Received: 3 November 2020 | Revised: 2 February 2021 | Accepted: 21 February 2021

DOI: 10.1049/ipr2.12150

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Learning adaptive spatial–temporal regularized correlation filters for visual tracking

Jianwei Zhao[1,2] | Yangxiao Li[1] | Zhenghua Zhou[1]

[1] Department of Mathematics and Information Sciences, China Jiliang University, Hangzhou 310018, PR China

[2] Key Laboratory of Intelligent Manufacturing Quality Big Data Tracing and Analysis of Zhejiang Province, China Jiliang University, Hangzhou 310018, PR China

Correspondence
Zhenghua Zhou, Department of Mathematics and Information Sciences, China Jiliang University, Hangzhou 310018, PR China.
Email: zzhzjw2003@163.com

Funding information: Zhejiang Provincial Nature Science Foundation, Grant/Award Number: LY18F020018, LSY19F020001; National Nature Science Foundation of China, Grant/Award Number: 61571410

Funding information
Zhejiang Provincial Nature Science Foundation, Grant/Award Numbers: LSY19F020001, LY18F020018; National Nature Science Foundation of China, Grant/Award Number: 61571410

Abstract
Recently, there have been many visual tracking methods based on correlation filters. These methods mainly enhance the tracking performances by considering the information of background, space, or time in the appearance model. This paper proposes an effective tracking method, named adaptive spatial–temporal regularized correlation filter (ASTRCF) tracker, based on the popular adaptive spatially regularized correlation filter (ASRCF) tracker. That is, the continuity of object's motion in the process of tracking is considered by introducing a temporal-regularized term in the appearance model of ASRCF tracker. Furthermore, its solution is inferred by applying the alternating direction method of multipliers. The proposed appearance model contains a background-awareness term, a spatially regularized term, an adaptive-weight term, and a temporal-regularized term. Therefore, it can not only keep the good performances of ASRCF tracker, such as learning the background information and the spatial information adaptively to enhance the discriminating ability, but also take advantage of the relation of correlation filters in the last frame and the current frame for addressing the complex cases, such as occlusion, and fast motion. Extensive experimental results on various challenging databases show that the proposed ASTRCF tracker achieves better tracking performances than some state-of-the-art trackers.

1 | INTRODUCTION

Visual tracking is a hot topic in computer vision because of its popular applications in real life. Its goal is to design an effective tracker for determining the position of the object automatically only with the given initial position in the first frame. As the tracker is usually trained by some features extracted from the last frame, some appearance variations of the object, such as occlusion, fast motion, motion blur, and deformation, will affect the tracking results seriously. Therefore, it is still a challenging task to design a robust and efficient tracker.

Recently, correlation filter (CF)-based trackers have caught people's attentions due to their excellent tracking performances. This type of trackers are designed under the principle that similar signals should have similar response maps. In 2010, Bolme et al. firstly introduced correlation filter into visual tracking, and proposed a minimum output sum of squared error (MOSSE) tracker [1]. It mainly learns the correlation filter with a set of extracted features using fast Fourier transform (FFT), then predicts the position of the object in the next frame by the learned correlation filter. Since then, numerous improved trackers have been proposed based on different prior constraints. Henriques et al. exploited the circulate structure of the shifted image patches in kernel space and proposed a kernelized correlation filter (KCF) tracker [2]. Danelljan et al. proposed a discriminative scale space tracker (DSST) for visual tracking by introducing a scale regularization term in the appearance model of the MOSSE tracker. Its appearance model for learning the desired correlation filter h can be described as follows:

$$\arg\min_{h} \frac{1}{2}\left\| y - \sum_{k=1}^{K} x^k * h^k \right\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K} \left\| h^k \right\|_2^2, \tag{1}$$

where x^k and h^k denote the k-th channel of the vectorized feature image $x = [x^1, x^2, ..., x^K]$ and correlation filter $h = [h^1, h^2, \ldots, h^K]$, respectively; $k = 1, 2, \ldots, K$, y is the desired Gaussian response; $*$ denotes the convolution operation; and λ is a regularization parameter.

In order to relieve the unwanted boundary effect on the tracking results, Danelljan et al. proposed a spatial regularized correlation filter (SRDCF) tracker [3] by fusing the spatial information in the scale regularization term in the appearance model of the DSST tracker. Its appearance model for learning the correlation filter can be described as follows:

$$\arg\min_{h} \frac{1}{2}\left\| y - \sum_{k=1}^{K} x^k * h^k \right\|_2^2 + \frac{\lambda}{2}\sum_{k=1}^{K} \left\| w \odot h^k \right\|_2^2, \tag{2}$$

where w is a negative Gaussian-shaped spatial weight matrix to make the learned filter have a high response around the center of the tracked object, and $\odot$ is the dot multiplication.

Considering the robustness requirement of a tracker for the complicated scenarios, Bertinetto et al. proposed a Staple tracker [4] by combining the complementary cues in the ridge regression framework for speeding up the tracking. Mueller et al. presented a context aware correlation filter (CACF) tracker [5] that considered the surrounding context information in the appearance model for learning the correlation filter. In the meanwhile, Galoogahi et al. proposed a background-aware correlation filter (BACF) tracker that trained the correlation filter from negative training examples to remit the boundary effect [6].

Recently, with the powerful ability of extracting high-level features, deep learning is also introduced in the area of visual tracking. Some people design effective deep networks to learn the relation between the frame and the object. For example, Zhang et al. proposed an effective convolutional network without training (CNT) tracker [7] for determining the position of target. Although deep network based trackers can improve the tracking results, they cannot use very deep networks for tracking the object because of the real-time requirement in the process of tracking. Therefore, some people combine the advantages of deep learning and correlation filter. They use the pre-trained deep networks to extract the high-level features as the training samples for learning the correlation filter. For example, Qi et al. proposed a hedged deep tracking (HDT) tracker [8] that took full advantage of features from different convolutional layers and used the adaptive Hedge method to hedge several DSST trackers into a single stronger one.

With the combination of deep features and hand-craft features, many improved CF-based trackers have been proposed. Based on SRDCF tracker, Li et al. proposed a spatially temporal regularized correlation filter (STRCF) tracker [9] by introducing a temporal regularization term in the appearance model of the SRDCF tracker. Its appearance model for learning the correlation filter can be described as follows:

$$\arg\min_{h} \frac{1}{2}\left\| y - \sum_{k=1}^{K} x^k * h^k \right\|_2^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K} \left\| w \odot h^k \right\|_2^2 + \frac{\lambda_2}{2}\left\| h - h_{t-1} \right\|_2^2, \tag{3}$$

where $x = [x^1, x^2, ..., x^K]$ is the combination of deep features and hand-craft features, h_{t-1} denotes the correlation filter learned from the $(t-1)$-th frame, and $\| h - h_{t-1} \|_2^2$ denotes the temporal regularization term.

Based on BACF tracker, Yuan et al. proposed a temporal regularization background-aware correlation filter (TRBACF) tracker [10] by introducing a temporal regularization term in the appearance model of the BACF tracker to efficiently adapt the change of the tracking scenes. Subsequently, considering that an adaptive weight matrix is more appropriate for a tracker, Dai et al. proposed an adaptive spatially regularized correlation filter (ASRCF) tracker [11] by introducing the adaptive spatial regularization terms in the appearance model of the BACF tracker. Its appearance model can be described as follows:

$$\arg\min_{h,w} \frac{1}{2}\left\| y - \sum_{k=1}^{K} x^k * (P^\top h^k) \right\|_2^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K} \left\| w \odot h^k \right\|_2^2 + \frac{\lambda_2}{2}\left\| w - w^r \right\|_2^2, \tag{4}$$

where P is a diagonal binary matrix to make the correlation filter directly apply on the true foreground and background samples, and w^r is a reference weight.

From above analysis, we found that the STRCF tracker considered the spatial information and the temporal information, not considering the background information and the variation of the spatial weight matrix w for the whole tracking process. While the TRBACF tracker considered the background information and the temporal information, not considering the spatial information and the variation of the spatial weight matrix, and ASRCF tracker considered the background information, spatial information, and the variation of the spatial weight matrix, not considering the temporal information.

This paper proposes a novel CF-based tracker, named an adaptive spatial–temporal regularized correlation filter (ASTRCF) tracker, by introducing the temporal regularization term in the appearance model of the ASRCF tracker. Hence, the proposed ASTRCF tracker not only considers the background information, spatial information, and the variation of the spatial weight matrix, but also considers the temporal information. By means of above prior constraints, our proposed ASTRCF tracker shows the good robustness for the complicated scenes. The main contributions of this paper are as follows:

- We propose a novel CF-based tracker, named an adaptive spatial–temporal regularized correlation filter (ASTRCF)

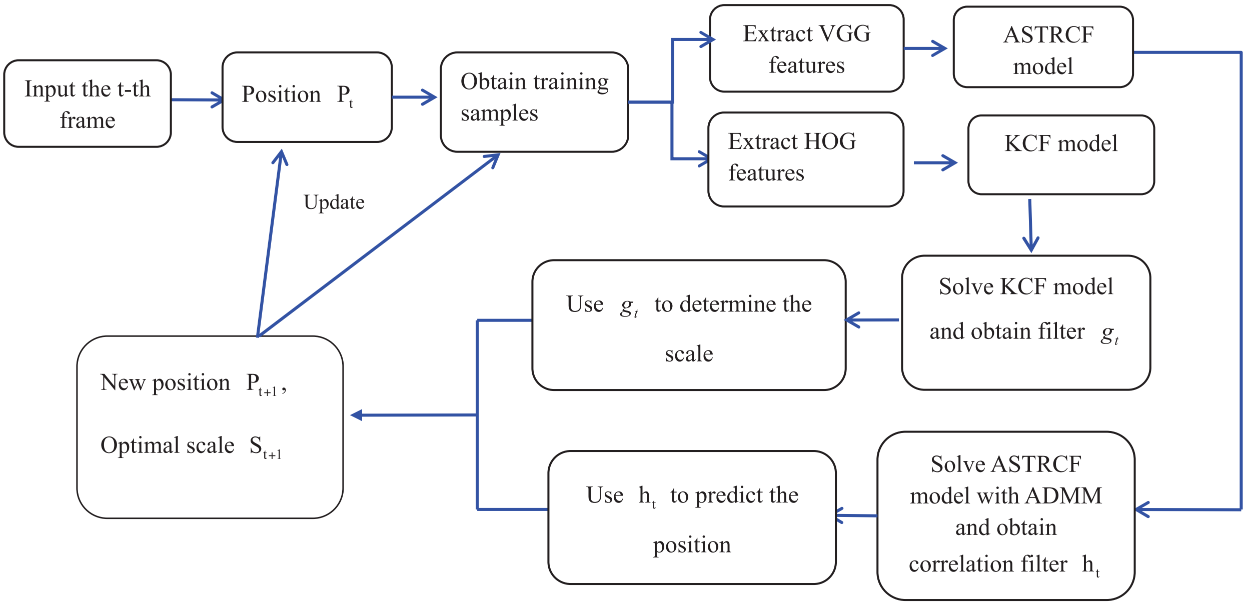


FIGURE 1 A flowchart of our proposed ASTRCF tracker

tracker, by introducing a temporal regularization term in the appearance model of the ASRCF tracker.

- For the proposed new appearance model, we apply the alternating direction method of multipliers (ADMM) to deduce the iterative solutions.
- The proposed ASTRCF tracker can not only take advantage of the background information, spatial information, and the variation of the spatial weight matrix, but also explore the relation between the last frame and the current frame. Therefore, our tracker has good robustness for complicated scenes.

The rest of the paper is organized as follows: Section 2 gives the details of our proposed ASTRCF tracker. Section 3 compares the results of our ASTRCF tracker with some other state-of-the-art trackers. Section 4 concludes the paper.

2 | PROPOSED ASTRCF TRACKER

In order to enhance the robustness of ASRCF tracker for the complicated scenes, we propose a novel CF-based tracker, ASTRCF tracker, by taking use of the relation of the last correlation filter and the current correlation filter. Firstly, our ASTRCF tracker designs a new appearance model by introducing a temporal regularization term in the appearance model of ASRCF tracker. Secondly, we apply ADMM [12] to deduce the iteration solution of above appearance model. In the meanwhile, we still use KCF model for the multiple scale searching. The flowchart of our proposed ASTRCF tracker is shown in Figure 1.

2.1 | Proposed ASTRCF appearance model for learning CF

Let I_t be the t-th frame in a video sequence with a tracked target (position P_t and scale S_t), then the task of our tracker is to determine the target (position P_{t+1} and scale S_{t+1}) in the subsequent frame I_{t+1} effectively. Here, the position P_t is the coordinate of the target's center, and the scale S_t is the target's bonding box with a height and a width. Since our proposed ASTRCF tracker is based on the theory of correlation filter, the first thing is to collect some training samples for learning the correlation filter. In order to avoid transmitting the sampling errors because of the tracking drift to the subsequent tracking process, and learn some background information around the target to avoid the interferences from some similar patches, we apply the circular shift operation on the whole t-th frame for collecting the training samples.

For each training sample $(x_i, y_i) \in \mathbb{R}^{M\times M} \times \mathbb{R}^{M\times M}$, we extract the VGG feature $[x_i^1, x_i^2, \dots, x_i^K] \in \mathbb{R}^{M_0\times M_0\times K}$ with K channels for x_i, $i = 1, 2, \dots, N$. In order to facilitate the calculation, we vectorized x_i^k and y_i, and still used the original symbols to denote them. Let x denote the matrix whose (i, j)-th block is the vector x_i^j, and use x^k to denote the k-th column of the matrix x, $k = 1, 2, \dots, K$. Denote $y = (y_1^\top, y_2^\top, \dots, y_N^\top)^\top$, where $\top$ is the transposition operation for a matrix.

In order to reflect the appearance variations of an object for the whole tracking process and use the relation between the last frame and current frame, we propose a novel adaptive spatial–temporal regularized correlation filter appearance model for learning the effective correlation filter. Its optimal model can be described as follows:

$$\begin{aligned}\arg\min_{h,w} E(h, w) \; := \arg\min_{h,w} \frac{1}{2}\left\| y - \sum_{k=1}^{K} x^k * (P^\top h^k) \right\|_2^2 \\ + \frac{\lambda_1}{2} \sum_{k=1}^{K} \| w \odot h^k \|_2^2 \\ + \frac{\lambda_2}{2} \| w - w^r \|_2^2 + \frac{\lambda_3}{2} \| h - h_{t-1} \|_2^2,\end{aligned} \tag{5}$$

where $P \in \mathbb{R}^{M_0^2N \times M_0^2N}$ is a diagonal binary matrix to make the correlation filter directly apply on the true foreground and background samples, w^r is a reference weight to make the learned filter have a high response around the center of the tracked object, w is the spatial weight matrix to be learned, h_{t-1} is the correlation filter learned for the $(t-1)$-th frame, and λ_i is a regularization parameter, $i = 1, 2, 3$.

In above model (5), the first term is a ridge regression term that uses the background information. The second term emphasizes the foreground information. The third term attempts to make the adaptive spatial weight w be similar to a reference weight w^r. This constraint introduces a prior information on w and avoids the model degradation. The last term is a temporal regularization term that makes the correlation filter h be similar to the learned correlation filter h_{t-1} because of the continuity of object's movement. Therefore, our proposed ASTRCF model can not only use the background and foreground of the object's variation effectively via the adaptive spatial weight, but also reflect the relation of the correlation filters between the last frame and the subsequent frame, which can relieve the drifting problem caused by the quick movement of the object.

Generally, CF-based trackers learn correlation filter and predict the position of the object on the frequency domain. Hence, we also solve the optimization problem (5) on the frequency domain. Firstly, we convert (5) into the following optimal form with equality constraints:

$$\begin{aligned} E(h, w, \hat{G}) = {} & \frac{1}{2}\left\| \hat{y} - \sum_{k=1}^{K} \hat{x}^k \odot \hat{g}^k \right\|_2^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K} \| w \odot h^k \|_2^2 \\ & + \frac{\lambda_2}{2}\sum_{k=1}^{K} \| w - w^r \|_2^2 \\ & + \frac{\lambda_3}{2}\| h - h_{t-1} \|_2^2 \ s.t. \ \hat{g}^k \\ = {} & \sqrt{M_0^2 N} F(P^\top h^k), \quad k = 1, 2, \dots, K, \end{aligned} \tag{6}$$

where $\hat{G} = [\hat{g}^1, \hat{g}^2, \dots, \hat{g}^K]$ is an auxiliary variable matrix, and $F \in \mathbb{R}^{M_0^2N \times M_0^2N}$ is the discrete Fourier matrix corresponding to the Fourier transform on $\mathbb{R}^{M_0^2N}$.

For the bi-convex model (6), we apply ADMM [12] to obtain its local optimal solution. The augmented Lagrangian function for the problem (6) is

$$\begin{aligned} L(h, w, \hat{G}, \hat{V}) = {} & E(h, w, \hat{G}) + \sum_{k=1}^{K} \hat{v}_k^\top \left(\hat{g}^k - \sqrt{M_0^2 N} F(P^\top h^k) \right) \\ & + \frac{\mu}{2}\sum_{k=1}^{K} \left\| \hat{g}^k - \sqrt{M_0^2 N} F(P^\top h^k) \right\|_2^2, \end{aligned} \tag{7}$$

where $V = [v_1, v_2, \dots, v_K] \in \mathbb{R}^{M_0^2N \times K}$ is the Lagrange multiplier, and $\hat{V} = [\hat{v}_1, \hat{v}_2, \dots, \hat{v}_K] \in \mathbb{R}^{M_0^2N \times K}$ is its Fourier transform. Let $u_k = \frac{1}{\mu} v_k, k = 1, 2, \dots, K$, then the optimization of (7) is equivalent to the following form:

$$\begin{aligned} L(h, w, \hat{G}, \hat{U}) = {} & \frac{1}{2}\left\| \hat{y} - \sum_{k=1}^{K} \hat{x}^k \odot \hat{g}^k \right\|_2^2 \\ & + \frac{\lambda_1}{2}\sum_{k=1}^{K} \| w \odot h^k \|_2^2 + \frac{\lambda_2}{2}\| w - w^r \|_2^2 \\ & + \frac{\lambda_3}{2}\| h - h_{t-1} \|_2^2 + \frac{\mu}{2}\sum_{k=1}^{K} \left\| \hat{g}^k \right. \\ & \left. - \sqrt{M_0^2 N} F(P^\top h^k) + \hat{u}_k \right\|_2^2, \end{aligned} \tag{8}$$

where $\hat{U} = [\hat{u}_1, \hat{u}_2, \dots, \hat{u}_K] \in \mathbb{R}^{M_0^2N \times K}$.

Now we apply ADMM to solve above optimization problem (8). Using ADMM, above optimization problem (8) will be converted to following three subproblems:

$$\begin{cases} h^k(j+1) = \arg\min\limits_{h^k} \frac{\lambda_1}{2}\| w(j) \odot h^k \|_2^2 + \frac{\lambda_3}{2}\left\| h^k - h_{t-1}^k \right\|_2^2 \\ \qquad + \frac{\mu}{2}\left\| \hat{g}^k(j) - \sqrt{M_0^2 N} F(P^\top h^k) + \hat{u}_k \right\|_2^2; \\ \hat{G}(j+1) = \arg\min\limits_{\hat{G}} \frac{1}{2}\left\| \hat{y} - \sum_{k=1}^{K} \hat{x}^k \odot \hat{g}^k \right\|_2^2 \\ \qquad + \frac{\mu}{2}\sum_{k=1}^{K} \left\| \hat{g}^k - \sqrt{M_0^2 N} F\left(P^\top h^k(j+1)\right) + \hat{u}_k \right\|_2^2; \\ w(j+1) = \arg\min\limits_{w} \frac{\lambda_1}{2}\sum_{k=1}^{K} \left\| w \odot h^k(j+1) \right\|^2 + \frac{\lambda_2}{2}\| w - w^r \|_2^2. \end{cases} \tag{9}$$

Now we give the concrete steps for solving above subproblems.

Subproblem h. For $k = 1, 2, \dots, K$, when $\hat{G}$, w, and $\hat{U}$ are given, the optimal correlation filter for the k-th channel h^k can be computed as follows:

$$\begin{aligned} h^k(j+1) = {} & \arg\min_{h^k} \frac{\lambda_1}{2}\| w(j) \odot h^k \|_2^2 + \frac{\lambda_3}{2}\left\| h^k - h_{t-1}^k \right\|_2^2 \\ & + \frac{\mu}{2}\left\| \hat{g}^k(j) - \sqrt{M_0^2 N} F(P^\top h^k) + \hat{u}_k \right\|_2^2 \\ = {} & \left[\lambda_1 W(j)^\top W(j) + \mu M_0^2 N P^\top P + \lambda_3 I \right]^{-1} \\ & \times \left[\mu M_0^2 N P\left(u_k + g^k(j)\right) + \lambda_3 h_{t-1}^k \right], \end{aligned} \tag{10}$$

where $W(j) = diag(w(j)) \in \mathbb{R}^{M_0^2N \times M_0^2N}$ denotes the diagonal matrix, and I is the identity matrix with M_0^2N order.

Subproblem $\hat{G}$: When $h^k (k = 1, 2, \ldots, K)$, w, and $\hat{U}$ are given, the optimal $\hat{G}$ in formula (9) is

$$\hat{G}(j+1) = \arg\min_{\hat{G}} \frac{1}{2} \left\| \hat{y} - \sum_{k=1}^{K} \hat{x}^k \odot \hat{g}^k \right\|_2^2 + \frac{\mu}{2} \sum_{k=1}^{K} \left\| \hat{g}^k - \sqrt{M_0^2NF}\left(P^\top h^k(j+1)\right) + \hat{u}_k \right\|_2^2. \tag{11}$$

Due to high computational complexity, it is difficult to optimize above formula (11) directly. Hence, we solve this problem by means of pixel, and describe above optimization (11) as follows:

$$\begin{aligned} V_i(\hat{G}(j+1)) = \arg\min_{V_i(\hat{G})} & \frac{1}{2} \left\| V_i(\hat{y}) - V_i(\hat{x})^\top V_i(\hat{G}) \right\|_2^2 \\ & + \frac{\mu}{2} \sum_{k=1}^{K} \left\| V_i(\hat{G}) + V_i(\hat{M}) \right\|_2^2 \\ s.t.\ \ & V_i(\hat{M}) = V_i(\hat{U}) \\ & - V_i\left(\sqrt{M_0^2NF}\left(P^\top h^k(j+1)\right) \right), \end{aligned} \tag{12}$$

where $V_i(\hat{G}) \in \mathbb{R}^K$ denotes the vector composed of the i-th row of the matrix $\hat{G}$. Then the solution of above problem (12) is

$$\begin{aligned} & V_i(\hat{G}(j+1)) \\ & = \frac{1}{\mu M_0^2 N} \left(I - \frac{V_i(\hat{x}) V_i(\hat{x})^\top}{\mu M_0^2 N + V_i(\hat{x}) V_i(\hat{x})^\top} \right) \\ & \quad \times \left(\hat{y} V_i(\hat{x}) + \mu V_i\left(\sqrt{M_0^2NF}\left(P^\top h^k(j+1)\right) \right) - \mu V_i(\hat{U}) \right). \end{aligned} \tag{13}$$

Subproblem w: If h is fixed, the solution of w can be computed by

$$\begin{aligned} w(j+1) &= \arg\min_{w} \frac{\lambda_1}{2} \sum_{k=1}^{K} \left\| w \odot h^k(j+1) \right\|^2 + \frac{\lambda_2}{2} \left\| w - w^r \right\|_2^2 \\ &= \left(\lambda_1 \sum_{k=1}^{K} \hat{h}^k(j+1) \odot \hat{h}^k(j+1) + \lambda_2 I \right)^{-1} \lambda_2 w^r \\ &= \frac{\lambda_2 w^r}{\lambda_1 \sum_{k=1}^{K} \hat{h}^k(j+1) \odot \hat{h}^k(j+1) + \lambda_2 I}. \end{aligned} \tag{14}$$

In this paper, the initial reference weight w^r can be taken as the negative Gaussian-shaped spatial weight matrix, and is updated by $w(j)$.

For the Lagrangian multipliers $\hat{U}$, we use the following formula to update

$$\hat{U}(j+1) = \hat{U}(j) + \hat{G}(j+1) - h(j+1), \tag{15}$$

where $\hat{U}(j+1)$ is the Fourier transform of $U(j+1)$.

Up to now, the optimization problem can be computed by above four steps iteratively. Now the position of the tracked object can be obtained in the Fourier domain by

$$\hat{r}_1 = \sum_{k=1}^{K} \hat{x}^k \odot \hat{g}^k, \tag{16}$$

where r_1 is the response map and $\hat{r}_1$ is its Fourier transform. With the obtained response map, the location of the object can be determined by the maximum response value.

2.2 | KCF model for scale

In this subsection, we explain KCF model for learning the correlation filter g_t briefly for the scale estimation. For each training sample $(x_i, y_i) \in \mathbb{R}^{M \times M} \times \mathbb{R}^{M \times M}$, we extract the HOG feature $[z_i^1, z_i^2, \ldots, z_i^L] \in \mathbb{R}^{M_1 \times M_1 \times L}$ with L channels for $x_i, i = 1, 2, \ldots, N$. Let z be the matrix whose the (i, j)-th block is the vector z_i^j, and use z^ℓ to denote the ℓ-th column of the matrix z, $\ell = 1, 2, \ldots, L$. Put $y = (y_1^\top, y_2^\top, \ldots, y_N^\top)^\top$.

During the tracking process, we apply KCF model as follows to learn the scale correlation filter g_t:

$$g_t = \arg\min_{g} \frac{1}{2} \left\| y - \sum_{\ell=1}^{L} z^\ell * g^\ell \right\|_2^2 + \frac{\lambda}{2} \sum_{\ell=1}^{L} \left\| g^\ell \right\|_2^2. \tag{17}$$

For above optimization (17), we give a closed form solution in the primal domain by

$$g_t = (z^\top z + \lambda I)^{-1} z^\top y. \tag{18}$$

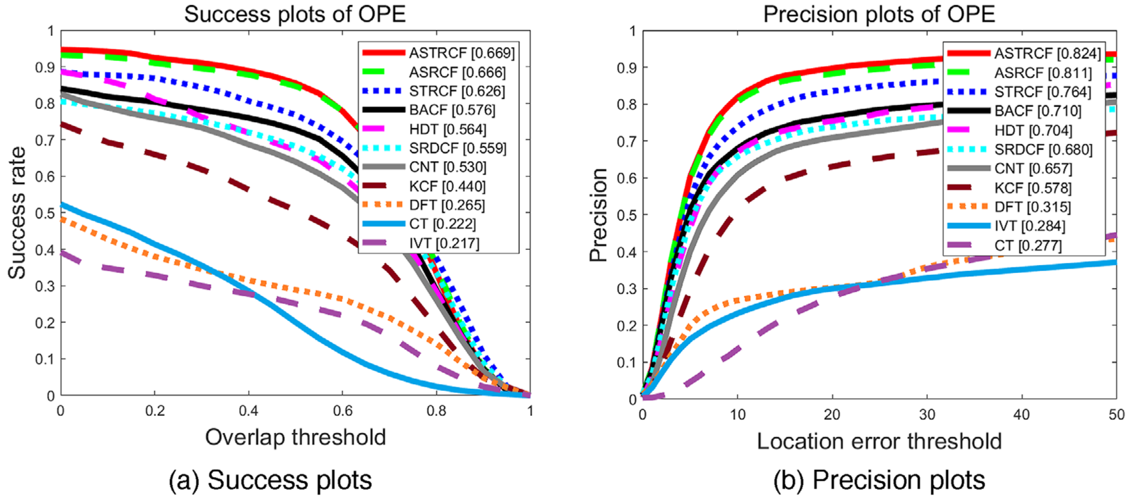
During the tracking process, we apply this correlation filter g_t on five scale search regions and obtain their corresponding response maps. Then, the best scale can be determined according to the maximum score of five response maps.

3 | EXPERIMENTAL RESULTS

In this section, we evaluate our proposed ASTRCF tracker by carrying some quantitative comparison and qualitative comparison with some other 10 popular trackers on OTB50 [13], OTB100 [14], DTB70 [15] and TC128 [16] databases. Here, 10

TABLE 1 Comparison of success rate and precision rate of our proposed ASTRCF tracker with 10 trackers on OTB50, OTB100, DTB70, and TC128 databases

	IVT	CT	DFT	KCF	CNT	SRDCF	HDT	BACF	STRCF	ASRCF	Our
OTB50.AUC	0.217	0.222	0.265	0.440	0.530	0.559	0.564	0.576	0.626	0.666	0.669
OTB50.Pre	0.284	0.277	0.315	0.578	0.657	0.680	0.704	0.710	0.764	0.811	0.824
OTB100.AUC	0.315	0.272	0.327	0.479	0.563	0.592	0.577	0.613	0.654	0.688	0.688
OTB100.Pre	0.409	0.347	0.403	0.635	0.701	0.715	0.734	0.753	0.793	0.831	0.837
DTB70.AUC	0.167	0.150	0.230	0.280	0.334	0.385	0.370	0.414	0.437	0.469	0.472
DTB70.Pre	0.218	0.235	0.331	0.428	0.469	0.517	0.519	0.562	0.606	0.646	0.657
TC128.AUC	0.116	0.133	0.202	0.274	0.301	0.359	0.339	0.399	0.426	0.448	0.461
TC128.Pre	0.148	0.211	0.281	0.419	0.409	0.475	0.478	0.540	0.581	0.603	0.624



(a) Success plots

(b) Precision plots

FIGURE 2 Comparison of success plot and precision plot of our proposed ASTRCF tracker with some popular trackers on **OTB50** database

trackers are KCF [2], SRDCF [3], BACF [6], CNT [7], HDT [8], STRCF [9], ASRCF [11], CT [17], IVT [18], and DFT [19].

Databases: This paper uses three popular video databases: OTB50, OTB100, DTB70 and TC128 in computer vision for experimental comparison. Here, OTB50 and OTB100 databases contain 100 challenging video sequences, DTB70 database consists of 70 challenging color video sequences, and TC128 database contains 128 challenging video sequences. All these video sequences contain 11 different attributes: low resolution (LR), in-plane rotation (IPR), out-of-plane rotation (OPR), scale variation (SV), occlusion (OCC), deformation (DEF), background clutter (BC), illumination variation (IV), motion blur (MB), fast motion (FM), and out of view (OV). The one pass evaluation (OPE) is employed to evaluate different trackers based on two criteria: overlap success and distance precision.

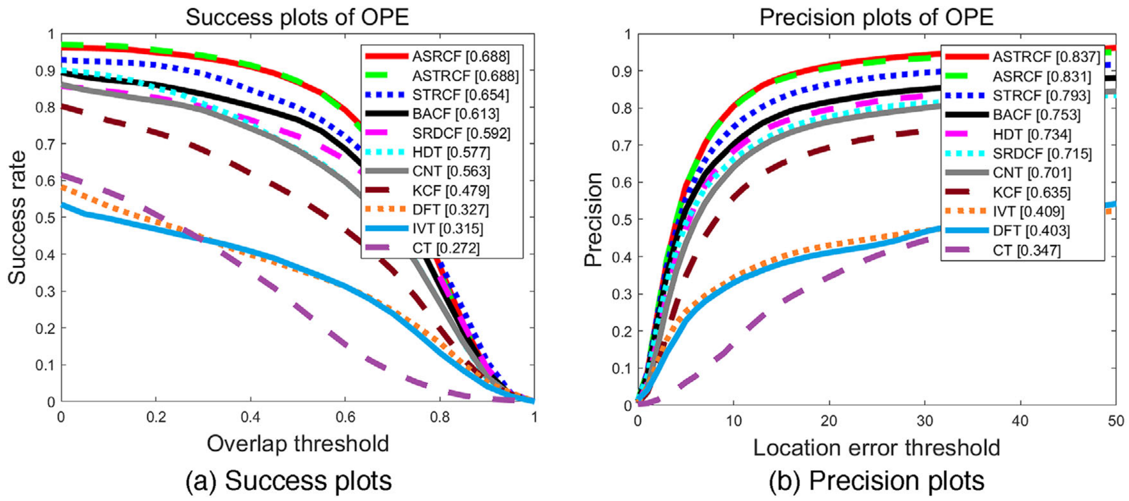
Parameters setting: In the following experiments, the parameters in our proposed ASTRCF tracker are set as follows: the number of scale is 5, the scale-step is 1.01, the number of iterations for ADMM is 2, $\beta = 10$, and the learning rate $\theta = 0.02$. When $\lambda_1 = 0.5$, $\lambda_2 = 1.5$, and $\lambda_3 = 0.01$, our tracker can achieve the best performance on OTB50, OTB100, DTB70, and TC128 databases.

All the experiments are carried out in Matlab2015b on a computer with Intel Xeon CPU E5-1620 v3 @ 3.50GHz.

3.1 | Overall performances

Table 1 shows the success rate and precision rate of our proposed ASTRCF tracker with 10 state-of-the-art trackers on OTB50 database, OTB100 database, DTB70 database, and TC128 database, and Figures 2, 3, 4, and 5 are the corresponding success plots and precision plots of the trackers in Table 1.

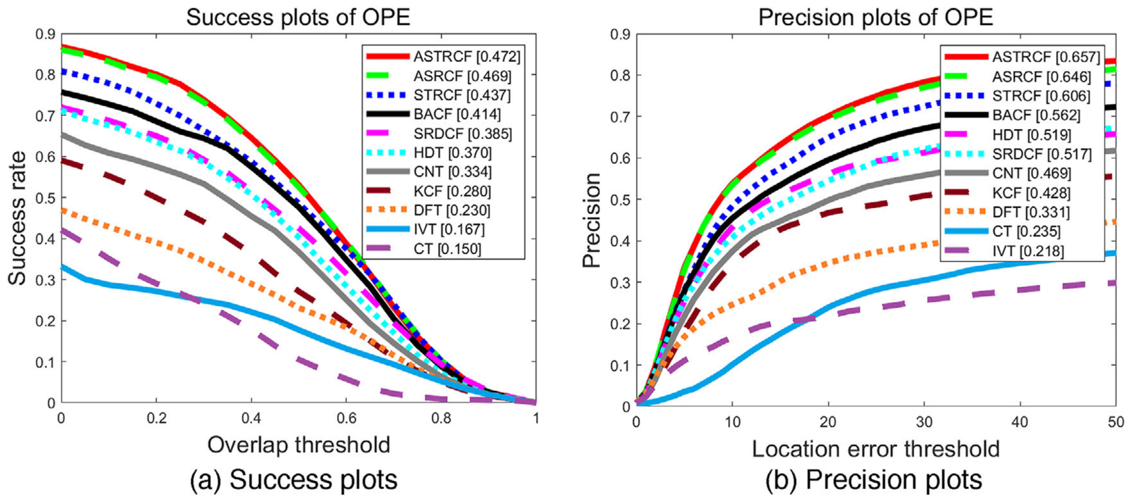
As observed from Table 1 and Figures 2–5 our proposed ASTRCF tracker achieves the best performances on success rate and precision rate among 11 popular trackers on OTB50, OTB100, DTB70, and TC128 databases, respectively. For example, on OTB50 database, the success rate and precision rate of our proposed ASTRCF tracker can reach 0.669 and 0.824, respectively. They are higher than those of the second top tracker ASRCF with 0.006 and 0.015, and the third top tracker STRCF with 0.052 and 0.063, respectively. On TC128 database, the success rate and precision rate of our proposed ASTRCF tracker can reach 0.461 and 0.624, respectively. They are higher than those of the second top tracker ASRCF with 0.013 and



(a) Success plots

(b) Precision plots

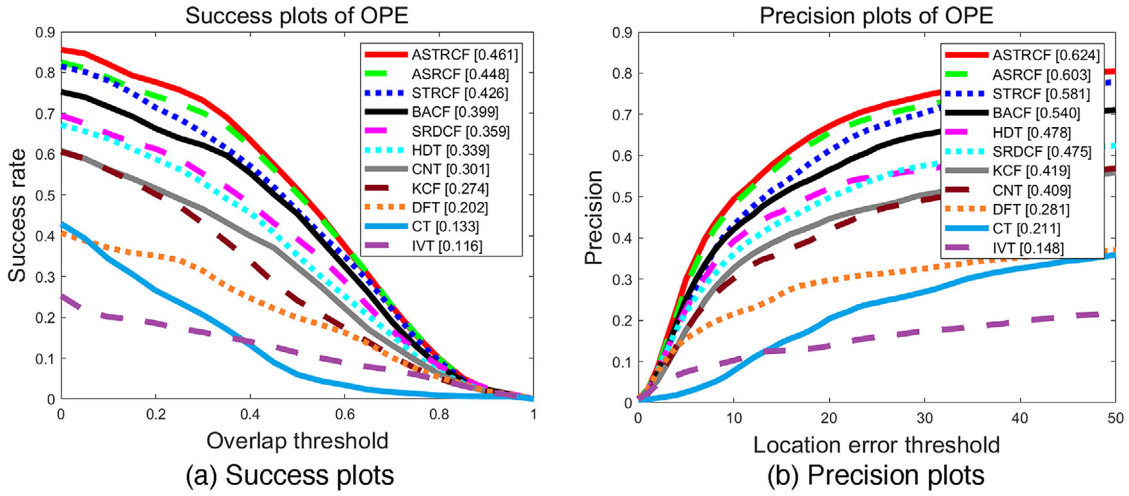
FIGURE 3 Comparison of success plot and precision plot of our proposed ASTRCF tracker with some popular trackers on **OTB100** database



(a) Success plots

(b) Precision plots

FIGURE 4 Comparison of success plot and precision plot of our proposed ASTRCF tracker with some popular trackers on **DTB70** database



(a) Success plots

(b) Precision plots

FIGURE 5 Comparison of success plot and precision plot of our proposed ASTRCF tracker with some popular trackers on **TC128** database

TABLE 2 Comparison of **success rate on 11 attributes** of our proposed ASTRCF tracker with some other 10 popular trackers on OTB100 database

Attributes	IVT	CT	DFT	KCF	CNT	SRDCF	HDT	BACF	STRCF	ASRCF	Ours
LR	0.261	0.232	0.192	0.313	0.466	0.499	0.484	0.516	0.579	0.597	0.600
IPR	0.309	0.273	0.324	0.470	0.528	0.547	0.556	0.561	0.597	0.644	0.646
OPR	0.309	0.279	0.334	0.457	0.555	0.543	0.558	0.575	0.627	0.668	0.670
SV	0.296	0.257	0.267	0.417	0.520	0.572	0.531	0.584	0.647	0.662	0.663
OCC	0.284	0.290	0.337	0.453	0.553	0.549	0.550	0.569	0.620	0.668	0.672
DEF	0.243	0.261	0.322	0.430	0.520	0.520	0.509	0.570	0.602	0.662	0.660
BC	0.295	0.287	0.345	0.479	0.532	0.539	0.530	0.576	0.628	0.662	0.671
IV	0.314	0.275	0.314	0.464	0.541	0.572	0.532	0.603	0.637	0.677	0.685
MB	0.211	0.224	0.260	0.473	0.558	0.616	0.613	0.602	0.669	0.667	0.672
FM	0.216	0.213	0.272	0.466	0.531	0.603	0.571	0.602	0.634	0.657	0.657
OV	0.220	0.296	0.294	0.405	0.533	0.477	0.514	0.539	0.600	0.658	0.647
Overall	0.315	0.272	0.327	0.479	0.563	0.592	0.577	0.613	0.654	0.688	0.688

TABLE 3 Comparison of **precision rate on 11 attributes** of our proposed ASTRCF tracker with some other 10 popular trackers on OTB100 database

Attribute	IVT	CT	DFT	KCF	CNT	SRDCF	HDT	BACF	STRCF	ASRCF	Ours
LR	0.410	0.337	0.262	0.514	0.643	0.625	0.686	0.659	0.723	0.740	0.744
IPR	0.423	0.369	0.413	0.633	0.663	0.672	0.723	0.709	0.736	0.795	0.803
OPR	0.414	0.367	0.429	0.616	0.697	0.665	0.717	0.717	0.776	0.813	0.821
SV	0.399	0.336	0.342	0.596	0.672	0.693	0.703	0.730	0.787	0.806	0.816
OCC	0.373	0.361	0.417	0.584	0.677	0.660	0.694	0.690	0.752	0.791	0.804
DEF	0.326	0.316	0.400	0.563	0.653	0.649	0.648	0.707	0.764	0.821	0.823
BC	0.382	0.351	0.402	0.622	0.642	0.651	0.651	0.703	0.775	0.793	0.815
IV	0.391	0.345	0.373	0.622	0.670	0.681	0.670	0.712	0.756	0.800	0.819
MB	0.229	0.245	0.266	0.565	0.640	0.709	0.706	0.709	0.770	0.769	0.786
FM	0.245	0.243	0.302	0.583	0.639	0.708	0.692	0.720	0.741	0.777	0.787
OV	0.268	0.312	0.333	0.479	0.637	0.565	0.632	0.660	0.718	0.786	0.783
Overall	0.409	0.347	0.403	0.635	0.701	0.715	0.734	0.753	0.793	0.831	0.837

0.021, respectively. The main reason is that the temporal regularization in our tracker can learn more appropriate correlation filter by taking full use of more information of the previous correlation filter.

3.2 | Attribute-based performances

In this subsection, we want to exhibit the performances of our proposed ASTRCF tracker on 11 attributes of video sequences. Table 2 shows the comparison of success rate on 11 attributes of our proposed ASTRCF tracker with KCF [2], SRDCF [3], BACF [6], CNT [7], HDT [8], STRCF [9], ASRCF [11], CT [17], IVT [18], and DFT [19] on OTB100 database.

As observed from Table 2, the rank of top three success rates on 11 attributes are our ASTRCF, ASRCF, and STRCF tracker successively. The reason is that our proposed ASTRCF tracker considers the background information, spatial information, adaptive weight, and temporal information simultaneously. In detail, our proposed ASTRCF tracker ranks the first on nine

attributes, *andranksthesecondonthelefttwo*

attributes: OV and DEF. In the aspect of OV attribute, the value of our ASTRCF tracker is 0.011 less than that of ASRCF tracker. The reason is that there is a temporal regularization term in the appearance model of our ASTRCF, but not in ASRCF tracker. While OV attribute means that there is a huge variation on object between the adjacent frames, so the temporal regularization may influence the tracking result.

Table 3 shows the comparison of precision rate on 11 attributes of our proposed ASTRCF tracker with KCF [2], SRDCF [3], BACF [6], CNT [7], HDT [8], STRCF [9], ASRCF [11], CT [17], IVT [18], and DFT [19] on OTB100 database.

As observed from Table 3, the rank of top three precision rates on 11 attributes are our ASTRCF, ASRCF, and STRCF tracker successively. The reason is that our proposed ASTRCF tracker considers the background information, spatial

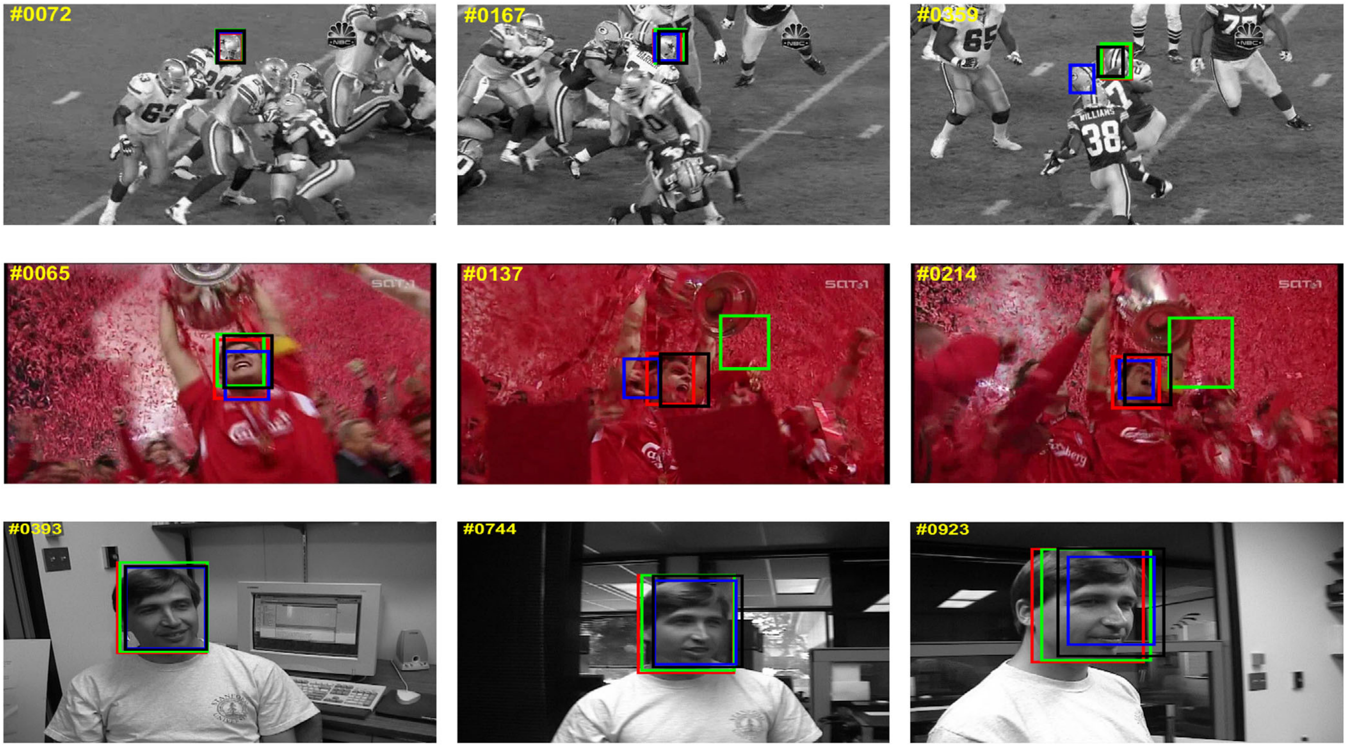


FIGURE 6 Tracking results of our proposed ASTRCF tracker (red box) with three trackers: HDT (blue box), BACF (black box), and STRCF (green box) over Football, Soccer and Dudek sequences with **background clutters**

information, adaptive weight, and temporal information simultaneously. In detail, our proposed ASTRCF tracker ranks the first on 10 attributes, and ranks the second on the left OV attribute. The reason is that the video sequences with OV attribute have drastic variations on the objects' appearances between the adjacent frames, while the temporal regularization in our ASTRCF tracker maybe interrupt the tracking results.

3.3 | Qualitative evaluation

In this subsection, we perform the qualitative evaluation of our proposed ASTRCF tracker with three outstanding trackers on five attributes: background clutters, deformation, fast motion, occlusion, and scale variation.

3.3.1 | Background clutters (BC)

Figure 6 shows several tracking results of our proposed ASTRCF tracker with HDT, BACF, and STRCF trackers on nine

frameso fthree

challenging video sequences with **background clutters**. In these frames, the colors of the backgrounds around the targets are similar to the targets over the whole processing. For example, the background and foreground of the "Soccer" sequence have extremely high similarity. The man wearing the red T-shirt is surrounded by the red flags (e.g. #65, #137, and #214) and some occlusions (e.g. golden cup).

As observed from Figure 6, only our ASTRCF tracker (red box) keeps grasping the target accurately. Although STRCF tracker can also locate the target, the precision of location is lower than ours. While BACF tracker and HDT tracker emerge different degree of tracking drift. Concretely, in the "Football" sequence, there are some partial occlusions (e.g. #72 and #167) and multiple clutters in the background (e.g. #359). Although our ASTRCF tracker and STRCF tracker both have good performances for slight occlusion, only our tracker can track the target well in the case of background clutters. While BACF tracker and HDT tracker drift into the background apparently. In the "Dudek" sequence, the cluttered background has a slight influence on target tracking, especially in frame #744 and frame #923. Obviously, our tracking results are more accurate than those of other three trackers.

3.3.2 | Deformation (DEF)

Figure 7 shows several tracking results of our proposed ASTRCF tracker with HDT, BACF, and STRCF trackers on nine

frameso fthree

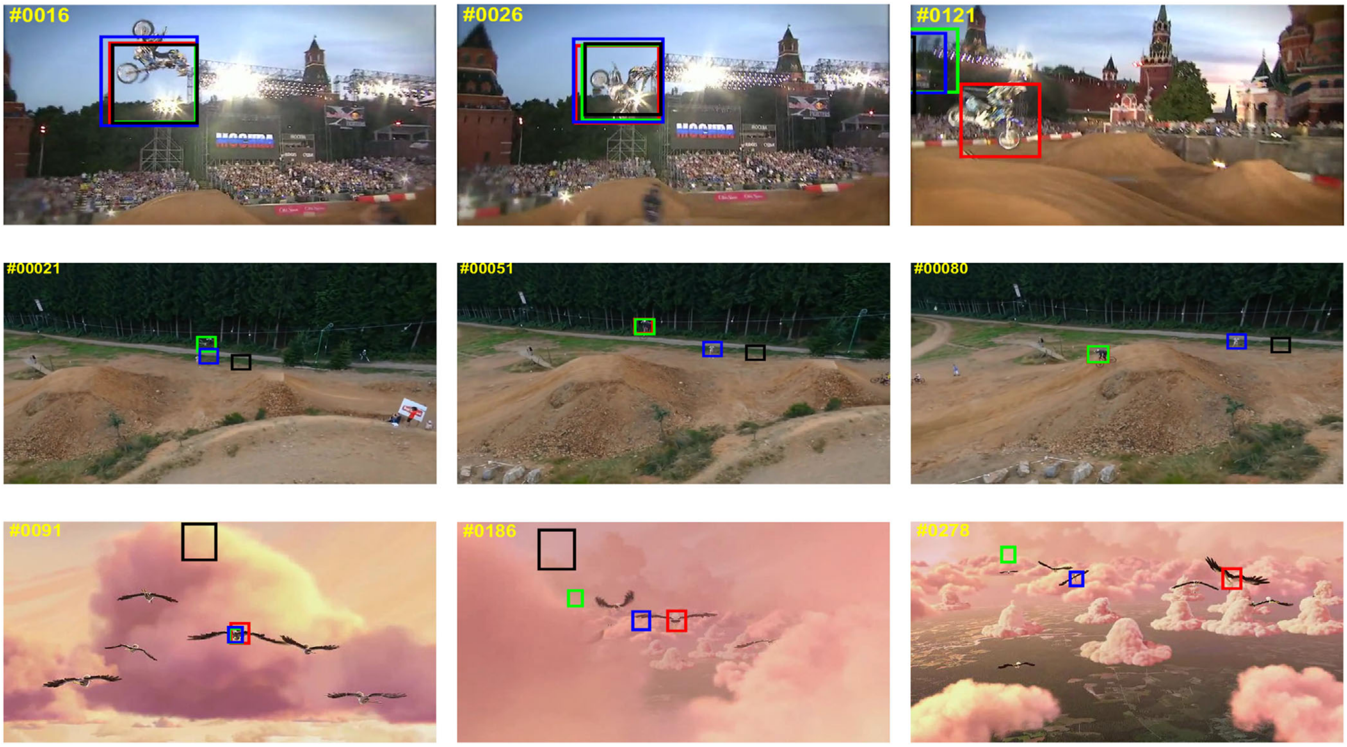


FIGURE 7 Tracking results of our proposed ASTRCF tracker (red box) with three trackers: HDT (blue box), BACF (black box), and STRCF (green box) over MotorRolling, MountainBike6, and Bird1 sequences with **deformation**

challenging video sequences with **deformation**. In these frames, the appearances of the targets happen to deform over the whole processing. For example, in the "MountainBike6" sequence, athletes constantly deform during running (e.g. #21, #51, and #80). In the "Bird1" sequence, the birds have large deformations during flying through the heavy clouds, and there are some other birds as background clutters. Hence, deformation and occlusion are two important attributions for this sequence (e.g. #91, #186, and #278).

As observed from Figure 7, only our ASTRCF tracker (red box) keeps grasping the target accurately. In the "MotorRolling" sequence, the motor's movement changes dramatically within three frames. Even the target is surrounded by the clutters closely, our tracker (red box) can still succeed to track the target accurately, while HDT (blue box), BACF (black box), and STRCF (green box) trackers fail to locate the target (e.g. #26 and #121). In the "MountainBike6" sequence, our ASTRCF tracker does not lose the target no matter how the bike's movement changes. In the "Bird1" sequence, even interfered by deformation and occlusion, our proposed ASTRCF tracker can still track the target accurately. While HDT (blue box), BACF (black box), and STRCF (green box) trackers sometimes fail to locate the target.

3.3.3 | Fast motion (FM)

In order to further evaluate the performances of our ASTRCF tracker in handling the target with fast motion, we compare our proposed ASTRCF tracker with HDT, BACF, and STRCF trackers on nine

frameso fthree

challenging video sequences with large **fast motion**. The tracking results are shown in Figure 8.

As observed from Figure 8, only our ASTRCF tracker (red box) keeps grasping the target accurately. In the "Surfing12" sequence, the ship are moving fast, which leads to the fast motion of the target (e.g. #21, #63, and #109). Our ASTRCF tracker can track the target accurately, while HDT, BACF, and STRCF trackers emerge tracking drift (e.g. #63 and #109). In the "Surfing03" sequence, besides fast motion, in-plane rotation (e.g. #39) and the scale variation (e.g. #64 and #111) also occur. Obviously, our proposed ASTRCF tracker can still track the target accurately, while HDT, BACF, and STRCF trackers emerge tracking drift (e.g. #64 and #111). In the "DragonBaby" sequence, baby moves so fast that motion blur occurs (e.g. #50). For this case, our proposed ASTRCF tracker can still grasp the face of baby accurately, while HDT, BACF, and STRCF trackers emerge tracking drift (e.g. #50 and #93).

3.3.4 | Occlusion (OCC)

Generally, the object is easily susceptible to heavy occlusion. Figure 9 shows the tracking results of our proposed ASTRCF

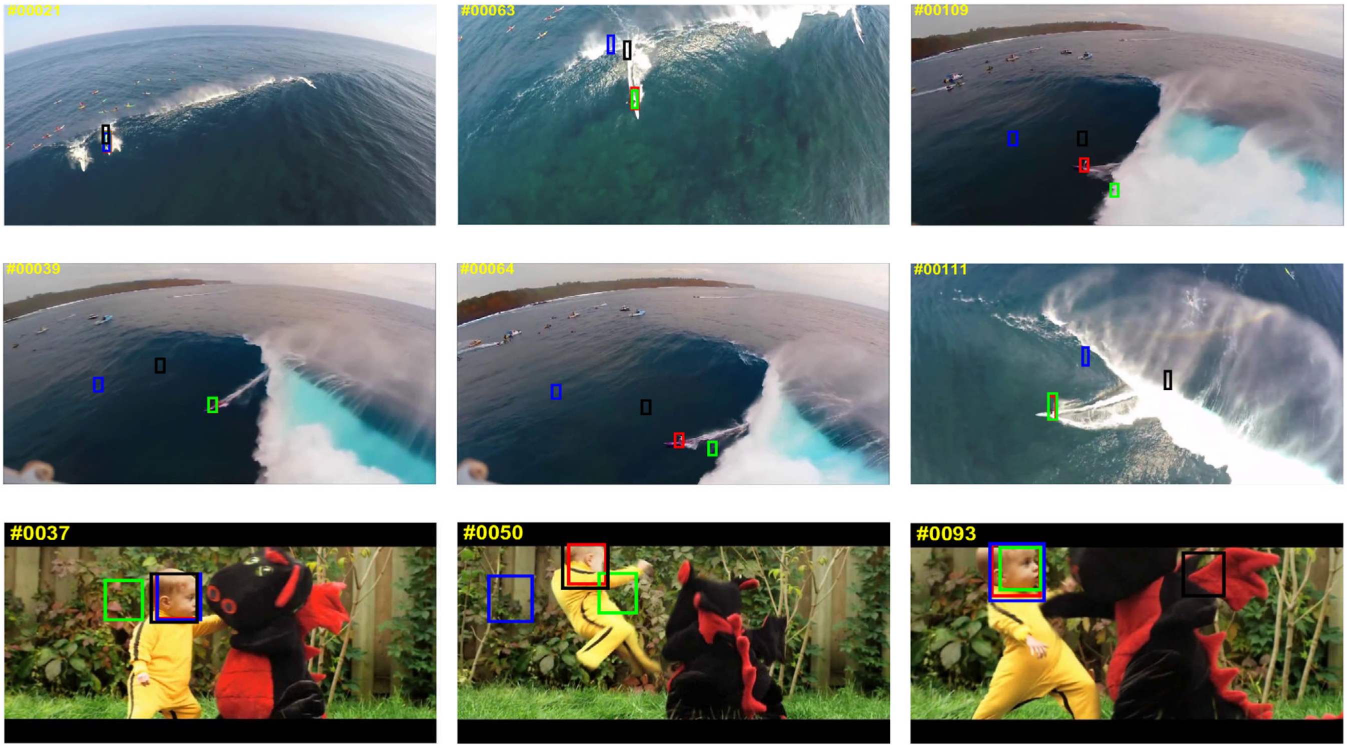


FIGURE 8 Tracking results of our proposed ASTRCF tracker (red box) with three trackers: HDT (blue box), BACF (black box), and STRCF (green box) trackers over Surfing12, Surfing03, and DragonBaby sequences with **fast motion**

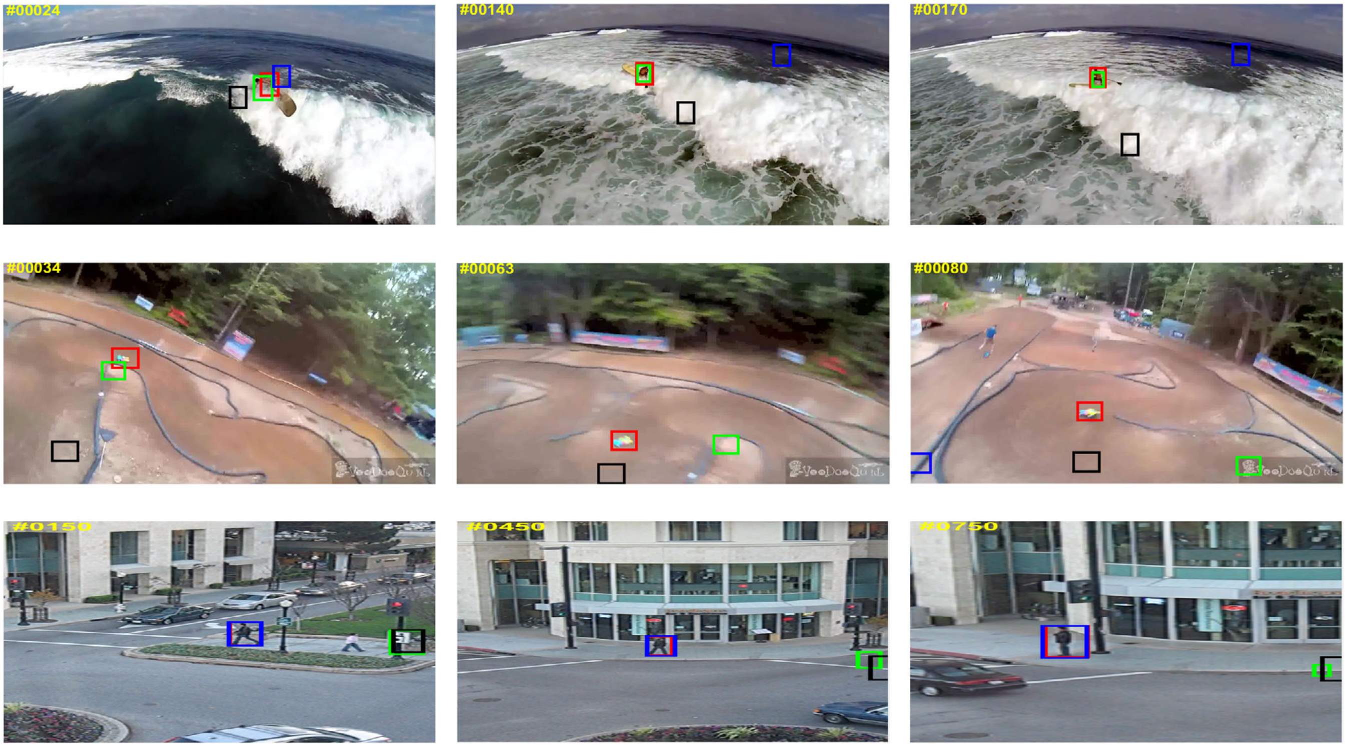


FIGURE 9 Tracking results of our proposed ASTRCF tracker (red box) with three trackers: HDT (blue box), BACF (black box), and STRCF (green box) trackers over Sup5, Rccar4, and Human3 sequences with **Occlusion**

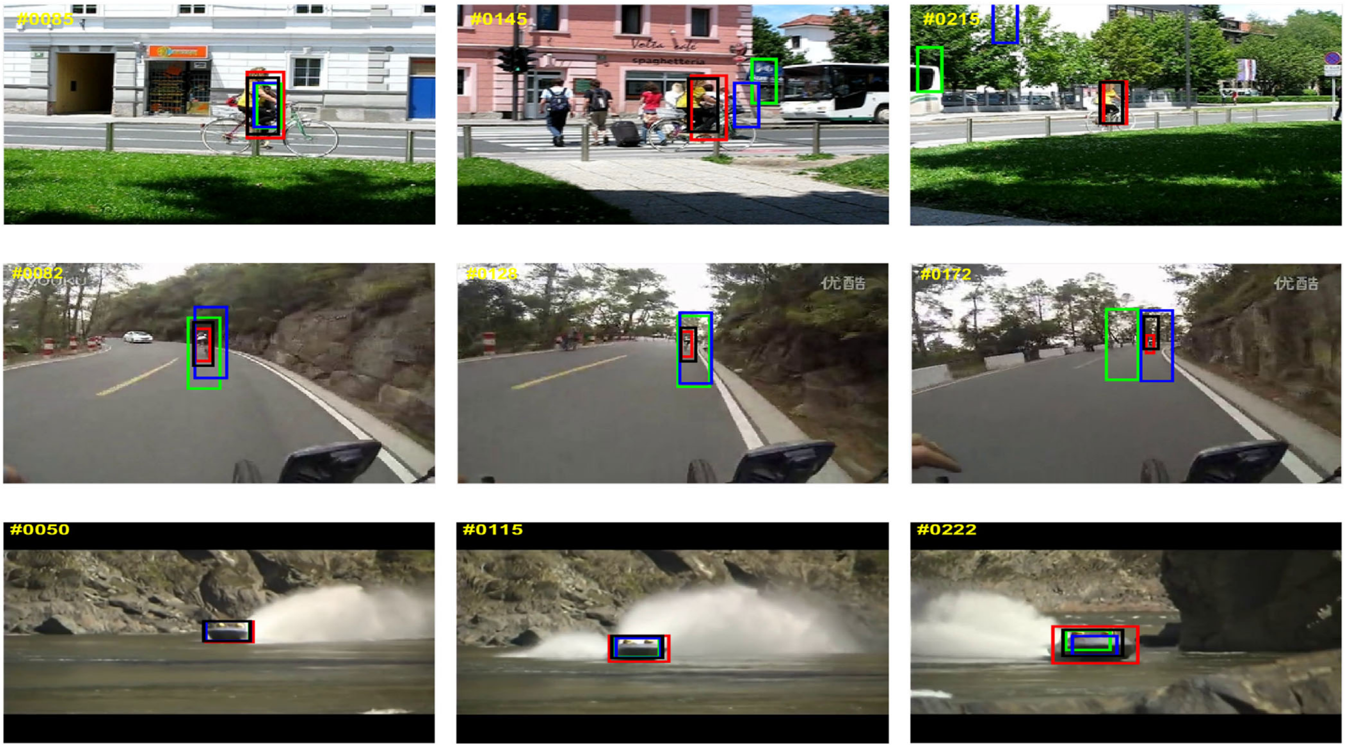


FIGURE 10 Tracking results of our proposed ASTRCF tracker (red box) with three trackers: HDT (blue box), BACF (black box), and STRCF (green box) trackers over Bicycle1, Bicycle2, and Boat with **Scale Variation**

tracker with HDT, BACF, and STRCF trackers on 9 frames of 3 challenging video sequences with **Occlusion**.

As observed from Figure 9, only our ASTRCF tracker (red box) keeps grasping the target accurately. In the "Sup5" sequence, it is seen that even in the case of background disturbance, our tracker still lock the target precisely (e.g. #24, and #140). While the other three trackers, especially HDT and BACF trackers, appear the phenomena of tracking drift. In the "Rcccar4" sequence, only our tracker tracks the target well (e.g. #34, and #80). While the other three trackers lose the target obviously. In the "Human3" sequence, the target is constantly occluded by different objects and some other pedestrians. Our tracker and HDT tracker grasp the target well, while BACF and STRCF trackers lose the target obviously.

3.3.5 | Scale variation (SV)

Generally, the target often suffers from the scale variation in the tracking processing. Figure 10 shows the comparison results of our proposed ASTRCF tracker with HDT, BACF, and STRCF trackers on nine

frameso fthree

challenging video sequences with **scale variation**.

As observed from Figure 10, our ASTRCF tracker (red box) can always keep grasping the targets accurately. For example, in the "Bicycle1" sequence, even the target has the scale variation, our tracker (red box) and BACF (black box) can still succeed to track the target, while HDT (blue box) and STRCF (green box) trackers emerge tracking drift (e.g. #145 and #215). In the "Bicycle2" sequence, our ASTRCF tracker does not lose the target no matter how the bicycle's scale changes. Furthermore, the scale of tracking box by our tracker can be adjusted according to the object's size.

3.4 | Tracking speed

Table 4 shows the comparison of tracking speed (frames per second, FPS) of our proposed ASTRCF tracker with KCF [2], SRDCF [3], BACF [6], CNT [7], HDT [8], STRCF [9], ASRCF [11], CT [17], IVT [18], and DFT [19] on DTB70 database.

As observed in Table 4, we can see that KCF tracker owns the highest tracking speed because of its simple appearance model, then BACF, ASRCF, and our ASTRCF trackers follow up. Our tracker can reach 27 FPS, which satisfies the requirement of real-time. The reason is that our ASTRCF tracker takes the deep features and has complicated appearance model.

4 | CONCLUSIONS

This paper has proposed an ASTRCF tracker for visual tracking. The proposed tracker designs an improved appearance model

TABLE 4 Comparison of tracking speed (FPS) of our proposed ASTRCF tracker with some other 10 popular trackers

Trackers	IVT	CT	DFT	KCF	CNT	SRDCF	HDT	BACF	STRCF	ASRCF	Ours
FPS	8.4	12.9	13.2	171	1.5	5	2.7	35.3	5.3	28	27

for learning the correlation filter based on ASRCF tracker with regularization theory. The appearance model contains a background awareness term, an adaptive spatial regularization term, and a temporal regularization. Therefore, it can not only take full advantage of background information, the spatial information, and the adaptivity of weight matrix, but also use the temporal information for optimizing the correlation filter. The designed appearance model can improve the robustness of our method against the occlusion, large appearance deformation, fast motion and background clutter effectively. Extensive experimental results illustrate that our proposed ASTRCF tracker is superior to several state-of-the-art trackers in terms of accuracy and robustness.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Jianwei Zhao https://orcid.org/0000-0001-9566-2178

REFERENCES

1. Bolme, D.S. et al.: Visual object tracking using adaptive correlation filters. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2544–2550. IEEE, Piscataway, NJ (2010)
2. Henriques, J.F., et al.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. 37(3), 583–596 (2014)
3. Danelljan, M. et al.: Learning spatially regularized correlation filters for visual tracking. In: IEEE International Conference on Computer Vision, pp. 4310–4318. IEEE, Piscataway, NJ (2015)
4. Bertinetto, L. et al.: Staple: complementary learners for real-time tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1401–1409. IEEE, Piscataway, NJ (2016)
5. Mueller, M., Smith, N., Ghanem, B.: Context-aware correlation filter tracking. In: IEEE International Conference on Computer Vision, pp. 1387–1395. IEEE, Piscataway, NJ (2017)
6. Galoogahi, H.K., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1144–1152. IEEE, Piscataway, NJ (2017)
7. Zhang, K.H., Liu, Q.S., Wu, Y.: Robust visual tracking via convolutional networks without training. IEEE Trans. Image Process. 25(4), 1779–1792 (2016)
8. Qi, Y.K., Zhang, S.P., Qin, L.: Hedged deep tracking. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 4303–4311. IEEE, Piscataway, NJ (2016)
9. Li, F. et al.: Learning spatial-temporal regulairzed correlation filters for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4904–4913. IEEE, Piscataway, NJ (2018)
10. Yuan, D., Shu, X., He, Z.Y.: TRBACF: learning temporal regularized correlation filters for high performance online visual object tracking. J. Visual Commun. Image Represent. 72, (2020), 102882 https://doi.org/10.1016/j.jvcir.2020.102882
11. Dai, K.N. et al.: Visual tracking via adaptive spatially-regularized correlation filter. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4665–4674. IEEE, Piscataway, NJ (2019)
12. Boyd, S., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3(1), 1–122 (2011)
13. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2411–2418. IEEE, Piscataway, NJ (2013)
14. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. 37(9), 1834–1848 (2015)
15. Li, S.Y., Yeung, D.Y.: Visual object tracking for unmanned aerial vehicles: a benchmark and new motion models. In: 31st AAAI Conference on Artificial Intelligence, pp 4140–4146 (2017)
16. Liang, P.P., Erik, B., Ling, H.B.: Encoding color information for visual tracking: algorithms and benchmark. IEEE Trans. Image Process. 24(12), 5630–5644 (2015)
17. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: European Conference on Computer Vision, pp. 864–877. Springer, Berlin (2012)
18. Ross, D.A., et al.: Incremental learning for robust visual tracking. Int. J. Comput. Vision 77(1-3), 125–141 (2008)
19. Learnedmiller, E., Sevillalara, L.: Distribution fields for tracking. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1910–1917. IEEE, Piscataway, NJ (2012)

How to cite this article: Zhao J, Li Y, Zhou Z. Learning adaptive spatial–temporal regularized correlation filters for visual tracking. IET Image Process. 2021;15:1773–1785. https://doi.org/10.1049/ipr2.12150