kernels." European conference on computer vision. Springer, Berlin, Heidelberg, 2012.

[15] Henriques, Joo F., et al. "High-speed tracking with kernelized correlation filters." IEEE Transactions on Pattern Analysis and Machine Intelligence 37.3 (2015): 583-596.

[16] Danelljan, Martin, et al. "Adaptive color attributes for real-time visual tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[17] Li, Yang, and Jianke Zhu. "A scale adaptive kernel correlation filter tracker with feature integration." European conference on computer vision. Springer, Cham, 2014.

[18] Kiani Galoogahi, Hamed, Terence Sim, and Simon Lucey. "Correlation filters with limited boundaries." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[19] Danelljan, Martin, et al. "Learning spatially regularized correlation filters for visual tracking." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[20] Lukezic, Alan, et al. "Discriminative Correlation Filter with Channel and Spatial Reliability." CVPR. Vol. 6. 2017.

[21] Mueller, Matthias, Neil Smith, and Bernard Ghanem. "Context-aware correlation filter tracking." Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 2. No. 3. 2017.

[22] Danelljan, Martin, et al. "Accurate scale estimation for robust visual tracking." British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press, 2014.

[23] Danelljan, Martin, et al. "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[24] Zhang, Le, and Ponnuthurai Nagaratnam Suganthan. "Robust visual tracking via co-trained Kernelized correlation filters." Pattern Recognition 69 (2017): 82-93.

[25] Danelljan, Martin, et al. "Convolutional features for correlation filter based visual tracking." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015.

[26] Song, Yibing, et al. "Crest: Convolutional residual learning for visual tracking." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.

[27] Zhu, Zheng, et al. "End-to-end flow correlation tracking with spatial-temporal attention." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[28] Ma, Chao, et al. "Hierarchical convolutional features for visual tracking." Proceedings of the IEEE international conference on computer vision. 2015.

[29] Wang, Lijun, et al. "Visual tracking with fully convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.

[30] Qi, Yuankai, et al. "Hedged deep tracking." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[31] Danelljan, Martin, et al. "Beyond correlation filters: Learning continuous convolution operators for visual tracking." European Conference on Computer Vision. Springer, Cham, 2016.

[32] Danelljan, Martin, et al. "ECO: Efficient Convolution Operators for Tracking." CVPR. Vol. 1. No. 2. 2017.

[33] He, Zhiqun, et al. "Correlation Filters with Weighted Convolution Responses." ICCV Workshops. 2017.

[34] Wang, Mengmeng, Yong Liu, and Zeyi Huang. "Large margin object tracking with circulant feature maps." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA. 2017.

[35] Zhu, Yingjie, et al. "Dynamic salience processing in paraventricular thalamus gates associative learning." Science 362.6413 (2018): 423-429.

[36] Hong, Seunghoon, et al. "Online tracking by learning discriminative saliency map with convolutional neural network." International Conference on Machine Learning. 2015.

[37] Tao, Ran, Efstratios Gavves, and Arnold WM Smeulders. "Siamese instance search for tracking." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[38] Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." European conference on computer vision. Springer, Cham, 2016.

[39] Wang, Naiyan, and Dit-Yan Yeung. "Learning a deep compact image representation for visual tracking." Advances in neural information processing systems. 2013.

[40] Ning, Jifeng, et al. "Object tracking via dual linear structured SVM and explicit feature map." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[41] Zhang, Jianming, Shugao Ma, and Stan Sclaroff. "MEEM: robust tracking via multiple experts using entropy minimization." European Conference on Computer Vision. Springer, Cham, 2014.

[42] Li, Yang, and Jianke Zhu. "A scale adaptive kernel correlation filter tracker with feature integration." European conference on computer vision. Springer, Cham, 2014.

[43] Liu, Ting, Gang Wang, and Qingxiong Yang. "Real-time part-based visual tracking via adaptive correlation filters." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

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

[45] Wu, Yi, Jongwoo Lim, and Ming-Hsuan Yang. "Object tracking benchmark." IEEE Transactions on Pattern Analysis and Machine Intelligence 37.9 (2015): 1834-1848.

[46] Guo, Qing, et al. "Learning dynamic siamese network for visual object tracking." The IEEE International Conference on Computer Vision (ICCV).(Oct 2017). 2017.