IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 2, 2019, accepted December 30, 2019, date of publication January 7, 2020, date of current version January 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964716

Spatial Adaptive Regularized Correlation Filter for Robust Visual Tracking

LEI PU[1], **XINXI FENG**[2], **AND ZHIQIANG HOU**[3]
[1]Graduate College, Air Force Engineering University, Xi'an 710077, China
[2]Information and Navigation College, Air Force Engineering University, Xi'an 710077, China
[3]School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Corresponding author: Zhiqiang Hou (hzq@xupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571458 and Grant 61703423.

ABSTRACT Correlation filter is a simple yet efficient method to deal with the visual tracking task. However, the unwanted boundary effects hinder further performance improvement. Spatially Regularized DCF (SRDCF) has been proposed to address this problem with a pre-computed spatial penalty matrix, which improves the tracking performance greatly. In this paper, aiming to achieve more accurate spatial regularization, we present our spatial adaptive regularized correlation filter (SARCF). A coarse-to-fine scale estimation approach is proposed to change the spatial penalty area, which can efficiently deal with large scale variation. Moreover, temporal regularization is introduced for long-term tracking. Experimental results show that the proposed algorithm outperforms most advanced algorithms in tracking accuracy and success rate.

INDEX TERMS Visual tracking, correlation filter, scale estimation, boundary effect, spatial regularization.

I. INTRODUCTION

Visual tracking is a fundamental and challenging problem in computer vision with many applications, such as robotic service, video surveillance, human interaction, motion analysis, and autonomous driving, to name a few. Given only the initial location and scale, tracking task is to estimate the state of a target throughout the sequence. Despite the significant effort that has been made in recent years, it is still a difficult problem due to compliment situations like scale variation, full occlusion, fast motion and so on.

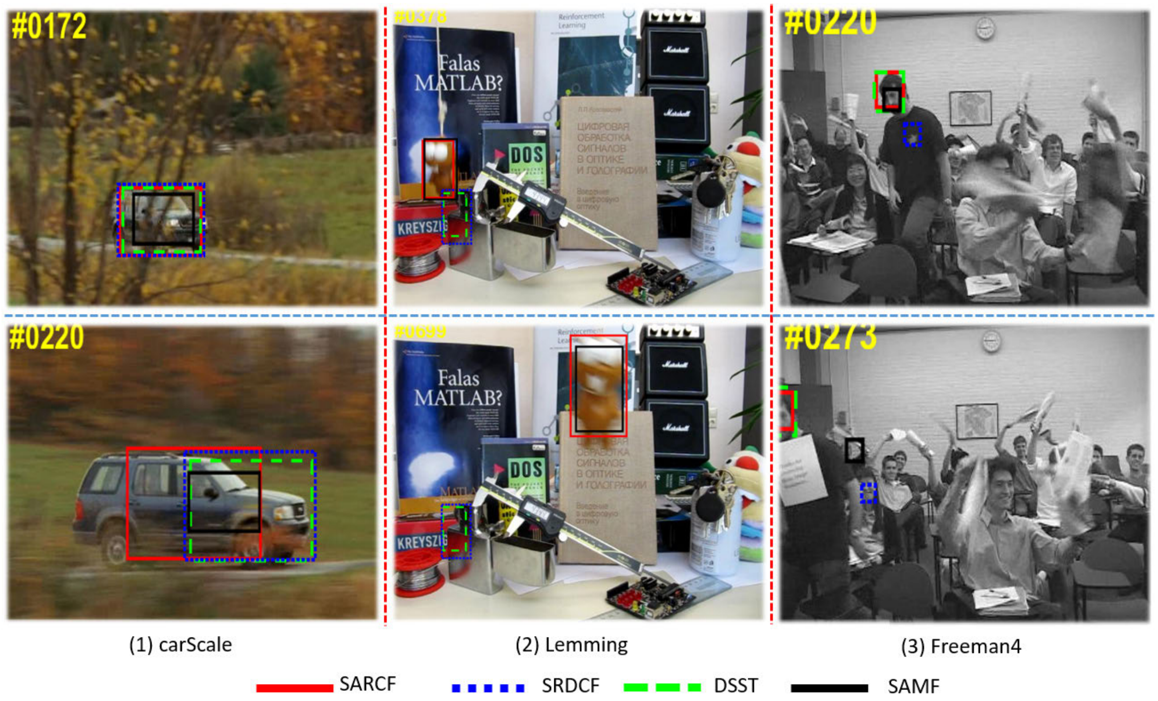
Recently, correlation filters have been very popular in the visual tracking community due to their high speed and well performance. Compare to the traditional methods that are troubled by the lack of training data, correlation filters can be learned with ample training samples. In 2010, Bolme *et al.* [3] firstly proposes the MOSSE tracker with the speed over 700 frames per second. Furthermore, many improved versions of MOSSE tracker have been proposed. One way is to improve the representation ability of features [6], [11], [21], the other way is to optimize the filter learning [7], [13], [14].

The standard correlation filters are learned from circular shifted examples, which can be processed efficiently in the frequency domain via Fast Fourier Transform (FFT). However, the periodic assumption of training samples produces unwanted boundary effects, which have been shown to have a drastic impact on tracking performance. Due to the circularity, the filter is trained on many inaccurate and unrealistic image patches which reduce the discriminative capability. Moreover, the detection scores are heavily influenced by the periodic assumption, which limits the searching region size at the location step.

The boundary effect problems are proven to have a significant impact on tracking performance. Recently, many methods have been proposed to address the disadvantage of learning from shifted examples [7], [13], [14], [16], [23]. The pioneering work is proposed by Galoogahi *et al.* [14] who use a larger training area and a smaller filter size to learn correlation filter from cropped samples, which can significantly improve the number of samples that are not contaminated by boundary effects. Moreover, Danelljan *et al.* [7] propose the well-known SRDCF tracker to penalize the boundary of the filter coefficient with an inverse Gaussian-shaped spatial map. But both methods use the fixed rectangle or invariable weights to implement regularization and not changed during filter learning, which cares little to the target scale variation.

In this paper, we introduce the SARCF, the Spatial Adaptive Regularized Correlation Filter. By incorporating scale estimation and spatial regularization, our tracker can achieve better scale estimation and a more robust filter model. On the

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.



(1) carScale (2) Lemming (3) Freeman4

FIGURE 1. Comparisons of the proposed tracking algorithm with the most related correlation filter trackers (SRDCF [7], DSST [5] and SAMF [18]) on the *carScale, Lemming*, and *Freeman4* sequences. These methods perform differently as various scale estimation methods and regulations are used for the scenes such as complex background, scale variations, and partial occlusions.

one hand, we introduce a rectangle-shaped spatial regularization component to address the unwanted boundary effects within the CF formulation. On the other hand, we construct a scale pyramid representation to learn discriminate correlation filter that can estimate the target scale efficiently. Then we update the spatial penalty area using the estimated target scale. Moreover, we introduce the temporal regularization for long-term tracking, and the filter learning function can be solved by the ADMM algorithm.

We evaluate our tracker on a popular benchmark dataset [29], The result demonstrates very competitive accuracy of our method compared to many state-of-the-art CF-based trackers and shown notable improvements towards SRDCF.

The contributions of this paper are as follows:

- A spatial adaptive regularization correlation filter is presented by establishing the relationship between scale estimation and spatial penalty area.
- A coarse-to-fine scale estimation approach is proposed to change the spatial penalty area.
- Our proposed SARCF can be effectively optimized via the alternating direction method of multipliers (ADMM), where each sub-problem has the closed-form solution.
- Our tracker achieves very remarkable performance compared to many state-of-the-art CF-based trackers and shown significant improvements towards SRDCF.

II. RELATED WORKS

This section provides a brief review of some relevant tracking methods. A more comprehensive review can be found in [4], [17], [25].

A. VISUAL TRACKING

Visual tracking can be categorized into two groups, generate trackers and discriminate trackers. The generate trackers construct an appearance model to represent the target and search for it in the detection area with the highest similarity. Many generate algorithms have been presented, such as mean shift tracker, incremental tracker (IVT), multi-task tracker (MTT) and low-rank dictionary learning based tracker [33], to name a few. Different to generate tracker, the discriminate tracker formulates visual tracking task as a regression [20] or classification problem. Many recently published algorithms come into this category including support vector tracking [12], [28], [30], deep learning based tracking [15], [22], [26], [27] and correlation filter based trackers [3], [7], [11]

B. CORRELATION FILTER

Correlation filters have achieved great success in visual tracking due to the balance of computational efficiency and robustness. The first attempt to employ CFs for visual tracking starts with MOSSE [3], which only uses single channel grayscale feature to learn the filter and achieves an impressive speed over 700 FPS. Notable improvements have been by introducing better features or more robust filter models. The CSK tracker [10] exploits the circulant structure of training samples and add the kernel trick into the CF formulation. The following KCF tracker [11] extends the single channel grayscale feature to the multi channels HOG descriptors. For robust tracking, more discriminative multi-channel features are widely used, such as color attributes [6], deep CNN features [8], [21], [24]. In this paper, we combine color attributes and HOG descriptors to learn a filter that is robust to both color changes and deformations. To cope with the scale changes, several scale-adaptive trackers [5], [18] are further investigated. Different from SAMF and DSST that only estimate simple axis-aligned bounding boxes, Li *et al.* [19] employ an efficient phase correlation scheme to deal with both scale and rotation changes simultaneously in log-polar coordinates.

C. BOUNDARY EFFECTS

Unwanted boundary effects in correlation filter based tracking lead to inaccurate representation and insufficient discrimination of the object, especially in the cluttering background. The main reason behind this is that the training data is only generated from the regions that are tightly surrounding the object. But all other information belonging to the background is discarded. Some works [7], [16], [32] wanted to solve this limitation by investigating the relationship between the training samples and filters, e.g. the filter coefficients, are penalized in terms of spatial locations [7] to achieve more robust appearance modeling suitable for large variations. Other attempts focused on the binary mask as a solution, Galoogahi *et al.* [13], [14] try to solve this challenge by using dot product operation on the image patches with a fixed binary mask containing the object regions. Similar to using an alternating direction method of multipliers(ADMM) for constrained optimization. Different from those methods, our SARCF establishes the bridge between scale estimation and spatial regularization, which can learn a more robust appearance model and better scale estimation results.

III. PROPOSED METHOD

In this section, we first revisit the CF tracker in Sec. III-A. The proposed scale estimation method is described in Sec. III-B. Our SARCF model is presented in Sec. III-C. Finally, we describe our tracking framework in Sec. III-D.

A. CORRELATION FILTER

Before the detailed discussion of the proposed method, we first revisit the basic correlation filter framework. The tracking-by-detection framework is very popular in tracking community. The main task of this framework is to train a model that can better distinguish target from background. In the past years, the number of training samples is a large problem. However, by using the circular matrix for dense sampling and transferring to frequency domain for fast calculation, the correlation filters have gained much attention. The CFs can be seen as the following ridge regression problem:

$$\mathcal{L}(\mathrm{h}) = \min_{\mathrm{h}} \|\mathrm{f} \otimes \mathrm{h} - \mathrm{g}\|^2 + \lambda \|\mathrm{h}\|^2 \tag{1}$$

Here, f denotes the training samples, and the learned filter is represented by h, ⊗ is the spatial correlation operator. g is the desired correlation response, λ is a regularization parameter. To solve the problem effectively, we express the objective function Eq.(1) in frequency domain (using Parseval's theorem):

$$\mathcal{L}(\mathrm{h}) = \min_{\mathrm{h}} \left\| \sum_{\mathrm{d}=1}^{\mathrm{D}} \mathrm{diag}\left(\hat{\mathrm{f}}^{\mathrm{d}}\right) \overline{\hat{\mathrm{h}}}^{\mathrm{d}} - \hat{\mathrm{g}} \right\|^2 + \lambda \sum_{\mathrm{d}=1}^{\mathrm{D}} \left\| \hat{\mathrm{h}}^{\mathrm{d}} \right\|^2 \tag{2}$$

where, d is the channel index, and the D is the channel number. the $\hat{\mathrm{f}}$ denotes the Fourier transformation of f. In this way, the circular correlation in Eq.(1) is replaced by point dot production, which significantly reduces the computation. The solution of one single channel filter can be deduced as follows:

$$\hat{\mathrm{h}}^{\mathrm{d}} = \left(\mathrm{diag}\left(\hat{\mathrm{f}}^{\mathrm{d}}\right) \overline{\hat{\mathrm{g}}}\right) \odot^{-1} \left(\sum_{\mathrm{d}=1}^{\mathrm{D}} \mathrm{diag}\left(\hat{\mathrm{f}}^{\mathrm{d}}\right) \overline{\hat{\mathrm{f}}}^{\mathrm{d}} + \lambda \right) \tag{3}$$

However, there are many limitations of the aforementioned correlation filter framework. The main limitation is that the filter is trained on many unreal samples generated by circular shift. This problem reduces the discrimination power and hinders to learn the filter from a larger region. The circulant shifted samples in CF-based trackers always suffer from boundary effect. Danelljan *et al.* [7] propose a spatial regularization term to penalize the CF coefficients depending on their spatial locations, The proposed SRDCF is formulated by minimizing the following objective:

$$\arg\min_{\mathbf{h}} \sum_{k=1}^{T} \alpha_k \left\| \sum_{d=1}^{D} \mathbf{f}_k^d \otimes \mathbf{h}^d - \mathbf{g}_k \right\|^2 + \sum_{d=1}^{D} \left\| \mathbf{w} \odot \mathbf{h}^d \right\|^2 \tag{4}$$

where ⊙ denotes the Hadamard product, **w** is the spatial regularization matrix. α_k indicates the weight to each sample $\mathbf{f}_k$ and is set to emphasize more to the recent samples. T is the number of samples.

Another limitation is the scale estimation. Nowadays many works have focused on the target location and care little to target estimation. In this paper, we take the spatial regularization and scale estimation into account and propose our SARCF tracker.

B. COARSE-TO-FINE SCALE ESTIMATION

Scale variation is a common problem in visual tracking. Current scale estimation methods mostly adopt a scale pyramid strategy. There are two most widely used approaches: DSST [5] and SAMF [18]. DSST adopts an independent scale filter to estimate the best scale after the target position is obtained. Unlike DSST, SAMF uses multiple-scale templates to seek the optimal position and scale at the same time. To further improve the accuracy of scale estimation, this paper combines the advantages of these two methods. First, SAMF

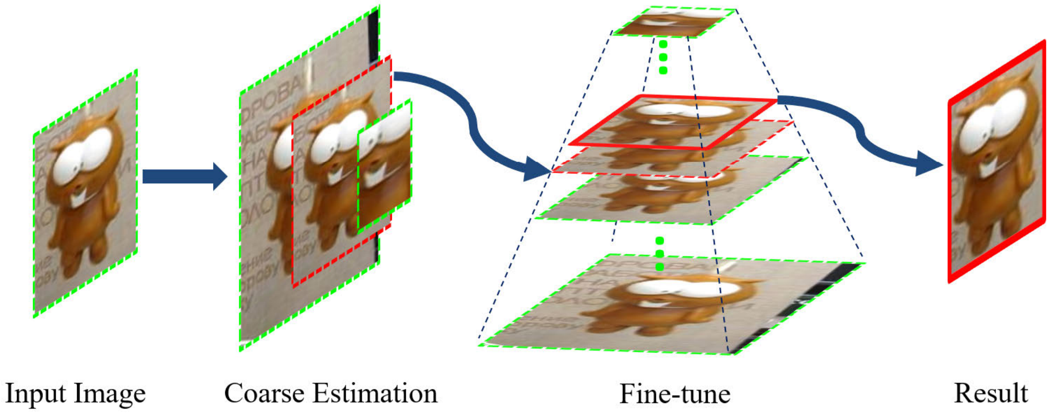


FIGURE 2. The proposed coarse-to-fine scale estimation. First, SAMF method is used for coarse estimation to get the approximate target scale, then DSST method is used for fine-tuning and correlation filter regression.

method is used for coarse estimation to get the approximate target scale, then DSST method is used for fine-tuning and correlation filter regression.

The target size is s_{t-1} in the last frame, and the scaling pool is defined as $\mathbf{S}_c = \{t_1, t_2, \ldots, t_k\}$. In this paper, we set $k = 3$. At present frame, we sample k sizes in $\{t_i \mathbf{s_t} | t_i \in \mathbf{S}\}$. We adopt bilinear interpolation method to adjust sizes of samples to be the same. The response can be obtained as follows:

$$\arg\max \mathbf{F}^{-1}\hat{\mathbf{h}}\left(\mathbf{f}^{t_i}\right) \tag{5}$$

where $\mathbf{f}^{t_i}$ is the features of i_{th} samples. Once the coarse scale estimation is finished, we learn an independent scale correlation filter based on it to achieve the accurate scale estimation, as shown in Figure 2.

After obtaining the target scale s_t, we adopt it to adjust the spatial regularization area adaptively. As we suppress all of the other positions except the target area in filter training, when the target becomes smaller, if we still suppress in this way, it will result in too few parameters, which makes the model easy to underfitting. For this reason, this paper only adjusts the regularization area when the target becomes larger, and does not adjust the area when the target becomes smaller. Set m as the spatial penalty matrix, which can be obtained as below:

$$\mathrm{m}_{i,j} = \begin{cases} 0.001, & \textit{if } \mathrm{m}_{i,j} \in T \\ 100000, & \textit{otherwise} \end{cases} \tag{6}$$

where T represents the target area, i, j are the position index.

C. SPATIAL ADAPTIVE REGULIZED CORRELATION FILTER

As multi-channel features are widely used in visual tracking, we extend Eq.(1) to multi-channel features:

$$\arg\min_{\mathrm{h}} \left\| \sum_{d=1}^{D} \mathrm{f}_t^d \otimes \mathrm{h}^d - \mathrm{g} \right\|^2 + \lambda \sum_{d=1}^{D} \left\| \mathrm{h}^d \right\|^2 \tag{7}$$

where t denotes the frame index. Then we adopt the spatial penalty matrix obtained in Sec. III-C to implement spatial regularization. In order to achieve long-term tracking, we introduce a temporal regularization term into the objective function, resulting in our scale adaptive regularized correlation filter model:

$$\arg\min_{\mathrm{h}} \left\| \sum_{d=1}^{D} \mathrm{f}_t^d \otimes \mathrm{h}^d - \mathrm{g} \right\|^2 + \sum_{d=1}^{D} \left\| \mathrm{m}_t \odot \mathrm{h}^d \right\|^2 + \gamma \left\| \mathrm{h} - \mathrm{h}_{t-1} \right\|^2 \tag{8}$$

where m_t is the spatial adaptive regularization matrix in the (t)-th frame, h_{t-1} denotes the CFs utilized in the $(t-1)$-th frame, and γ denotes the regularization parameter. To solve the Equation (8), we introduce an auxiliary variable $\mathrm{h}_c = \mathrm{h}$ and formulate an augmented Lagrangian equation as:

$$\begin{aligned} \arg\min_{\mathbf{h}} & \left\| \sum_{d=1}^{D} \mathrm{f}_t^d \otimes \mathrm{h}^d - \mathrm{g} \right\|^2 + \sum_{d=1}^{D} \left\| \mathrm{m}_t \odot \mathrm{h}_c^d \right\|^2 \\ & + \sum_{d=1}^{D} \left(\mathrm{h}^d - \mathrm{h}_c^d \right)^{\mathrm{T}} \mathrm{w}^d + \lambda \sum_{d=1}^{D} \left\| \mathrm{h}^d - \mathrm{h}_c^d \right\|^2 \\ & + \gamma \left\| \mathrm{h} - \mathrm{h}_{t-1} \right\|^2 \end{aligned} \tag{9}$$

where w, γ are the Lagrange multiplier and penalty factor, λ is the stepsize. Set $\mathrm{r} = \frac{1}{\lambda}\mathrm{w}$, we can reformulate Equation (9) as

$$\begin{aligned} \arg\min_{h} & \left\| \sum_{d=1}^{D} \mathrm{f}_t^d \otimes \mathrm{h}^d - \mathrm{g} \right\|^2 + \sum_{d=1}^{D} \left\| \mathrm{m}_t \odot \mathrm{h}_c^d \right\|^2 \\ & + \lambda \left\| \mathrm{h} - \mathrm{h}_c + \mathrm{r} \right\|^2 + \gamma \left\| \mathrm{h} - \mathrm{h}_{t-1} \right\| \end{aligned} \tag{10}$$

Equation (10) can be solved iteratively using the ADMM technique. By this way, Equation (10) can be divided into following subproblems:

$$\begin{aligned} \mathrm{h}^{(i+1)} = \arg\min_{\mathrm{h}} & \left\| \sum_{d=1}^{D} \mathrm{f}_t^d \otimes \mathrm{h}^d - \mathrm{g} \right\|^2 + \lambda \left\| \mathrm{h} - \mathrm{h}_c + \mathrm{r} \right\|^2 \\ & + \gamma \left\| \mathrm{h} - \mathrm{h}_{t-1} \right\|^2 \end{aligned} \tag{11}$$

$$\mathrm{h}_c^{(i+1)} = \arg\min \sum_{d=1}^{D} \left\| \mathrm{m}_t \odot \mathrm{h}_c^d \right\|^2 + \gamma \left\| \mathrm{h} - \mathrm{h}_c + \mathrm{r} \right\|^2 \tag{12}$$

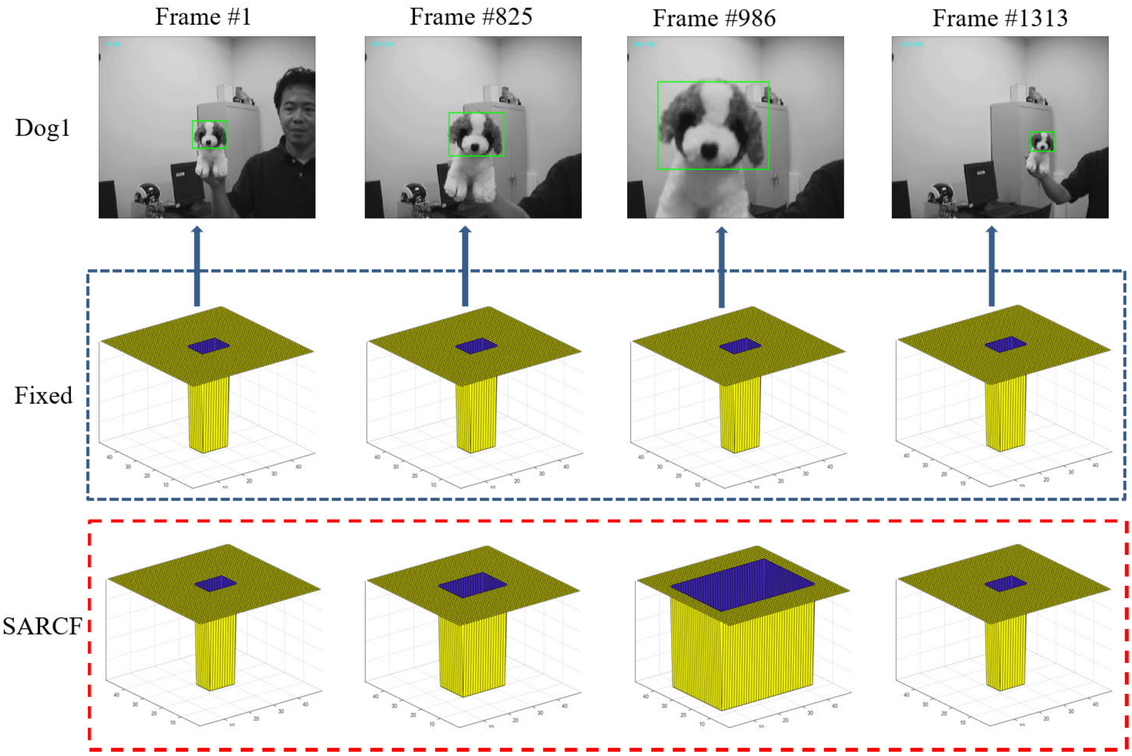


FIGURE 3. The illustration of the proposed scale adaptive regularized correlation filter (SARCF). When the target size is changed, the spatial penalty area is modified accordingly.

$$\mathbf{r}^{(i+1)} = \mathbf{r}^{(\mathrm{i})} + \mathrm{h}^{(\mathrm{i+1})} - \mathrm{h}_{\mathrm{c}}^{(\mathrm{i+1})} \tag{13}$$

Firstly, the Equation (11) is transformed into Fourier domain by using the Parseval's theorem:

$$\hat{\mathrm{h}}^{(\mathrm{i+1})} = \arg\min_{\hat{\mathrm{h}}} \left\| \sum_{\mathrm{d=1}}^{\mathrm{D}} \hat{\mathrm{f}}_{\mathrm{t}}^{\mathrm{d}} \odot \hat{\mathrm{h}}^{\mathrm{d}} - \hat{\mathrm{g}} \right\|^2 + \lambda \left\| \hat{\mathrm{h}} - \hat{\mathrm{h}}_{\mathrm{c}} + \hat{\mathrm{r}} \right\|^2 + \gamma \left\| \hat{\mathrm{h}} - \hat{\mathrm{h}}_{\mathrm{t-1}} \right\|^2 \tag{14}$$

where $\hat{\mathrm{h}}$ represents the Fourier transform of h. To simplify the problem, we can take Equation (14) from another perspective. The features $\hat{\mathrm{f}}$ can be formulated as $c_h \times c_w$ cells and each contains one D-dimensional vector. $\mathcal{V}_i(\mathbf{f}) \in \mathbb{R}^D$ denotes the i-th cells. Then Equation (14) can be divided into $c_h \times c_w$ subproblems. Finally, we can get the close solution of $\mathcal{V}_i(\hat{\mathbf{h}})$ as

$$\mathcal{V}_i(\hat{\mathbf{h}}) = \frac{1}{\gamma + \lambda}\left(I - \frac{\mathcal{V}_i\left(\hat{\mathbf{f}}_t\right)\mathcal{V}_i\left(\hat{\mathbf{f}}_t\right)^{\top}}{\gamma + \lambda + \mathcal{V}_i\left(\hat{\mathbf{f}}_t\right)^{\top}\mathcal{V}_i\left(\hat{\mathbf{f}}_t\right)}\right)\mathbf{p} \tag{15}$$

where $\mathbf{p} = \mathcal{V}_i\left(\hat{\mathbf{f}}_t\right)\hat{\mathrm{g}}_i + \lambda\mathcal{V}_i(\hat{\mathrm{h}}_{\mathrm{c}}) - \lambda\mathcal{V}_i(\hat{\mathbf{r}}) + \gamma\mathcal{V}_i\left(\hat{\mathbf{h}}_{t-1}\right)$. Equation (15) can be solved efficiently with only dot product operations. Moreover, the closed-form solution of g can be obtained as below,

$$\hat{\mathrm{h}}_{\mathrm{c}} = \left(\mathbf{M}^{\top}\mathbf{M} + \gamma I\right)^{-1}(\gamma\mathbf{h} + \gamma\mathbf{r}) \tag{16}$$

where $\mathbf{M}$ denotes the diagonal matrix concatenated with D diagonal matrices Diag(m).

D. ONLINE TRACKING

In this subsection, we describe our proposed tracking framework based on adaptive spatial regularization. The illustration of the proposed tracker is shown in Figure 3.

1) LOCALIZATION STAGE

In the localization step, the position of the target in a new frame t is estimated by the previous filter and current frame. Firstly, we extract the multi-channel features and feed them into correlation filters, and the target is localized by summing the correlation response maps, the maximum value of the responses is the center position of the target. To estimates the target scale, we apply the proposed coarse-to-fine method to estimate the scale changes.

2) MODEL UPDATE STAGE

Once the position is estimated, the training region centered at the position is extracted to update the model and estimate the spatial penalty matrix. Instead of linear combination updating, we adopt a temporal regularization method to passively update the CF model. On the one hand, this method can make the updated model similar to the previous one. On the other hand, it can also guarantee the new sample be used to modify the CF model, thus leading to more robust models in the case

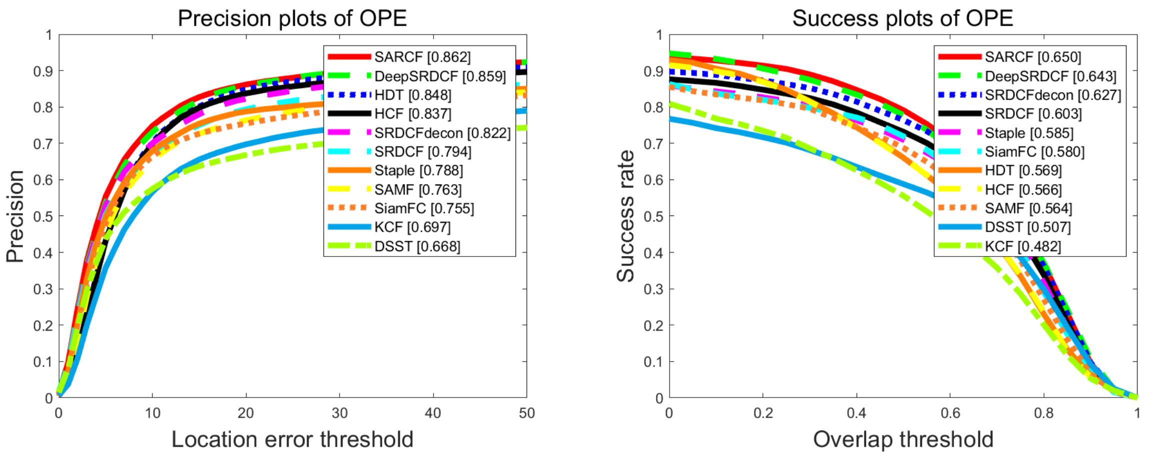


FIGURE 4. **Precision and success plots on OTB2015 [29] dataset. The legend contains the average distance precision score at 20 pixels and the area-under-the-curve (AUC) score for each tracker. Our proposed algorithm performs favorably against the state-of-the-art trackers.**

of large appearance variations. Finally, the training filter is used for the target detection and tracking of the next frame.

IV. EXPERIMENTS

To validate the effectiveness and robustness of our proposed tracker, we present extensive experimental evaluations on the popular OTB2015 [29] datasets. Implementation details are discussed in IV-A, the overall performance is presented in IV-B. Section IV-C and IV-D report the attribute-based evaluation and quantitative evaluation. Moreover, the ablation study is presented in Section IV-E.

A. IMPLEMENTATION DETAILS

We use HOG and Color names as feature representation for target location and HOG features for scale estimation. The tracking speed is about 18 fps without deep features. Similar to the setting of SRDCF [7], we crop a lager training and detection area centered at the target. The extracted features are weighted by a cosine window. For tracking configurations, the regularization parameter is set to $\lambda = 0.01$, the augmented Lagrangian optimization parameters are set to $\gamma = 16$, the number of scales in SAMF is set to $k = 3$, the scale pool is 0.95, 1.00, 1.05, the number of scales in DSST is set to $s = 33$, the scale factor $\alpha = 1.02$. Since the correlation filter responses are sensitive to scale variation than spatial translation, a coarse-to-fine method is employed to estimate the object scale in the proposed tracking. All the other related parameters are set according to [7].

B. OVERALL PERFORMANCE

The OTB2015 benchmark [29] contains 100 annotated sequences. We evaluate our tracker on the OTB2015 [29] dataset using the one-pass evaluation protocol with distance precision and overlap success rate. The precision is an average of Euclidean distance in pixels between the center points of the tracked and the groundtruth boxes, the center location error with a threshold of 20 pixels. The success rate of a tracker is the proportion of the successful frames with an overlap rate larger than a given threshold of 0.5. The trackers in success plots are ranked based on the area under the curve (AUC).

1) COMPARISON WITH SPATIAL REGULARIZATION BASED METHODS

We evaluate the proposed tracker with three spatial regularization based trackers including SRDCF [7], DeepSRDCF [8], SRDCFdecon [9]. Fig. 4 shows the OPE results on the OTB100 dataset, our tracker outperforms SRDCF by 6.8% in terms of precision and 4.7% in terms of success rate.

2) COMPARISON WITH CNN BASED METHODS

We compare the proposed tracking algorithm with the current CNN based tracking including HCF [21], HDT [24], SiamFC [2]. We note that our handcrafted features based method has shown better performance. It indicated that filter model has equal importance with feature representation. To build a more robust correlation filter model can make great tracking performance improvement.

3) COMPARISON WITH OTHER METHODS

We compare the proposed tracking method with scale estimation based tracking including DSST [5], and SAMF [18], and others including KCF [11] and Staple [1]. These methods are much simpler and faster than ours, but have no competitiveness in tracking performance.

C. ATTRIBUTE-BASED EVALUATION

There are 11 annotated attributes in the OTB100 [29] dataset, including: illumination variation (IV35), out-of-plane rotation (OR59), scale variation (SV61), occlusion (OCC44), deformation (DEF39), motion blur (MB29), fast motion (FM37), in-plane rotation (IR51), out-of-view (OV14),

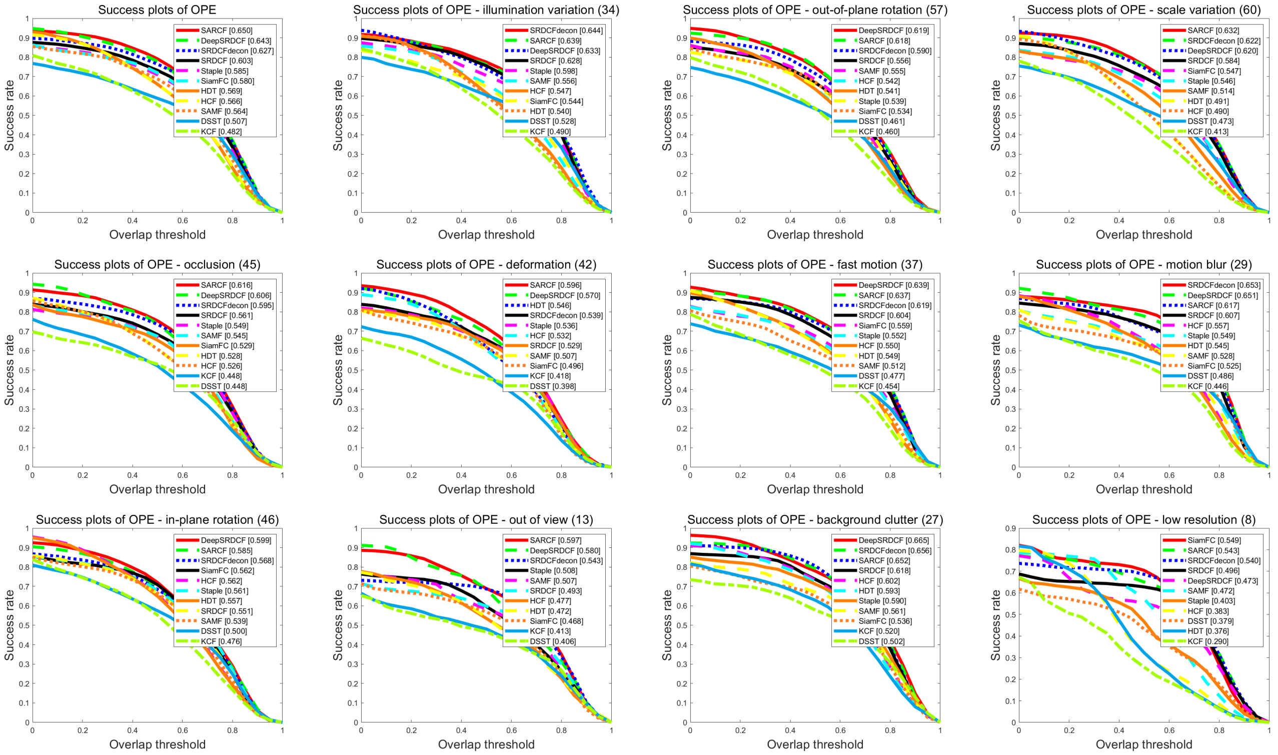


FIGURE 5. Evaluation of different trackers with 11 attributes on the OTB2015 dataset [29], where the legend of overlap success contains area-under-the-curve score for each tracker. For completeness, we also include the overall results. The legend contains the area-under-the-curve score for each tracker. The proposed algorithm performs well against state-of-the-art results.

background clutter (BC31), and low resolution (LR9) (number of videos for each attribute is appended to the end of each abbreviation).

We present the comparison results in terms of different attributes in Fig. 5 and Fig. 6. Red line represents the optimal result and green represents the suboptimal result. It can be seen from Fig. 5 and Fig. 6 that the algorithm in this paper achieves the optimal or suboptimal tracking results on almost all attributes. Especially in dealing with scale variation (SV), the algorithm in this paper achieves 0.632 in success rate and 0.844 in tracking accuracy, which is far superior to similar tracking algorithms based on SAMF and DSST. This further verifies that the the better scale estimation method used in this paper has a good ability to deal with scale variation. Overall, our tracker obtains the best performance in challenges of scale variation, occlusion, deformation and out of view situations. This can be attributed to the better scale estimation method and the more robust filter model. By adaptive spatial regularization, our tracker can handle many challenging situations.

We can also see that the DeepSRDCF performs well in deformation and rotation situations. This can indicate that CNN features can deal with hard appearance changes.

D. QUANTITATIVE EVALUATION

Fig. 7 shows the qualitative comparisons with the performing tracking methods: SRDCF [7], DeepSRDCF [8], SRDCFdecon [9], HCF [21], HDT [24], SiamFC [2], DSST [5], SAMF [18], KCF [11], Staple [1] and MEEM [31] and the proposed method on four challenging image sequences including *BlurOwl, carScale, Freeman4, Human3*. Overall, our tracker is able to locate the target well in these completed scenes. In the *Freeman4* sequence, the target suffers from scale variation and occlusion, our tracker can have shown better performance than SRDCF. In the *human3* sequence, the target is small and suffers from scale variation, occlusion and in-plane rotation. Moreover, this sequence is much longer than the others. Many tracker (SRDCF [7], HCF [21], HDT [24] and MEEM [31]) lose the target after a few frames. From Figure 7, we can see that our tracker can well handle long-term tracking and the scale variation.

E. ABLATION STUDY

To demonstrate the effectiveness of the proposed tracking method based on spatial adaptive regularization, we compare SARCF against its component trackers. Compared with the SRDCF, There are three additional components: (1) the coarse-to-fine scale estimation method, (2) the spatial adaptive regularization method,(3) the temporal regularization method.

We demonstrate the effectiveness of these three components with an ablation study on the OTB2015 data set [29]. In order to evaluate the effectiveness of scale adaptive spatial and temporal regularization methods, we denote SARCF without scale adaptive spatial constraint as SARCF-SA,

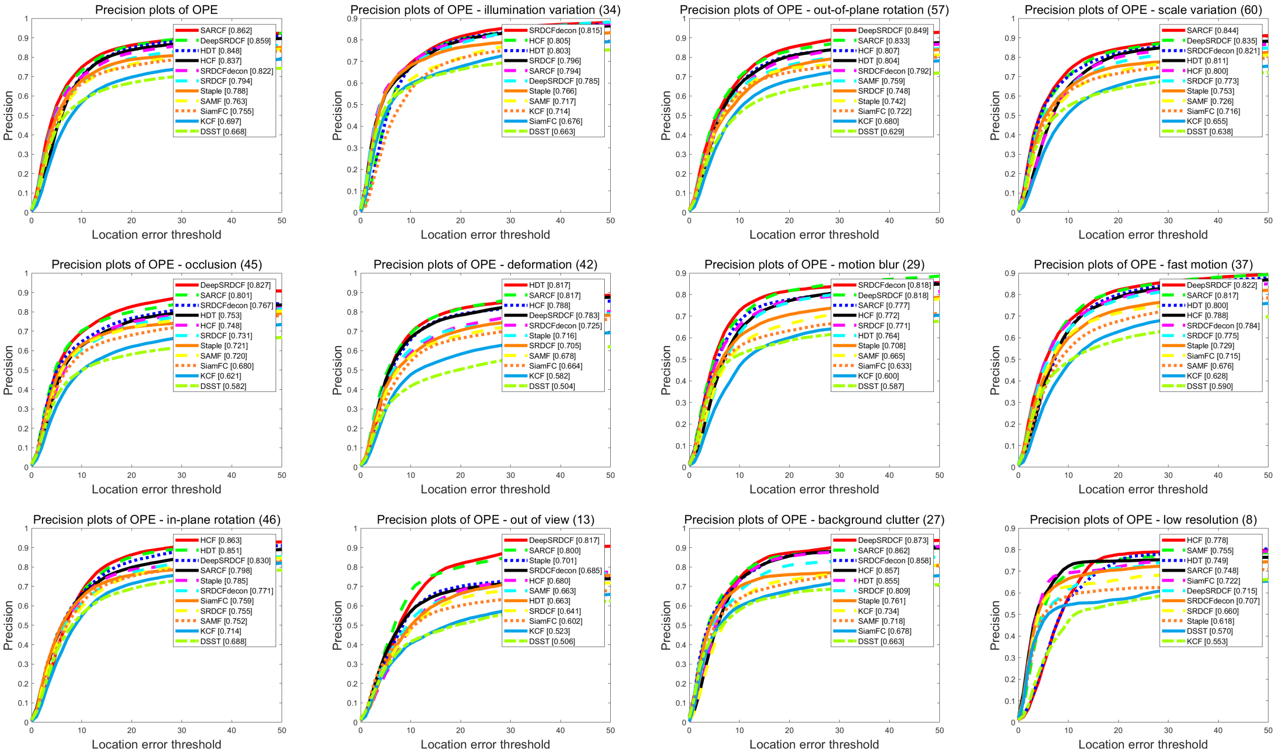


FIGURE 6. Evaluation of different trackers with 11 attributes on the OTB2015 dataset [29], where the legend of distance precision contains threshold scores at 20 pixels for each tracker. For completeness, we also include the overall results. Our proposed algorithm performs favorably against the state-of-the-art trackers.

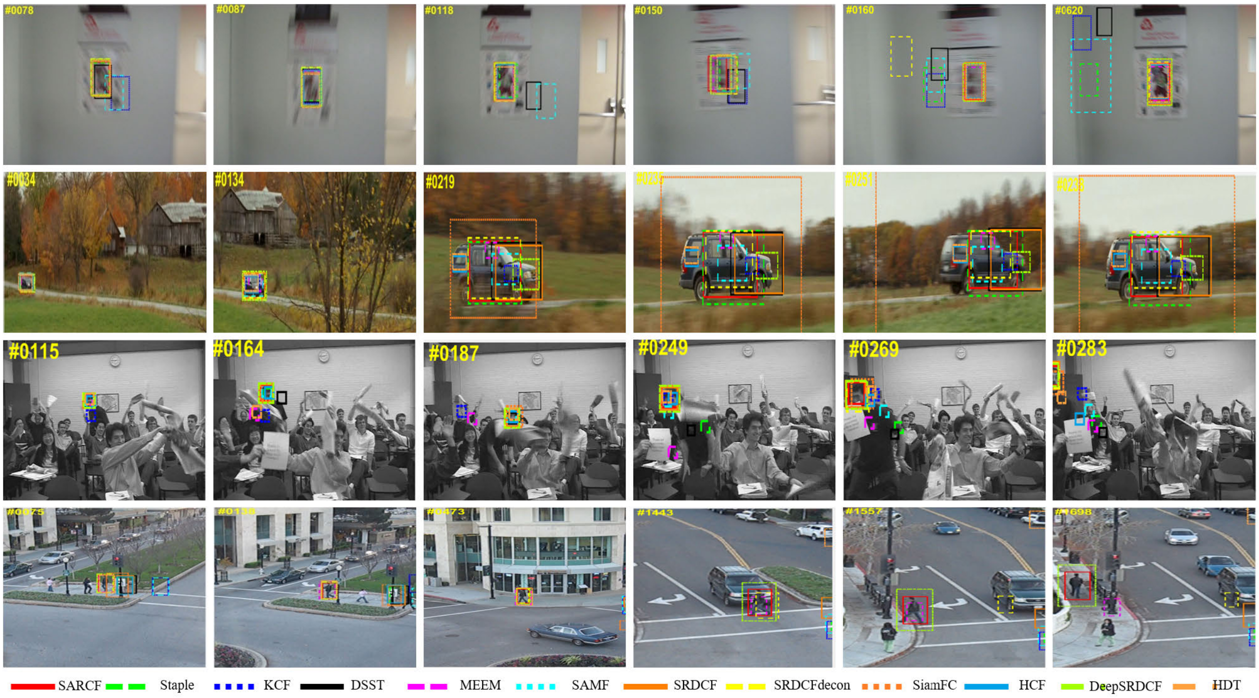


FIGURE 7. Sample tracking results on challenging image sequences (from top to down are *BlurOwl, carScale, Freeman4, Human3*). We show some tracking results of SRDCF [7], DeepSRDCF [8], SRDCFdecon [9], HCF [21], HDT [24], SiamFC [2], DSST [5], SAMF [18], KCF [11], Staple [1] and MEEM [31] methods as well as the proposed algorithm.

without temporal regularization as SARCF-T, without both of these as SARCF-SAT. The comparison results are shown in Table 1, which shows that the success score of the proposed SARCF is decreased by 1.9% without the scale adaptive spatial constraint and decreased by 2.7% without the temporal regularization.

TABLE 1. Ablation study. Component analysis on the OTB2015 dataset [29].

	SARCF	SARCF-T	SARCF-SA	SARCF-SAT
AUC(%)	65.0	62.3	63.1	61.1
DP(%)	86.2	84.5	85.6	81.2

These results demonstrate that both scale adaptive spatial and temporal regularization facilitate the SARCF to perform better.

V. CONCLUSION

In this paper, we propose a correlation filter tracking algorithm based on adaptive spatial regularization. A tight connection is built up to link scale estimation and spatial regularization. We propose a coarse-to-fine scale estimation approach to obtain the spatial penalty matrix, which is used to address the boundary effect problem. In addition, we introduce the temporal term to update the target model, which can well handle occlusion and long-term tracking problem. Experimental evaluations demonstrate the effectiveness and robustness of the proposed method.

REFERENCES

[1] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1401–1409.

[2] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2016, pp. 850–865.

[3] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[4] Z. Chen, Z. Hong, and D. Tao, "An experimental survey on correlation filter-based tracking," 2015, *arXiv:1509.05520*. [Online]. Available: https://arxiv.org/abs/1509.05520

[5] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Nottingham, U.K., 2014.

[6] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Coloring channel representations for visual tracking," in *Proc. Scandin. Conf. Image Anal.* Springer, 2015, pp. 117–129.

[7] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[8] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015.

[9] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1430–1438.

[10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer-Verlag, 2012, pp. 702–715.

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

[12] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, Jun. 2015, pp. 597–606.

[13] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017.

[14] H. K. Galoogahi, T. Sim, and S. Lucey, "Correlation filters with limited boundaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.

[15] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[16] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[17] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking," *Pattern Recogn.*, vol. 76, pp. 323–338, Apr. 2018.

[18] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCV)*, 2014, pp. 254–265.

[19] Y. Li, J. Zhu, S. C. Hoi, W. Song, Z. Wang, and H. Liu, "Robust estimation of similarity transformation for visual object tracking," in *Proc. Conf. Assoc. Adv. Artif. Intell. (AAAI)*, vol. 33, Aug. 2019, pp. 8666–8673.

[20] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 369–386.

[21] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[22] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.

[23] L. Pu, X. Feng, and Z. Hou, "Learning temporal regularized correlation filter tracker with spatial reliable constraint," *IEEE Access*, vol. 7, pp. 81441–81450, 2019.

[24] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4303–4311.

[25] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.

[26] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3119–3127.

[27] L. Wang, W. Ouyang, X. Wang, and H. Lu, "STCT: Sequentially training convolutional networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1373–1381.

[28] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 21–26.

[29] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[30] H. Zhang, Z. Gao, X. Ma, J. Zhang, and J. Zhang, "Hybridizing teaching-learning-based optimization with adaptive grasshopper optimization algorithm for abrupt motion tracking," *IEEE Access*, vol. 7, pp. 168575–168592, 2019.

[31] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 188–203.

[32] T. Zhou, H. Bhaskar, F. Liu, and J. Yang, "Graph regularized and locality-constrained coding for robust visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2153–2164, Oct. 2017.

[33] T. Zhou, F. Liu, H. Bhaskar, and J. Yang, "Robust visual tracking via online discriminative and low-rank dictionary learning," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2643–2655, Sep. 2018.

LEI PU was born in Sichuan, China, in 1991. He received the M.S. degree from Air Force Engineering University (AFEU), Xi'an, China, in 2017. He is currently pursuing the Ph.D. degree. His main research interests include visual tracking, pattern recognition, and computer vision.

XINXI FENG received the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1987 and 1991, respectively. He is currently a Full Professor with the Information and Navigation College, AFEU. His research interests include information fusion and image processing.

ZHIQIANG HOU received the Ph.D. degree from Xi'an Jiaotong University, in 2005. He was a Visiting Scholar with the University of Bristol, U.K., in 2009. He is currently a Full Professor with Air Force Engineering University and Xi'an University of Posts and Telecommunications. His research interests include pattern recognition, computer vision, image processing, and information fusion.