SPATIALLY ATTENTIVE CORRELATION FILTERS FOR VISUAL TRACKING

Huai Qin[1], *Zhixiong Pi*[1], *Changqian Yu*[1], *Changxin Gao*[1], *Jin-Gang Yu*[2], *Nong Sang*[1]

[1]Key Laboratory of Ministry of Education for Image Processing and Intelligent Control,
School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China.
[2]South China University of Technology, Guangzhou 510640, China.

ABSTRACT

Although correlation filter based trackers have recently demonstrated excellent performance, they still suffer from the boundary effects. The cosine window is introduced to alleviate the boundary affects, which however may result in poor performance in case of occlusion or fast motion. To address this problem, we propose a simple yet effective framework, which builds a spatially attentive model with multiple features to guide the detection of the correlation filter based trackers. The proposed method not only can breakthrough the spatial extent of cosine window, but also can provides prior information about the target object. Moreover, to model a robust object prior, we propose a generic strategy for adaptive fusion and update of multiple features. Extensive experiments over multiple tracking benchmarks demonstrate the superior accuracy and real-time performance of our methods compared to the state-of-the-art trackers.

***Index Terms*—** Visual Tracking, Correlation Filter, Spatially Attentive Model, Adaptive Feature Fusion

1. INTRODUCTION

The Correlation Filter (CF) based trackers have attracted wide attention [4, 5, 6, 3, 7, 8, 9], due to their superior accuracy and real-time performance benefiting from cyclic shifts model and ridge regression objective equation. The cyclic shifts model increases the number of samples, which enhances the discriminative ability of correlation filters. And the form of loss function makes it possible to get a closed-form solution in frequency domain with high speed.

Despite all the advantages, the CF trackers still have some limitations. Bolme et al. [4] indicates that the cyclic shifts connect the sample's boundaries and create artifacts that do not exist at the boundary which is called *boundary effect*. This effect undermines discriminative ability of the correlation filters. To mitigate the boundary effect, the image is multiplied by a cosine window which sets the values of boundary pixels to zero. However, the cosine window introduces some new problems due to its restricted search area, especially with the challenges of *fast motion* and *occlusion*. Both of them lead to the abrupt motion of the target in two contiguous frames. It is obvious that the target will be lost, if the range of motion is larger than the search area of the cosine window.

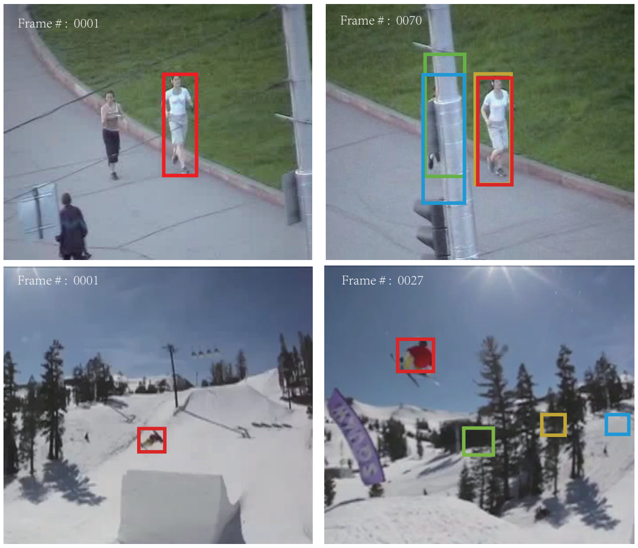

Fig. 1. Tracking results of our spatially attentive CF of the baseline Staple tracker, denoted as Staple$_{SA}$, and a comparison with recent state-of-the-art tracking algorithms Staple [1], BACF [2], SRDCF [3] on the Jogging-2 and Skiing sequences from OTB-100.

To eliminate these limitations of fast motion and occlusion, predecessors have done a lot of work. SRDCF [3] expands the search area while penalizing the filter coefficients of the boundary area to suppress background information interference. However, it undermines the closed-form solution of the correlation filter so that the objective equation can only be solved by iterative method which drastically retards the speed of the CF trackers. And [10, 2] propose a new cyclic shift method on a larger region that contains the target, and then crops out the block where the target was located as a negative sample. In addition, [11] propose an adaptive target response model during the training stage to solve the fast motion and

This work is supported by Natural Science Foundation of Hubei Province (2018CFB426), and the Fundamental Research Funds for the Central Universities No. 2017KFYXJJ179.

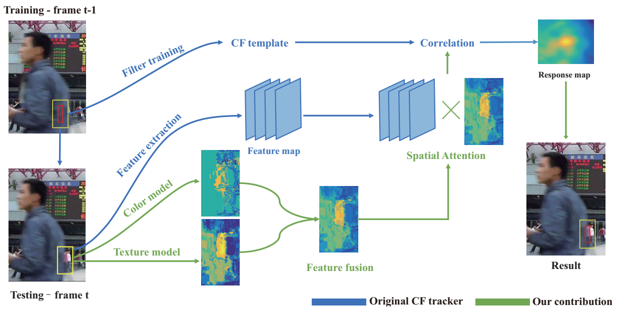

Fig. 2. Pipeline of our spatially attentive correlation filter.

occlusion challenge. All of the aforementioned algorithms focus on eliminating the limitations in the training stage. In contrast, we find that improvement during the detection stage is a more straightforward and effective way. The reasons are as follows. (a) Since fast motion and occlusion occurs in the detection stage rather than in the training stage, it is more essential to solve this problem in the detection stage. (b) The improvement in the detection stage does not affect the training of correlation filter, and it dose not disturb the closed-form solution.

Furthermore, to obtain more robust features, recent CF-based methods [8, 9, 1, 7] have taken the strategy of multi-features fusion. Most existing algorithms use a fixed weight strategy for feature fusion, such as Staple [1] which merge the color map and response map of CF with weight of 0.3 and 0.7, respectively. However, this strategy has some drawbacks. (a) Different sequences have a preference for different features. For example, the color feature makes sense when target and background have a significant difference in color, while a sequence with no difference in target and background colors may have a preference for other features. It is impossible to find a fixed weight optimal for all sequences. (b) The target and background also change during the tracking process, however, the fixed weight is not robust to these variations. With the change of target and scene, the fixed weight strategy based features cannot accurately represent target, which will result in tracking failure.

To this end, this paper propose a Spatially Attentive Correlated Filters (SACF) framework for real-time object tracking. As shown in Figure 1, our method has excellent performance in occlusion and fast motion situation without breaking the closed-form solution of the correlation filter, so as to achieve superior accuracy with real-time performance. The main contributions of our work can be summarized as follows:

- We propose a spatially attentive model to provide spatial guidance during the detection stage, which is simple yet effective, and without breaking the closed-form solution(see Figure 2).
- To model a robust object prior, we propose a generic strategy for multi-features adaptive fusion.
- Extensive experiments demonstrate that our Spatially Attentive framework is widely applicable to all CF based trackers and can significantly and steadily improve their performance at low computational cost.

2. PROPOSED APPROACH

2.1. Spatially Attentive Correlated Filters

Before the detailed discussion of our proposed approach, we first revisit the formulas of the conventional CF trackers. In the training stage, the goal is to obtain a CF template w that minimizes the squared error over samples x_i and their regression targets y_i,

$$\min_{w} \quad \sum_{i}(w^T \hat{x}_i - y_i)^2 + \lambda\|w\|^2 \tag{1}$$

where λ is a regularization parameter to avoid overfitting, and $\hat{x_i}$ can be obtained from the output of a cosine window C by $\hat{x_i} = x_i \odot C$, where the symbol $\odot$ denotes an element-wise product. However, the cosine window restricts the region of filter during training and detection stage which results in an unrecoverable tracking drift on certain categories. Naturally, the CF tracking performance can be significantly improved if we can provide the filter with prior information about the location of the target. Therefore, we propose a framework for CF trackers that adds spatial attention to the correlation filters during the detection stage.

In the training stage, we create feature histograms for the target and background respectively in the first frame and update them in the subsequent frames. In the testing stage, we calculate the probability for each pixel and get the foreground and background probability matrix P_f , P_b. The pixel belongs to target are expected with a high probability in P_f and a low probability in P_b. Therefore, the probabilities that the pixels in detection area belong to the target can be expressed as,

$$P = \frac{P_f}{P_f + P_b} \tag{2}$$

Then we can get a spatial attention based feature map: $x_i^{'} = x_i \odot P$. Noting that the probabilities of the boundary pixels in matrix P are close to zero, that is to say, the probability matrix P can also suppress the boundary effect as cosine window. We can rewrite the objective equation by substituting $x_i^{'}$ for $\hat{x_i}$ in (1):

$$\min_{w} \quad \sum_{i}(w^T x_i^{'} - y_i)^2 + \lambda\|w\|^2 \tag{3}$$

It's desirable that our proposal does not break the closed-form solution of the equation, thus the speed of CF tracking is almost unaffected.

2.2. Adaptive Features Fusion

To obtain an object prior which is robust to the variations of target and background, we propose an adaptive features fusion strategy. In each frame t, we define the formula as:

$$P = \sum_{d=1}^{D} P^d \omega_t^d \tag{4}$$

where P^d and ω_t^d denote the feature map and the adaptive weight of feature d, respectively.

For the ideal feature map, the pixels belonging to target and background have relatively high and low response scores, respectively. However, there are some background pixels confused with those from target, which disturb the tracking process. We call them as *critical pixels*. We propose a method to calculate ω_t^d by measuring the difference between the response scores of target pixels and critical pixels. For a 2-D feature map $P_{m\times n}$, we define the discriminating ability α :

$$\alpha = \frac{1}{k\theta_1}\sum_{i=1}^{k\theta_1} p_i - \frac{1}{k\theta_2}\sum_{j=k\theta_1+1}^{k(\theta_1+\theta_2)} p_j \tag{5}$$

where θ_1, θ_2 denote the ratio of target and critical pixels to all pixels, respectively. $k = m \times n$ denotes the number of pixels in matrix $P_{m\times n}$. Arranging the elements of matrix $P_{m\times n}$ in descending order $(p_1, p_2 ...... p_{m\times n})$, p_i represents the i th element. Then we can get the normalized weight ω^d :

$$\omega^d = \frac{\alpha^d}{\sum_{l=1}^{D} \alpha^l} \tag{6}$$

Finally, we update the weight ω_t^d in (4) with a learning rate η and ω^d calculating from Eq. (6),

$$\omega_t^d = (1-\eta)\omega_{t-1}^d + \eta\omega^d \tag{7}$$

3. EXPERIMENT

In this section, we integrate our framework with three popular CF trackers (Staple [1], KCF [5], DCF [5]) and have conducted extensive experiments on two public datsets, OTB-100 [19] and Temple Color 128 [20]. In the first part we present the details related to our experiment. In the second part, we quantitatively evaluate our framework and have a comparison to state-of-the-art trackers to validate the effectiveness of our framework.

3.1. Experiment Details

Evaluation methodology. All trackers are evaluated according to two measures, precision and success, as defined in OTB-50/OTB-100 [21]. Precision measures the center error between the tracked bounding box and ground truth. The common threshold of 20 pixels is used for ranking trackers. Success is measured as the intersection over union (IoU) of the tracked bounding box and the ground truth. The trackers are ranked by the area under the curve (AUC).

Table 1. Success rates (overlap threshold with AUC) of our framework-based trackers compared to state-of-the-art CF trackers without deep features. The first and second highest rates are highlighted in color. The $*$ denotes speeds from the original paper, not test on the same platform.

	Published	OTB100	TC128	FPS
Staple$_{\text{SA}}$	ours	0.626	0.537	35.83
Staple[1]	2016 CVPR	0.579	0.5	75.31
KCF$_{\text{SA}}$	ours	0.483	0.459	92.79
KCF[5]	2015 PAMI	0.469	0.389	175.59
DCF$_{\text{SA}}$	ours	0.452	0.46	127.79
DCF[5]	2015 PAMI	0.432	0.388	350.37
Staple$_{\text{CA}}$[12]	2017 CVPR	0.598	-	35.2*
LMCF[13]	2017 CVPR	0.568	-	85*
BACF[2]	2017 ICCV	0.63	0.52	35.3*
MUSTer[14]	2015 CVPR	0.575	-	4*
LCT[15]	2015 CVPR	0.562	0.439	27.4*
SRDCF[3]	2015 ICCV	0.598	0.517	5*

Table 2. Success rates (overlap threshold with AUC) of our best performing CF tracker(Staple$_{\text{SA}}$) compared to state-of-the-art CF trackers based on deep learning. The $*$ denotes speeds from the original paper, not test on the same platform.

	Published	OTB100	TC128	FPS
Staple$_{\text{SA}}$	ours	0.626	0.537	35.83
CREST[16]	2017 ICCV	0.623	-	1*
CFNet[17]	2017 CVPR	0.568	-	75*
ECO[8]	2017 CVPR	0.694	0.612	6*
C-COT[9]	2016 ECCV	0.686	0.583	0.3*
DeepSRDCF[18]	2015 ICCV	0.643	0.543	$<$1

Update strategy. We employ the color feature and the texture feature [22] in our spatially attentive framework with the initial weights 0.75 and 0.25 respectively. The update rate of feature fusion η is set to 0.1. The CF template is updated in the same way as the original trackers.

3.2. Quantitative Results

Ablation research. We conducted experiment on the baseline method of Staple, as shown in Figure 4. The total improvement of precision and success rate has reached an astonishing $\{5.3\%,\ 4.7\%\}$ in Staple. The results demonstrate the good performance of the proposed framework.

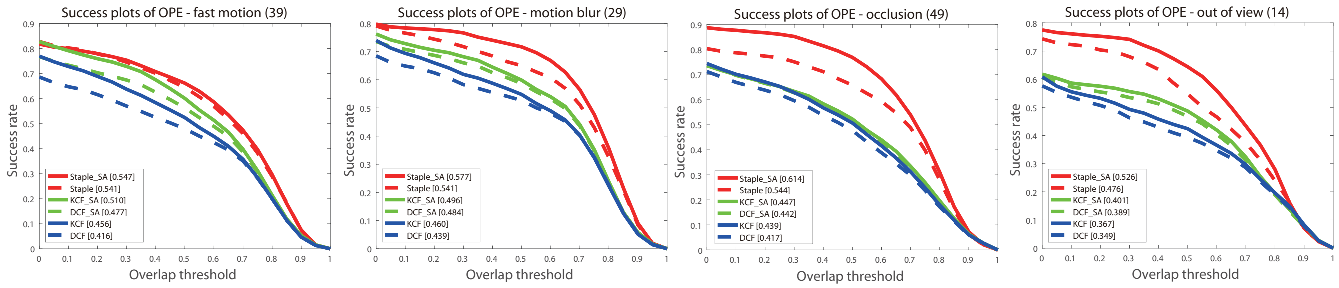

Fig. 3. Average performance on OTB-100 for 4 attributes.

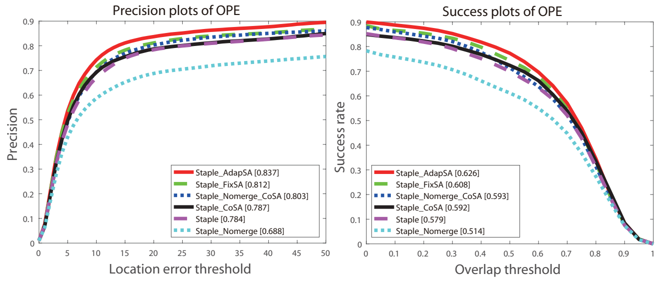

Fig. 4. Performance comparison on OTB-100 in different configurations. The suffix CoSA denotes the spatially attentive framework with color feature, FixSA denotes that color and texture features are fused with fixed weights, Nomerge denotes the Staple without response merge, AdapSA means spatially attentive framework with adaptive feature fusion.

We then evaluate the components of the proposed method. Noting that staple already has a color model in training stage, thus we take $\text{Staple}_{\text{Nomerge}}$ as a baseline, which can be approximately considered as a version of staple without the color model. Particularly, (1) $\text{Staple}_{\text{CoSA}}$ outperforms $\text{Staple}_{\text{Nomerge}}$, which demonstrates the effectiveness of the color based spatially attentive framework. (2) $\text{Staple}_{\text{CoSA}}$ outperforms Staple, which demonstrates that our spatially attentive framework is more effective than the response merge strategy in Staple [1]. (3) $\text{Staple}_{\text{FixSA}}$ outperforms $\text{Staple}_{\text{CoSA}}$, which demonstrates the complementary effect of the texture feature. (4) $\text{Staple}_{\text{AdapSA}}$ outperforms $\text{Staple}_{\text{FixSA}}$, which demonstrates the effectiveness of our adaptive feature fusion method.

Evaluation per attribute. While our framework improves tracking performance in most scenarios, there are certain categories that benefit more than others. Here we just present the result of four attributions which are related to the fast motion and occlusion. On the left of Figure 3, the spatially attentive based trackers has a significant improvement over the original one for the attribution of fast motion and motion blur, which proves that our framework is robust to fast motion. And in the other two graphs of Figure 3, our framework also achieves significant improvement in the cases of occlusion and out of view. This is largely due to the fact that compared to cosine window, our spatially attentive framework can provide a more accurate detection position before the detection stage.

Comparison to baselines. We evaluated the performance of our spatially attentive CF trackers and their baselines on the OTB-100 and Temple-128 datasets. As shown in Table 1, our framework has an improvement of $\{3.7\%, 7.0\%, 7.2\%\}$ on Temple-128 for the baseline Staple, KCF and DCF, respectively. As for the OTB-100, there are 23 grayscale videos which undermine performance of the color-based framework. Furthermore, the KCF and DCF do not have scale estimation, which also affect the reliability of feature extraction. Despite these unfavorable factors, there are still $\{4.7\%, 1.4\%, 2.0\%\}$ promotion on OTB-100. This further proves the stability and applicability of our framework for all CF trackers.

Comparison to state-of-the-art trackers. We compare our spatially attentive CF trackers and their baselines to state-of-the-art trackers without deep features in Table 1 and compare our best performing CF tracker ($\text{Staple}_{\text{SA}}$) to state-of-the-art CF trackers with deep features in Table 2. In Table 1, the tracker $\text{Staple}_{\text{SA}}$ based on our framework ranked first on Temple-128 and second on OTB-100. Furthermore, it has reached the real-time requirements (FPS$\geq$30). When compared to the state-of-the-art methods based on deep learning, $\text{Staple}_{\text{SA}}$ still gets the competitive performance while far exceeding the speed of other algorithms.

4. CONCLUSION

We propose a spatially attentive framework which is applicable for all correlation filter based trackers. The spatial guidance can locate the target precisely before the detection stage, capacitating CF based trackers more robust to fast motion and occlusion. And the adaptive feature fusion strategy enables better representation of features among different videos. Extensive experiments demonstrate that our framework can significantly and steadily improve the performance of CF based trackers at low computation cost. In addition, our framework may promote all detection-based trackers, which is subject to our further experiments to prove.

5. REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr, "Staple: Complementary learners for real-time tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1401–1409.

[2] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey, "Learning background-aware correlation filters for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1135–1143.

[3] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4310–4318.

[4] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui, "Visual object tracking using adaptive correlation filters," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2544–2550.

[5] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[6] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[7] Yang Li and Jianke Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *European Conference on Computer Vision*. Springer, 2014, pp. 254–265.

[8] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg, "Eco: Efficient convolution operators for tracking," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, 2017, pp. 21–26.

[9] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 472–488.

[10] Hamed Kiani Galoogahi, Terence Sim, and Simon Lucey, "Correlation filters with limited boundaries," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 4630–4638.

[11] Adel Bibi, Matthias Mueller, and Bernard Ghanem, "Target response adaptation for correlation filter tracking," in *European conference on computer vision*. Springer, 2016, pp. 419–433.

[12] Matthias Mueller, Neil Smith, and Bernard Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2017, pp. 1396–1404.

[13] Mengmeng Wang, Yong Liu, and Zeyi Huang, "Large margin object tracking with circulant feature maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, 2017, pp. 21–26.

[14] Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, and Dacheng Tao, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 749–758.

[15] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming-Hsuan Yang, "Long-term correlation tracking," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 5388–5396.

[16] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson WH Lau, and Ming-Hsuan Yang, "Crest: Convolutional residual learning for visual tracking," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2574–2583.

[17] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr, "End-to-end representation learning for correlation filter based tracking," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5000–5008.

[18] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 58–66.

[19] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[20] Pengpeng Liang, Erik Blasch, and Haibin Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5630–5644, 2015.

[21] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in *Computer vision and pattern recognition (CVPR), 2013 IEEE Conference on*. Ieee, 2013, pp. 2411–2418.

[22] Shengcai Liao, Guoying Zhao, Vili Kellokumpu, Matti Pietikäinen, and Stan Z Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1301–1306.