CONTENT-RELATED SPATIAL REGULARIZATION FOR VISUAL OBJECT TRACKING

Ruize Han, Qing Guo*, Wei Feng*†

School of Computer Science and Technology, Tianjin University, Tianjin, China
Key Research Center for Surface Monitoring and Analysis of Cultural Relics, SACH, China
{han_ruize, tsingqguo, wfeng}@tju.edu.cn

ABSTRACT

Spatial regularization (SR), being an effective tool to alleviate the boundary effects, can significantly improve the accuracy and robustness of correlation filters (CF) based visual object tracking. The core of SR is a spatially variant weight map that is used to regularize the online learned correlation filters by selecting more meaningful samples. However, most existing trackers apply a data-independent SR weight map. In this paper, we show that a content-related spatial regularization (CRSR) can help to further boost both the tracking accuracy and robustness. Specifically, we present to consider both frame saliency and spatial preference to online generate the CRSR weight map and propose a simple yet effective saliency-embedded CF objective function to simultaneously optimize the filters and CRSR weight map in spatial-temporal domain. Extensive experiments validate that our content-related SR outperforms the classical SR, with higher tracking accuracy and almost two times faster speed.

Index Terms— Object Tracking, Correlation Filters, Content-Related Spatial Regularization, Saliency Guidance

1. INTRODUCTION

Visual object tracking is an important task in computer vision and has many applications, such as robotic service, human motion analyses and autonomous driving. One of the main challenges of object tracking is to address the targets' appearance change over time. Despite great progress in recent years, it remains a challenging problem while handling all factors from the background and targets themselves such as occlusions, deformations and illumination variations [1].

Recently, correlation filters (CF) tracking [2, 3], being one of the best tracking frameworks, has shown continuous performance improvement in terms of accuracy and robustness on various benchmarks [1, 4]. There are two main ways to enhance CF tracking, i.e. using more discriminative features and building more effective filter learning method. In the first way, besides intensity feature [2], HOG [3], Color [5], deep features [6] and feature fusion [7] are used to improve the CF tracking. These methods, however, do not consider an inherent drawback of CF, i.e. boundary effects introduced by circular shifting, thus learn less discriminative filters. In the second way, several trackers e.g. SRDCF [8], CCOT [9] and ECO [10], is designed to address this problem by using a spatially variant weight map to regularize correlation filters and has achieved the best performance on popular benchmarks, e.g. OTB [1] and VOT [4]. However, the weight map for spatial regularization (SR) is generated according to the bounding box of the target, given at the first frame and fixed during the whole tracking process, which loses sight of the object content information. Such designing is clearly not suitable for object tracking that usually address irregular, nonrigid and temporally changing objects, e.g. a player shown in Fig. 1. As shown in Fig. 1, a basketball player keeps running and changing during the whole sequence and is represented by a bounding box containing a lot of background information, which makes the spatial regularization weight map constructed from such bounding box become less effective. SRDCF thus cannot locate and wrap the player accurately.

To alleviate such problem, we propose content-related spatial regularization for correlation filters (CRSRCF), which introduces the saliency information and online learned filters into the SR weight map shown in Fig. 1 with the consideration of target content information, i.e. the shape and variation. As a result, CRSRCF can track the irregular, nonrigid and temporally changing targets accurately. Specifically, we first propose static content-related SR by introducing target saliency map into the SR weight map to highlight the target while suppressing the surrounding at the first frame. We then propose a simple yet effective saliency-embedded CF objective function to simultaneously optimize filters and SR weight map. Experiments results show that our approach helps SRDCF track irregular, nonrigid and variational target accurately and gets much better performance than several state-of-the-art trackers on OTB-2015 [1].

2. BACKGROUND

Discriminative correlation filters. Given a d-dimension feature map $\mathbf{X}$ extracted from an image region and a desired output $\mathbf{Y}$ labeling the object likelihood of each location in

* are the equal contribution authors, † is the corresponding author. This work is supported by NSFC 61671325, 61572354, 61672376.

978-1-5386-1737-3/18/$31.00

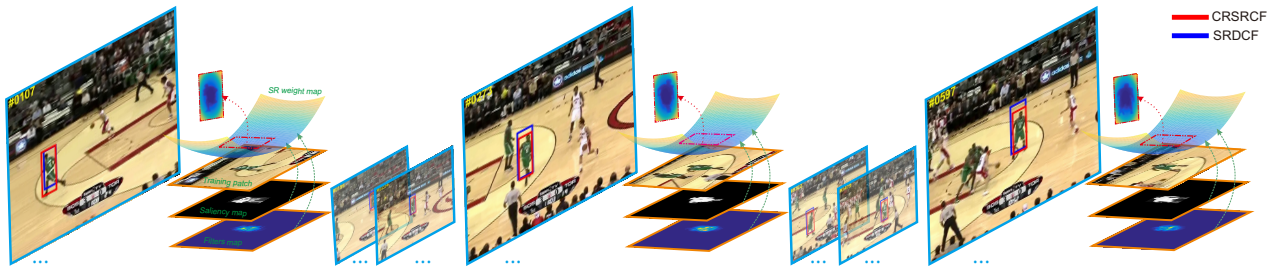

Fig. 1. Examples of our CRSR and SRDCF. CRSR updates the spatially variant weight map temporally via the saliency information and online learned filters. Our CRSRCF framework updates the correlation filters and the weight map alternately, which makes the weight map contain both object variation and the learned filters to adapt the object appearance suitably.

$\mathbf{X}$, DCF aims to learn multi-channel convolution filters $\mathbf{F}$ that can detect the same object in a new region. We denote $\mathbf{X}^l$ as the l-th channel $l \in \{1, ..., d\}$ of $\mathbf{X}$, and $\mathbf{F}^l$ as the l-th channel of $\mathbf{F}$ correspondingly. The correlation response of $\mathbf{F}$ and $\mathbf{X}$ is

$$\mathrm{C}(\mathbf{X}) = \sum_{l=1}^{d} \mathbf{X}^l * \mathbf{F}^l, \tag{1}$$

where '$*$' denotes circular convolution. The objective function of DCF is to minimize the L^2-error between the response $\mathrm{C}(\mathbf{X})$ and the desired output $\mathbf{Y}$

$$\mathrm{E}(\mathbf{F}) = \|\mathrm{C}(\mathbf{X}) - \mathbf{Y}\|^2 + \lambda \sum_{l=1}^{d} \left\|\mathbf{F}^l\right\|^2. \tag{2}$$

Spatially regularized discriminative correlation filters. SRDCF introduces a spatial regularization (SR) term within the DCF to address its boundary effects. SR replaces the regularization term in DCF and gets a new objective function

$$\mathrm{E}(\mathbf{F}) = \left\| \sum_{l=1}^{d} \mathbf{X}^l * \mathbf{F}^l - \mathbf{Y} \right\|^2 + \sum_{l=1}^{d} \left\|\mathbf{W} \odot \mathbf{F}^l\right\|^2, \tag{3}$$

where '$\odot$' is the element-wise multiplication. The spatial regularization weight map $\mathbf{W}$ penalizes $\mathbf{F}^l$ by assigning higher weights to the outside target region and lower weights to the inside target region to alleviate boundary effects. The weight map is constructed through following equation

$$\mathbf{W}_{\mathrm{SR}}(\mathbf{x}) = \xi + \eta \, (\frac{x - x_o}{w})^2 + \eta \, (\frac{y - y_o}{h})^2, \tag{4}$$

where $\mathbf{x} = (x, y)$ denotes the coordinate in the search region, and (x_o, y_o) denotes the center of the search region and $w \times h$ is the target size, while ξ and η are fixed parameters. By the construction of the weight map, the weight values are decided by spatial distance and target size in the first frame, without the consideration of the target shape and variation.

Other related work. Recently, several methods [11, 12, 10, 13, 14] are proposed to improve tracking accuracy by learning effective filters through mining information from target content. CFLB [11] proposes to alleviate boundary effects of CF by selecting effective training samples. CSR-DCF [12] improves CFLB by adopting a spatial reliability map to adapt the filters support to the parts of target suitable for tracking. C-COT [9] and ECO [10] improves the SRDCF by learning filters on the continuous domain and selecting effective features for efficient tracking. DSiam [13] maintains two online transformations for both target and background. Although success, all above methods do not consider the saliency of the target around its location, which is also useful information to separate the target from background.

3. CONTENT-RELATED SPATIAL REGULARIZATION FOR CF TRACKING

We propose content-related spatial regularization (CRSR) for CF by online constructing a saliency-embedded weight map. We first present static CRSR whose weight map is constructed from the saliency detection of the first frame and fixed during the whole tracking process. We then extent such method to a temporal version through a saliency-embedded objective function that can optimize both correlation filters and the weight map efficiently.

3.1. Static content-related spatial regularization

Given the first frame of a sequence, we crop a region that is 2^2 times larger than the bounding box of the target. We then adopt the saliency detection method proposed by [15] to detect saliency within the region and get a saliency map $\mathbf{S}$ with the values between $[0, 1]$. To make $\mathbf{S}$ work as a weight map that has low values on target and high values on background, we adjust it via

$$\mathbf{S}'(\mathbf{x}) = \frac{1}{1 + \mu \mathbf{S}(\mathbf{x})}, \tag{5}$$

where the values of $\mathbf{S}'(\mathbf{x})$ are within $[\frac{1}{1+\mu}, 1]$ with values of the object region approximating to $\frac{1}{1+\mu}$ and the surrounding region approximating to 1. We then obtain the weight map denoted as $\mathbf{W}_{\mathrm{CR}}$ via

$$\mathbf{W}_{\mathrm{CR}} = \mathbf{S}' \odot \mathbf{W}_{\mathrm{SR}}, \tag{6}$$

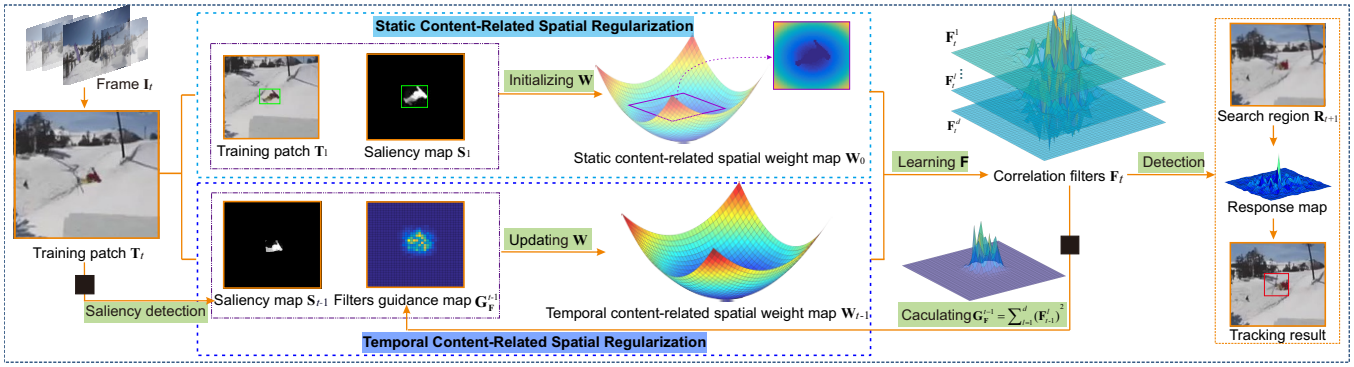

Fig. 2. Pipeline of CRSR based CF tracking. Static CRSR introduces the object saliency map of the first frame into the spatially variant weight map to highlight the target region. On this basis, temporal CRSR updates the weight map temporally through the saliency map and online learned filters to make the filters adapt the object variation better in filters learning.

where we pad $\mathbf{S}'$ with 1 to keep the same size with $\mathbf{W}_{\mathrm{SR}}$.

The objective function Eq. (3) thus can be replaced by

$$\mathrm{E}(\mathbf{F}) = \left\| \sum_{l=1}^{d} \mathbf{X}^l * \mathbf{F}^l - \mathbf{Y} \right\|^2 + \sum_{l=1}^{d} \left\| \mathbf{W}_{\mathrm{CR}} \odot \mathbf{F}^l \right\|^2. \quad (7)$$

With such objective function, we can construct a tracker that online learns the filters via the saliency-embedded weight map, i.e. $\mathbf{W}_{\mathrm{CR}}$. Since $\mathbf{W}_{\mathrm{CR}}$ is fixed during the tracking process after constructed at the first frame, we denote such method as static CRSR. We will show that static CRSR improves the tracking accuracy of classical SR while maintaining the speed in Section 4.

3.2. Temporal content-related spatial regularization

We have introduced the content information into the spatial weight map in the first frame (static CRSR), however, for the tracking problem, the shape and size of tracking object are always changing over time. The spatial weight map initialized in the first frame and fixed in subsequent frames is unconscionable. Here, we introduce the object variation information to update the spatial weight map temporally.

As discussed above, we propose a new objective function to replace the function Eq. (3)

$$\mathrm{E}(\mathbf{F}, \mathbf{W}_{\mathrm{T}}) = \left\| \sum_{l=1}^{d} \mathbf{X}^l * \mathbf{F}^l - \mathbf{Y} \right\|^2 + \sum_{l=1}^{d} \left\| \mathbf{W}_{\mathrm{T}} \odot \mathbf{F}^l \right\|^2 + \lambda_1 \|\mathbf{S} \odot \mathbf{W}_{\mathrm{T}}\|^2 + \lambda_2 \|\mathbf{W}_{\mathrm{T}} - \mathbf{W}_{\mathrm{SR}}\|^2, \quad (8)$$

where the $\mathbf{W}_{\mathrm{T}}$ and $\mathbf{S}$ are the spatial weight map and the object saliency map respectively, and $\mathbf{W}_{\mathrm{SR}}$ is the original weight map by Eq. (4). For this energy function, the first two terms are same to Eq. (3), while the last two terms are the bound terms of $\mathbf{W}_{\mathrm{T}}$. Specifically, the first bound term is to decrease the value of $\mathbf{W}_{\mathrm{T}}$ in object saliency region, due to the values of the saliency map in object region are positive while in non-object region are zeros. And the second bound term is to constraint the $\mathbf{W}_{\mathrm{T}}$ similar with the original spatial weight map to ensure the integral spatial regularization.

The objective function Eq. (8) involves two variables $\mathbf{F}$ and $\mathbf{W}_{\mathrm{T}}$, as for as $\mathbf{F}$, the function is same as Eq. (3), and as for as $\mathbf{W}_{\mathrm{T}}$, the objective function can be extracted as

$$\mathrm{E}(\mathbf{W}_{\mathrm{T}}) = \sum_{l=1}^{d} \left\| \mathbf{W}_{\mathrm{T}} \odot \mathbf{F}^l \right\|^2 + \lambda_1 \|\mathbf{S} \odot \mathbf{W}_{\mathrm{T}}\|^2 + \lambda_2 \|\mathbf{W}_{\mathrm{T}} - \mathbf{W}_{\mathrm{SR}}\|^2. \quad (9)$$

We utilize Eq. (9) to update the spatial weight $\mathbf{W}_{\mathrm{T}}$ temporally in each frame t, the gradient of E can be solved by

$$\frac{\partial \mathrm{E}}{\partial \mathbf{W}_{\mathrm{T}}} = 2\mathbf{W}_{\mathrm{T}} \odot \Big(\sum_{l=1}^{d} \left(\mathbf{F}^l\right)^2 + \lambda_1 \mathbf{S}^2 + \lambda_2\Big) - 2\lambda_2 \mathbf{W}_{\mathrm{SR}}. \quad (10)$$

By solving $\frac{\partial \mathrm{E}}{\partial \mathbf{W}_{\mathrm{T}}} = \mathbf{0}$ we get the closed-form solution

$$\mathbf{W}_{\mathrm{T}} = \frac{\lambda_2 \mathbf{W}_{\mathrm{SR}}}{\lambda_2 + (\mathbf{G}_{\mathbf{F}} + \lambda_1 \mathbf{G}_{\mathbf{S}})}, \quad (11)$$

where we denote $\mathbf{G}_{\mathbf{F}} = \sum_{l=1}^{d} \left(\mathbf{F}^l\right)^2$ as the filters guidance map and $\mathbf{G}_{\mathbf{S}} = \mathbf{S}^2$ as the saliency guidance map. By the solution of Eq. (11), we obtain the optimal $\mathbf{W}_{\mathrm{T}}$ which contains the object variation information for current frame.

We compare the solution of Eq. (11) of the t-th frame with the operation by Eq. (6) of the 1-st frame, and apply Eq. (11) to the first frame, due to the filters in initialization is not learned, the Eq. (11) can be transformed into

$$\mathbf{W}_{\mathrm{T}} = \frac{\lambda_2 \mathbf{W}_{\mathrm{SR}}}{\lambda_2 + \lambda_1 \mathbf{G}_{\mathbf{S}}}. \quad (12)$$

On the other hand, we substitute Eq. (5) into Eq. (6) and get

$$\mathbf{W}_{\mathrm{CR}} = \frac{\mathbf{W}_{\mathrm{SR}}}{1 + \mu \mathbf{S}}. \quad (13)$$

Therefore, we illustrate the $\mathbf{W}_{\mathrm{CR}}$ is the particular case of $\mathbf{W}_{\mathrm{T}}$, in the case of $\lambda_1 = \mu$, $\lambda_2 = 1$ and $\mathbf{G}_{\mathbf{S}} = \mathbf{S}$.

Furthermore, we observe the $\mathbf{W}_{\mathrm{T}}$ updated in each frame t by Eq. (11), which includes the saliency information obtained by the saliency guidance map, as well as the object spatial preference obtained by the online learned filters, due to the absolute values of the filters in the object region are higher and in the non-object region are approximate to zero. Ablation study in Section 4.2 will evaluate the impact of the saliency and filters guidance map respectively.

3.3. CRSR based CF tracking

Tracking algorithm. CRSR is a fundamental method by optimizing the SR, thus improves the performance of CF based trackers. We show our tracking framework in Algorithm 1 .

Algorithm 1: CRSR based CF tracking

Input: Frame $\{\mathbf{I}_t\}_1^T$, initial object bounding box $\mathbf{B}_1$
Output: Object bounding box of each frame $\{\mathbf{B}_t\}_2^T$

Initialization: initialize the correlation filters, initialize the spatial weight map $\mathbf{W}_0 = \mathbf{W}_{\mathrm{CR}}$ by Eq. (6).
Learn $\mathbf{F}_1$ by minimizing Eq. (8), update $\mathbf{W}_1$ by the solution Eq. (11) via the first frame with given bounding box, $t = 2$.
while $t \leq T$ **do**
    Crop an image region $\mathbf{R}_t$ from $\mathbf{I}_t$ at the last bounding box $\mathbf{B}_{t-1}$ and extract its feature map $\mathbf{X}_t$.
    Detect the object location $\mathbf{p}_t$ by calculating the response by Eq. (1) via $\mathbf{X}_t$ and $\mathbf{F}_{t-1}$ and the estimate the scale of the target as [8], thus get $\mathbf{B}_t$.
    Learn $\mathbf{F}_t$ by minimizing Eq. (8) using Gauss-Seidel iteration via $\mathbf{X}_t$ and $\mathbf{W}_{t-1}$.
    Update $\mathbf{W}_t$ by the closed-form solusion Eq. (11) via $\mathbf{S}_t$ and $\mathbf{F}_t$.
    $t = t + 1$
return $\{\mathbf{B}_t\}_2^T$.

Implementation details. We use HOG [16] as feature same as SRDCF [8]. We set $\mu = 2$ in Eq. (5). We calculate $\mathbf{G_F}$ in Eq. (11) and normalize it into $[0, 1]$ same as $\mathbf{G_S}$, and we set $\lambda_1 = 0.75,\ \lambda_2 = 1$ in Eq. (11). In our CRSRCF, we update the filters and SR weight map in every τ frames, and set $\tau = 2$ in our experiments. All parameters have straightforward interpretation, do not require fine-tuning.

4. EXPERIMENTAL RESULTS

4.1. Setup

Our Matlab implementation runs on an Intel Core i7 3.4GHz standard desktop. We validate our proposed method by performing comprehensive experiments on standard benchmarks OTB-2015 [1] containing 100 videos, for the OTB-2015 we use the one-pass evaluation (OPE) with precision and success plots metrics. We provide a comparison of our tracker with 8 well-known state-of-art methods including: KCF [3], SAMF [7], DSST [17], Staple [18], SRDCF [8], CSR-DCF [12], LMCF [19] and CFNet [20]. Most of them are designed with conventional hand-crafted features and the version of CFNet we use is *Baseline+CF-conv3* [20]. Among them, KCF is the fundamental CF based tracker, SAMF, DSST, SRDCF, Staple are follow-up trackers based on CF, CFNet, CSR-DCF and LMCF are the up to date CF based trackers.

Table 1. Ablation study of CRSRCF. Tracking results of SRDCF and SRDCF with the filters update frequency of 2 frames ($\mathrm{SRDCF}_{\tau=2}$), and the static content-related SRCF ($\mathrm{CRSRCF}_{\mathrm{static}}$) and temporal content-related SRCF (CRSRCF), as well as CRSRCF without saliency guidance map $\mathrm{CRSRCF}_{\mathrm{NS}}$, without filters guidance map $\mathrm{CRSRCF}_{\mathrm{NF}}$.

Trackers	Prec.	Succ.rate	Trackers	Prec.	Succ.rate
SRDCF	78.9	59.8	$\mathrm{CRSRCF}_{\mathrm{NS}}$	77.8	60.1
$\mathrm{SRDCF}_{\tau=2}$	77.0	58.7	$\mathrm{CRSRCF}_{\mathrm{NF}}$	78.2	59.9
$\mathrm{CRSRCF}_{\mathrm{static}}$	79.4	60.1	CRSRCF	**79.5**	**60.7**

4.2. Ablation study

In this section, we conduct ablation analysis to compare the performance of our static and temporal CRSRCF with the SRDCF. In addition, we also evaluate the impact of saliency guidance map and filters guidance map in Eq. (11) of temporal CRSR. Lastly, we discuss the influence of filters update frequency τ of CRSRCF and SRDCF.

The average distance precision score at 20 pixels and the area-under-the-curve (AUC) score for success rate on OTB-2015 of different trackers are shown in Table 1 . We first evaluate our static CRSR in Section 3.1 and temporal CRSR in Section 3.2 respectively on OTB-2015, AUC score for success rate rise from 59.8% (SRDCF) to 60.1% ($\mathrm{CRSRCF}_{\mathrm{static}}$) and 60.7% (CRSRCF). Furthermore, as shown in Table 1, we set the saliency guidance map in Eq. (11) to a constant map with uniform values within the bounding box and zeros elsewhere ($\mathrm{CRSRCF}_{\mathrm{NS}}$), which results in a drop of precision and success rate. Replacing the filters guidance map in Eq. (11) by the constant map ($\mathrm{CRSRCF}_{\mathrm{NF}}$) also results in a significant performance drop of the AUC score compared to CRSRCF. Finally, in CRSRCF we update the filters and spatial weight map in every 2 frames, we change the filters update frequency of SRDCF to 2 frames ($\mathrm{SRDCF}_{\tau=2}$) accordingly, the AUC score for success rate drops from 59.8% to 58.7%, which indirectly illustrates that our alternately update process of the spatial weight map and the correlation filters improves the robustness and adaptation of the filters.

4.3. Comparative results

State-of-the-art comparison. We provide a comparison of our approach with the state-of-the-art trackers. Fig. 3 shows the distance precision and success plot over all 100 videos

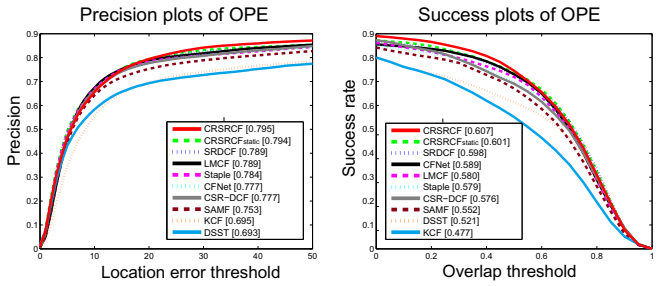

Fig. 3. Precision plots (left) and success plots (right) showing a comparison with state-of-the-art methods on OTB-2015. The legend contains the average distance precision score at 20 pixels and the AUC score for each tracker.

of OTB-2015. The left figure shows the precision of 8 comparative trackers and our CRSRCF, static CRSRCF tracker. Our CRSRCF tracker obtains the best performance at the precision of 79.5%, the following static CRSRCF of 79.4%. The right figure shows the success plot of the 10 trackers, among the state-of-the-art methods, SRDCF provides the best results with an AUC score of 59.8%, the following CFNet tracker achieves an AUC score of 58.9%. Our static CRSRCF, CRSRCF outperform other 8 trackers and obtain an AUC score of 60.1% and 60.7% for success plot respectively.

Attribute based comparison. We also perform an attribute based analysis of our tracker on the OTB-2015 dataset. The 100 videos are annotated with 11 attributes: scale variation, deformation, out-of-plane rotation, occlusion, etc. Fig. 5 shows example success plots of four typical attributes to illustrate the advantage of our method. The results indicate that our CRSRCF tracker is effective in handling scale variation, in which SRDCF get the state-of-the-art performance, due to the proposed SR enables an expansion of the image region used for training the filter. However, it does not perform well as our CRSRCF with the temporal object saliency information to highlight the object appearance, which makes significant improvements in object scale variation. We have also found similar performance in deformation and out-of-plane rotation, where Staple proposed a combination of template and histogram scores to deal with the object deformations and LMCF proposed multimodal target detection to prevent model drift introduced by background noise. But our CRSRCF with the temporal saliency information and online learned filters integrated into the framework and optimized as a whole, which introduces the object variation in filters learning therefor get the best performance in these two scenarios. Finally, for occlusion, our filters update frequency contributes to stabilizing the filters learning, especially in scenarios where the object is affected by sudden changes, such as occlusions.

Qualitative results. Fig. 4 shows the tracking results of CRSRCF and other four CF-based trackers, i.e. KCF [3], Staple [18], CSR-DCF [12] and SRDCF [8] on OTB-2015 in cases of scale variation (first row-*Human9*), deformation (second row-*shaking1*), rotation (third row-*Sylvester*) and occlusion (last row-*Human3*). In CRSRCF, we temporally update spatial weight map via the saliency information and online learned filters, which makes our tracker obtain more object content and track irregular object accurately. Experimental results confirm the strength of our method in tracking the irregular, nonrigid and variational target.

Fig. 4. Qualitative evaluation of our CRSRCF, SRDCF, CSR-DCF, Staple, KCF on three representative sequences.

Table 2. Success rates (% at IoU>0.50 and AUC score) of CRSRCF versus related trackers, and the weighted average speed on OTB-2015 dataset.

	KCF (PAMI2015)	CFLB (CVPR2015)	SRDCF (ICCV2015)	CSR-DCF (CVPR2017)	CRSRCF (Ours)
Succ.rate (IoU)	55.1	44.7	72.8	69.1	74.5
Succ.rate (AUC)	47.7	41.5	59.8	57.6	60.7
Avg.FPS	153.46	87.1	3.5	7.5	6.4

4.4. Speed analysis

Tracking speed is a considerable factor in tracking problems. Table 2 compares several related and well-known CF based trackers. Among these trackers, KCF [3] is the fundamental CF based tracker, and CFLB [11], CSR-DCF [12], SRDCF [8] are all proposed to adapt the filters support to the part of the object suitable for CF tracking. We calculate the average speed with the weights of the number of frames per video. Speed and performance measures on OTB-2015 are shown in Table 2 , results show that our CRSR can improve the original SRDCF tracker in both performance and speed as well as achieve the state-of-the-art tracking performance.

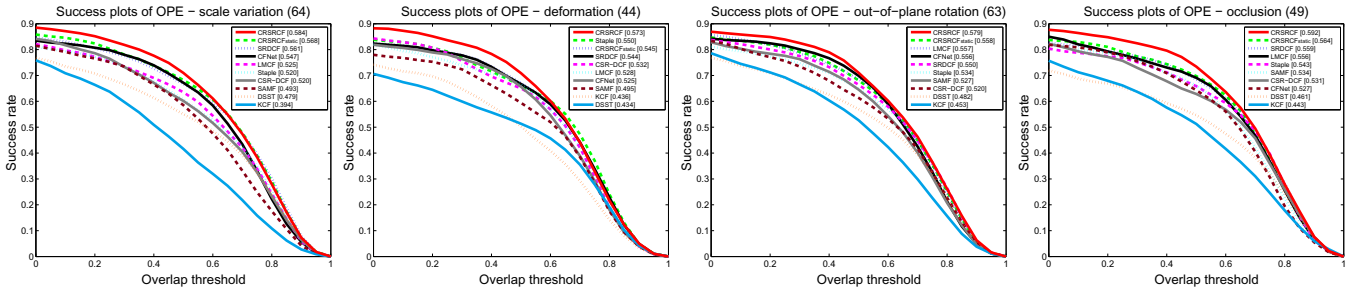

Fig. 5. Attribute-based analysis of our approach on the OTB-2015 dataset with 100 videos. Success plots are shown for four attributes. Our approach demonstrates superior performance compared to existing trackers in these scenarios.

5. CONCLUSION

In this paper, we have revisited the spatially regularized correlation filters (SRDCF) to conquer its drawbacks in tracking irregular, nonrigid and rapidly-changing objects. We have proposed an effective content-related spatial regularization for correlation filters (CRSRCF) by online constructing image saliency related spatial regularization (SR) weight map and a fast way to online learn both the filters and the regularization map. Experimental results on benchmark OTB-2015 dataset have showed that our approach significantly outperformed the state-of-the-art SRDCF with higher accuracy, robustness and speed. Our tracker performs especially well in tracking irregular and nonrigid objects. In the future, we want to investigate how to generally apply the proposed CRSR to more CF trackers, e.g. [21, 22], to achieve better tracking performance. We can further improve CRSR by using structure information from superpixel and object segmentation [23, 24].

6. REFERENCES

[1] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE TPAMI*, vol. 37, no. 9, pp. 1834–1848, 2015.

[2] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *CVPR*, 2010.

[3] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE TPAMI*, vol. 37, no. 3, pp. 583–596, 2015.

[4] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE TPAMI*, vol. 38, no. 11, pp. 2137–2155, 2016.

[5] M. Danelljan, F. S. Khan, M. Felsberg, and J.V.D. Weijer, "Adaptive color attributes for real-time visual tracking," in *CVPR*, 2014.

[6] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Hierarchical convolutional features for visual tracking," in *ICCV*, 2015.

[7] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *ECCV*, 2014.

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

[9] M. Danelljan, F. S. Robinson, A.and Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *ECCV*, 2016.

[10] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *CVPR*, 2016.

[11] H. K. Galoogahi, T. Sim, and S. Lucey, "Correlation filters with limited boundaries," in *CVPR*, 2015.

[12] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *ICCV*, 2017.

[13] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *ICCV*, 2017.

[14] Q. Guo, W. Feng, C. Zhou, C.-M. Pun, and B. Wu, "Structure-regularized compressive tracking with online data-driven sampling," *IEEE TIP*, vol. 26, no. 12, pp. 5692–5705, 2017.

[15] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *CVPR*, 2015.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[17] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *BMVC*, 2014.

[18] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *CVPR*, 2015.

[19] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," in *CVPR*, 2017.

[20] J. Valmadre, L. Bertinetto, Joao H., A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *CVPR*, 2017.

[21] C. Zhou, Q. Guo, L. Wan, and W. Feng, "Selective object and context tracking," in *ICASSP*, 2017.

[22] Z. Chen, Q. Guo, L. Wan, and W. Feng, "Background-suppressed correlation filters for visual tracking," in *ICME*, 2018.

[23] Q. Guo, S. Sun, X. Ren, F. Dong, B. Z. Gao, and W. Feng, "Freqeuncy-tuned active contour model," *Neurocomputing*, vol. 275, no. 31, pp. 2307–2316, 2018.

[24] W. Feng, J. Jia, and Z. Liu, "Self-validated labeling of markov random fields for image segmentation," *IEEE TPAMI*, vol. 32, pp. 1871–1887, 2010.