Visual Tracking via Spatial-Temporal Regularized Correlation Filters with Advanced State Estimation

Zhao-Qian Tang* and Kaoru Arakawa†
*Meiji University, Tokyo, Japan
E-mail: cs172017@meiji.ac.jp Tel: +81-3-5343-8340
†Meiji University, Tokyo, Japan
E-mail: kara@meiji.ac.jp Tel: +81-3-5343-8340

***Abstract*— Discriminative correlation filter (CF) based visual trackers achieves outstanding performance with the hand-crafted feature in visual tracking. In this work, based on the discriminative correlation filter, we propose a new Spatial-Temporal regularized correlation filters with advanced state estimation (CFASE) to achieve more significant tracking performance. First, we propose a new method to estimate correlation filters more precisely using prediction from the previous two filters, considering the drift during the tracking process. Second, we train two correlation filters models to obtain scale estimation and object location, respectively. The separated two correlation filter models help to reduce the adverse effects of scale changes on object location. Third, our tracker introduces average peak-to-correlation energy (*APCE*) to evaluate the accuracy of scale estimation and object location. Experimentally, the proposed tracker (CFASE) achieves outstanding and real-time performance for the challenging benchmark sequence (OTB2013, OTB2015, and TC128).**

I. INTRODUCTION

Visual tracking is one of the fundamental topics in computer vision. When the information on an arbitrary object from the first frame is given, the temporally changing object in the remaining image sequences is automatically detected. Visual tracking has been applied to UAV, self-driving, human-computer interaction, and video surveillance by the right of outstanding performance. It remains enormous challenges such as deformations, fast motions, occlusions, background clutter, and scale variations, and so on.

Correlation filter (CF) is one of the most successful frameworks in visual tracking. Trackers based on correlation filtering have significant speed and achieve outperformance in tracking. Henriques et al. proposed a kernelized correlation filter tracker (KCF) [2] becomes one of the best tracking baselines with state-of-the-art performance. KCF utilizes a cyclic sliding window operation to obtain a large number of samples, which improves the robustness of tracking filters. This method cleverly uses fast Fourier transform (FFT) to convert correlation operations in the spatial domain to dot product in the frequency domain, which greatly reduces the calculation burden. In order to improve the accuracy of

Fig. 1 Qualitative evaluation of CFASE and STRCF on the *Girl2*(upper) *and Soccer* (lower) videos sequence with occlusion and background clutter, respectively. CFASE obtains outperformance than STRCF on occlusion and background.

trackers based on correlation filtering, M. Danelljan proposed discriminative scale-space tracker (DSST) [3] and Yang Li proposed a scale pool technique [8] to solve scale variation, respectively. In order to take full advantage of the information of objects, conventional handcraft features (HOG and CN) are introduced to keep the robustness of model [21, 22].

There are still serious limitations for trackers based on correlation filters that obtain outperformance in visual tracking. As discussed above, trackers based on correlation filtering use a cyclic sliding window operation to obtain a set of training samples, so the training and detection samples have periodicity in the frequency domain; this periodic assumption produces boundary effects. Since the object search area of trackers based on correlation filtering is constrained, expansion of the search area makes no sense at all, resulting in a limited performance in tracking. Moreover, the boundary effects also reduce the distinguishability of the model. M. Danelljan proposed Spatially Regularized Discriminative Correlation Filters (SRDCF) [26] to solve the boundary effects. SRDCF introduces penalize correlation filter coefficients in learning and uses the iterative Gauss-Seidel method to the optimal formulation. However, the optimization strategy increases the burden of calculation, undermining the real-time performance of the trackers based on correlation filtering. Based on SRDCF, Feng Li proposed

Spatial-Temporal regularized correlation filters (STRCF) [4]. This method incorporates both spatial and temporal regularization into the correlation filtering framework to improve the robustness of the appearance model. Moreover, this method applies the alternating direction method of multipliers (ADMM) to the optimal formulation, achieving real-time performance in tracking.

In this paper, we proposed a real-time outperformance tracker based on STRCF. The contributions of our method are as follows.

First, we achieve a more precise estimation of the correlation filter by considering the temporal continuous change of correlation filters. While STRCF utilizes only one latest filter for the temporal regularization, our method predicts the current filter from two latest filers and applies the temporal regularization to the predicted one. Accordingly, our method can achieve a more precise estimation of the current filter. Moreover, the calculation additionally required for this improvement is sufficiently small. Thus, our new method is fast enough for real-time object tracking.

Second, to reduce the adverse effects of scale variation on object location, we distinguish the process of scale estimation from object location and train the two correlation filtering models, respectively. Considering the calculation efficiency, the correlation filtering model of scale estimation only uses the HOG feature. At the same time, we introduce average peak-to-correlation energy (APCE) [28] to estimate the accuracy of the object scale.

Finally, we develop the experiments on three challenging datasets: OTB-2013 [7] with 51 videos, OTB-2015 [15] with 100 videos, and TC128 [1] with 128 videos. our method achieves obvious improvements compared to STRCF in some tracking attributes. As shown in Fig. 1, we can see our method achieves better performance for attributes as occlusion and background clutter. Furthermore, our method obtains outstanding performance compared to all the state-of-the-art trackers.

II. Related Work

Recently, discriminative correlation filters (DCF) make rapid progress in visual tracking. Some trackers based on discriminative correlation filters achieve great tracking results on different benchmarks. For example, KCF [2] obtains success on both speed and accuracy in tracking. Based on KCF, some state-of-the-art trackers appear. Although the advantages of CF trackers are circular samples and fast calculations, the disadvantage is also caused by circular samples. This brings boundary effects. CF trackers only use a limited search region, too large a search region will increase the negative impact of background information.

Hamed Kiani Galoogahi proposed learning background-aware correlation filters (BACF) [6] to solve the unwanted boundary effects. BACF extracts an image search region to obtain more samples. To reduce the impact of negative samples, BACF only focuses on the circulant samples in the center area, the size of which is equivalent to the search field of the original CF trackers. In this way, BACF not only expands the search domain but also uses real samples while ensuring the circulant structure of samples. Unfortunately, due to the enforcement of a spatial constraint, BACF cannot efficiently solve correlation filter as KCF and must be solved in the spatial domain with the alternating direction method of multipliers (ADMM). However, BACF can still achieve real-time and sharply increase the performance of CF trackers in object tracking. Despite the unwanted boundary effects, the discriminative correlation filters trackers compare favorably with deep learning trackers in the performance of tracking.

SRDCF was also proposed to solve the boundary effects like BACF. The difference between BACF and SRDCF is that SRDCF introduces a spatial regularization weight function to penalize the magnitude of the correlation filter coefficients w in learning. The value of the weight depends on the spatial locations. The closer to the center, the lower the coefficient, and the closer to the surrounding, the higher the coefficient. w reduces the negative influence of boundary effects. Therefore, discriminative correlation filters can be learned in larger image regions. The formulation of SRDCF as following.

$$\min_{\boldsymbol{f}} \sum_{k=1}^{T} \left\| \sum_{d=1}^{D} x_k^d * f^d - y_k \right\|^2 + \sum_{d=1}^{D} \left\| w \cdot f^d \right\|^2 \qquad (1)$$

Here, $*$ denotes circular convolution, $\cdot$ denotes the Hadamard product. x_k denotes the samples with size of $N \times M$, where each sample consists of D feature maps. y_k is the desired Gaussian-shaped label. w is the spatial regularization weight, f is the correlation filter. Although the boundary effects are solved, SRDCF uses the iterative Gauss-Seidel method to minimize Equation (1). This method reduces the efficiency of calculation, and the real-time nature of the trackers is sacrificed.

III. Our Approach

A. Review STRCF

STRCF adopts the same methods as SRDCF to solve the boundary effects. Compared to SRDCF, STRCF considers both spatial and temporal regularization into the correlation filtering framework. STRCF is expressed as follow,

$$\min_{\boldsymbol{f}} \frac{1}{2} \| \sum_{d=1}^{D} x_t^d * f^d - y \|^2 + \frac{1}{2} \sum_{d=1}^{D} \| w \cdot f^d \|^2 + \frac{\mu}{2} \| f - f_{t-1} \|^2 \qquad (2)$$

Here μ is a regularization parameter, w is the spatial regularization weight, f_{t-1} denotes the correlation filter obtained in the previous frame.

STRCF can get a perfect correlation filter model f by minimizing Equation (2). Temporal regularization was used to prevent the corruption of the correlation filter model in each frame because the obtained filters are close to the previous one. STRCF develops the alternating direction method of multipliers (ADMM) to efficiently optimized formulation in which each sub-problem has the closed-form solution. Therefore, STRCF achieves real-time outperformance in tracking.

B. Consideration of Temporal State Change of Objects

STRCF solves the boundary effects with spatial-temporal regularization. As discussed in Equation (2), STRCF introduces temporal regularization into SRDCF and obtains the optimum correlation filter by minimizing this formula. Temporal regularization makes sure that the obtained f is as similar as possible to the filter in the (*t-1*)-th frame, to prevent the corruption of the correlation filter, and it can also play a good role against occlusion. However, since STRCF only considers the correlation filter in the previous one frame, it cannot consider the continuous state change of the object. A new method is proposed here to predict the current filter from the previous two filters f_{t-1} and f_{t-2}, and apply the regularization to the predicted filter so that the current filter can be close to the predicted filter. The predicted filter f_* is obtained as the following Equation (3) using a positive small value α.

$$f_* = f_{t-1} + \alpha(f_{t-1} - f_{t-2}) \tag{3}$$

Comparing to the temporal regularization of STRCF, the previous filter f_{t-1} is replaced with the predicted filter f_* in order to reflect the trends in change of the correlation filters from f_{t-2} to f_{t-1}. This method can consider the state change of the object more effectively by replacing f_{t-1} with f_*. The new STRCF is expressed as follows.

$$\min_{\boldsymbol{f}} \tfrac{1}{2}\|\textstyle\sum_{d=1}^{D} x_t^d * f^d - y\|^2 + \tfrac{1}{2}\sum_{d=1}^{D}\|w \cdot f^d\|^2 + \tfrac{\mu}{2}\|f - f_*\|^2 \tag{4}$$

The optimization process of STRCF is not changed. The robustness of temporal regularization is improved without increasing the burden of calculation.

C. New Scale Estimation

Scale change is also one of the factors of the temporal state change of objects, and accurate scale estimation is required to avoid the influence of the scale change and get high tracking performance. STRCF adopts the same scale estimation (Scale Pool) as SAMF [8] in which the object is detected by the translation correlation filter on the multi-scale image regions, and the translation position and best scale with the largest response are obtained. Thus, the scale pool technology can detect the change of object location and scale variation.

However, the largest response does not always correspond to the correct object location because of the state change around the object. To estimate the scale more precisely, a new method for scale estimation is proposed here.

In the new method, the scale estimation filter is trained by HOG feature, and the object location filter is trained with hand-craft (HOG+CN) features. During the process of scale estimation, the best scale of the object is obtained as S_{HOG}, and the object position as Pos_{HOG}. This best scale is used to object location filter to obtain the object position as Pos_{HOG+CN}. Through different features, we can get two object positions. Under normal conditions, these two positions should be similar. While obtaining the best object position, we can use the distance $Dist$ between the two positions to judge the reliability of the scale.

$$Dist = \sqrt{(Pos_{HOG} - Pos_{HOG+CN})^2} \tag{5}$$

To accurately estimate the scale, we also consider the change of the response map. In the method of discriminating the abnormality in object tracking, the average peak-to-correlation energy (*APCE*) shows the fluctuation of response map and reflects the reliability of tracking to a certain extent; the larger the value of *APCE* is, the more reliable the tracking is. *APCE* is defined as

$$APCE = \frac{\|R_{max} - R_{min}\|^2}{mean(\sum_{i,j}(R_{i,j} - R_{min})^2)} \tag{6}$$

where $\boldsymbol{R_{max}}$, $\boldsymbol{R_{min}}$, and $\boldsymbol{R_{i,j}}$ denote the maximum, minimum, and the i-row j-column elements of response map.

When $Dist$ is great and *APCE* is small, S_{HOG} is considered to be not reliable. Thus, the scale estimation method proposed here adopts the scale in the previous frame, if $Dist$ is greater than a threshold ($Dist$ > 20) and *APCE* is less than a threshold (*APCE* < 0.3).

IV. EXPERIMENTS

All the experiments are conducted on the Matlab R2020a platform, and a PC machine with an Intel (R) Core (TM) i7-9700F CPU (3.00GHZ), 16GB memory. In the proposed method, the size of the search region is set to 5 times the size of the object. HOG features are used for scale estimation and HOG and CN are used for object location. The regularization parameter μ is set to 15 and 13 for location filter and scale estimation filter, respectively. In Equation (3), α is set to 0.5 and 0.2 for the location filter and scale estimation filter, respectively.

We offer comprehensive assessments to evaluate the performance of the proposed CFASE on OTB-2013, OTB-2015 and Temple color 128 benchmark database. We also evaluate the effect of introducing the filter prediction in

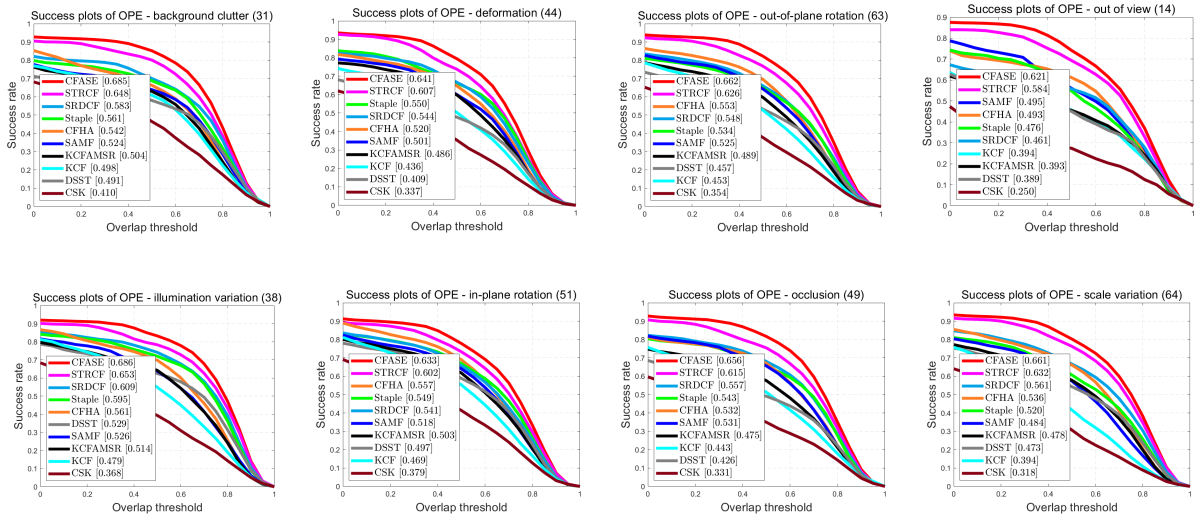

Fig.2 Evaluation of different trackers with eight attributes on OTB-2015.

temporal regularization as described in 3.2 and the advanced scale estimation as in 3.3 respectively on OTB-2013. To make orientation analysis about CFASE, we select eight attributes from OTB-2015 benchmark, to analyze the reliability of our method.

A. OTB-2013 and OTB-2015

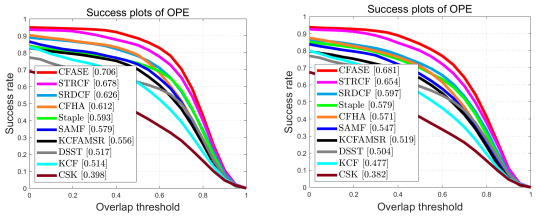

Fig.3 Success plot on OTB-2013 and OTB-2015, respectively.

Table 1. Success and speed of top-5 trackers on the OTB-2015. The best two results are shown in red and blue fonts, respectively.

	SRDCF	Staple	CFHA	STRCF	CFASE
Success	0.597	0.579	0.571	**0.654**	**0.681**
FPS	10.4	**105.5**	**115.1**	33.4	31.1

OTB-2013 and OTB-2015 contains 51, 100 video sequences, respectively. OTB benchmark database is annotated with 11 attributes to evaluate the performance of trackers, such as deformation (DEF), fast motion (FM), background clutters (BC), illumination variation (IV), motion blur (MB), scale variation (SV), in-plane rotation (IPR), low resolution (LR), occlusion (OCC), out of plane rotation (OPR), out of view (OV). Based on OTB benchmark, we compare our proposed tracker with nine state-of-the-art trackers (CSK [5], SRDCF [26], CFHA [29], Staple [25], SAMF [8], STRCF [9], DSST [3], KCFAMSR [17] and KCF [2]) by AUC (Area under the curve of success rate plots). As shown in Fig. 3, our method obtains a success score of 70.6% and 68.1% based on OTB-2013, OTB-2015, respectively. Compare to the baseline STRCF, our method achieves the improvement of 2.8% and 2.7% on OTB, respectively. As shown in Table 1, we show the success and speed of top-5 trackers on the OTB-2015. CFASE achieves the best performance in calculation speed 31.1 fps compared with the rest of state-of-the-art trackers.

We show the evaluation of different trackers with eight attributes on OTB-2015 in Fig.2. Our method CFASE obtains better performance than other trackers in these attributes. Especially in challenge attributes such as background clutter, deformation, out of view and occlusion, our method obtains a gain of 3.7%, 3.4%, 3.7%, and 4.1% better than the second-best tracker STRCF. This is mainly because we consider the predicted filter in temporal regularization and new scale estimation in STRCF. Since our method estimates the scale change more precisely, we can see that CFASE obtains an improvement of 2.9% over STRCF in attribute as scale variation. The results of the experiment demonstrate that the proposed method is more stable and efficient than STRCF.

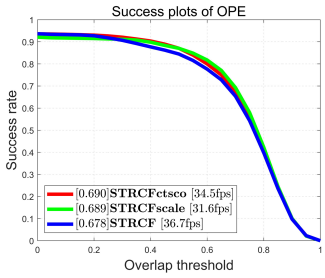

Fig.4 Success plots of STRCF, STRCF with Consideration of Temporal State Change of Objects (STRCFctsco), and Scale estimation filter (STRCFscale) on OTB-2013.

Fig. 4 shows the effect of introducing the filter prediction and the advanced scale estimation respectively on OTB-2013. Both of them get a gain of 1.2% over STRCF. Especially, the burden of calculation does not increase in introducing filter prediction.

B. Temple color 128

Temple color 128 (TC128) contains a large set of 128 color sequences. Most modern trackers use color information, while OTB benchmarks have some grayscale images. It is not enough for some state-of-the-art trackers with color features to evaluate their performance on OTB, so we also conducted experiment of evaluation on Temple color 128 benchmark with CFASE, STRCF [9], MEEM [22], Struck [15], ASLA [11], VTD [19], CN2 [13], DFT [10], CSK [5], KCF [2].

In addition to the success scores, we use the precision scores. As shown in Fig. 5, our method obtains the outperformance both in precision plot and success plot on TC128.

Compared to STRCF, CFASE obtains a gain of 2.1% and 2.9% in precision scores and success scores, respectively. Here the precision score indicates the ratio of the frames in which the distance between estimated locations and the ground-truth positions are within 20 pixels.

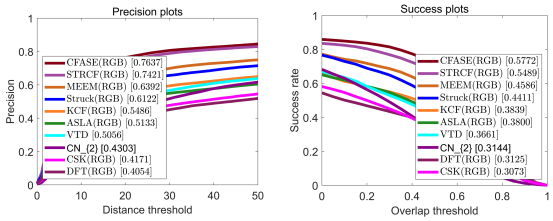

Fig.5 Precision plot and Success plot on TC128.

V. CONCLUSION

In this work, we propose a novel Spatial-Temporal regularized correlation filters with advanced state estimation (CFASE) based on CF to obtain excellent tracking performance. In this method, filter prediction is newly-introduced into temporal regularization to consider continuous temporal change of objects, and advanced scale estimation is introduced to estimate the scale more precisely. In the advanced scale estimation, two correlation filter models are trained respectively to obtain scale estimation and object location. We also introduce APCE to estimate the accuracy of scale and position from scale estimation filter and location filter, respectively. By experiments, we demonstrate CFASE gets more robust performance than STRCF and achieves outstanding and real-time performance for the challenging benchmark sequences. How to adjust the parameters utilized in CFASE appropriately for the video type is for further research.

REFERENCES

[1] PP. Liang, E. Blasch, and H. Ling. "En- coding color information for visual tracking: Algorithms and benchmark." *IEEE Transcations on Image Processing*, vol. 24, no. 12, pp.5630–5644, 2015.

[2] F. Henriques, R. Caseiro, P. Martins, and J. Batista. "High-speed tracking with kernelized correlation filters." *TPAMI*, vol. 37, no. 20, pp. 583-596, 2014.

[3] M. Danelljan, G. Häger, F. Khan, M. Felsberg. "Accurate Scale Estimation for Robust Visual Tracking." In *BMVC*, 2014.

[4] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang. "Learn- ing spatial-temporal regularized correlation filters for visual tracking." In *IEEE Conference on Computer Vision and Pat-tern Recognition*, 2018.

[5] F. Henriques, R. Caseiro, P. Martins, and J. Batista. "Exploiting the Circulant Structure of Tracking-by-detection with Kernels." In *ECCV LNCS*, vol.7575, pp.702-715,2012.

[6] H. K. Galoogahi, A. Fagg, and S. Lucey. "Learning background-aware correlation filters for visual tracking." In *ICCV*, 2017.

[7] Y. Wu, J. Lim, and M.H. Yang. "Online object tracking: A benchmark." In *CVPR*, pp.2411-2418,2013.

[8] F. Li, C. Tian, W. M. Zuo, L. Zhang, and M. H. Yang. "Learning spatial-temporal regularized correla- tion filters for visual tracking." In *CVPR*, 2018.

[9] H. K. Galoogahi, A. Fagg, and S. Lucey. "Learning background-aware correlation filters for visual tracking." *In ICCV*, 2017.

[10] H L. Sevilla-Lara and E. Learned-Miller. "Distribution Fields for tracking." In *CVPR*, 2012.

[11] X. Jia, H. Lu, and M.-H. Yang. "Visual Tracking via Adaptive Structural Local Sparse Appearance Model." In *CVPR*, 2012.

[12] M. Danelljan, F. Shahbaz Khan, M. Felsberg and J. van de Weijer. "Adaptive Color Attributes for Real-Time Visual Tracking." *In CVPR*, 2014.

[13] W. Yi., LJ. Yang. "Object tracking benchmark." *TPAMI*, vol. 37, no. 9, pp. 1834– 1848 (2015)

[14] S. Hare, A. Saffari, and P. H. S. Torr. "Struck: Structured Output Tracking with Kernels." In *ICCV*, 2011.

[15] ZQ. Tang, K. Arakawa. "Kernel Correlation Filter via Adaptive Model." *In ISPACS*, 2018.

[16] J. Kwon and K. M. Lee. "Visual Tracking Decomposition." In *CVPR*, 2010.

[17] Z. Kalal, K. Mikolajczyk, and J. Matas. "Tracking-learning-detection." *IEEE transactions on pattern analysis and ma- chine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[18] J. Kwon and K. M. Lee. "Visual tracking decomposition." In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1269–1276. IEEE, 2010.

[19] J. Zhang, S. Ma, and S. Sclaroff. "Meem: robust tracking via multiple experts using entropy minimization." In *ECCV*, 2014.

[20] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. Torr. "Staple: Complementary learners for real-time tracking." In *CVPR*, 2016.

[21] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. "Learning spatially regularized correlation filters for visual tracking." In *ICCV*, 2015.

[22] MM. Wang, Y. Liu, and ZY. Huang. "Large Margin Object Tracking with Circulant Feature Maps." In *CVPR*, 2017.

[23] ZQ. Tang, K. Arakawa. "Visual Tracking via Correlation Filter using Luminance Histogram and Adaptive Model." In *SISA*, 2019.