Adaptive Spatio-Temporal Regularized Correlation Filters for UAV-based Tracking

Libin Xu[1], Qilei Li[2], Jun Jiang[3],
Guofeng Zou[1], Zheng Liu[4], and Mingliang Gao[1⋆]

[1] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China
[2] School of Electronics and Information Engineering, Sichuan University, Chengdu, 610065, China
[3] School of Computer Science, Southwest Petroleum University, Chengdu, 610500, China
[4] Faculty of Applied Science, The University of British Columbia, Vancouver, BC V1V 1V7, Canada

Abstract. Visual tracking on unmanned aerial vehicles (UAVs) has enabled many new practical applications in computer vision. Meanwhile, discriminative correlation filter (DCF)-based trackers have drawn great attention and undergone remarkable progress due to their promising performance and efficiency. However, the boundary effect and filter degradation remain two challenging problems. In this work, a novel Adaptive Spatio-Temporal Regularized Correlation Filter (ASTR-CF) model is proposed to address these two problems. The ASTR-CF can optimize the spatial regularization weight and the temporal regularization weight simultaneously. Meanwhile, the proposed model can be effectively optimized based on the alternating direction method of multipliers (ADMM). Experimental results on DTB70 and UAV123@10fps benchmarks have proven the superiority of the ASTR-CF tracker compared to the state-of-the-art trackers in terms of both accuracy and computational speed.

Keywords: UAV Tracking · Correlation Filter · Spatio-Temporal Regularization .

1 Introduction

The advance of visual tracking has provided UAV with the intriguing capability for various practical applications. Differing from the generic tracking, UAV-based tracking poses new challenges to the tracking problem, *e.g.*, rapid changes in scale and perspective, limited pixels in the target region, and multiple similar disruptors [1].

Recently, discriminative correlation filter (DCF)-based trackers brought the performance of tracking into a new level [2,3,4,5,6]. One of the prominent merits that highlights the DCF-based trackers is that DCF is efficient in the training

⋆ Corresponding author, Email: *mlgao@sdut.edu.cn*

and detection stage as they can be transferred into the Fourier domain and operated in element-wise multiplication, which is of significance for the real-time tracking. However, it is still challenging to achieve high-performance tracking for an arbitrary object in unconstrained scenarios. The main obstacles include spatial boundary effect and temporal filter degeneration [6].

Learning DCF in the frequency domain comes at the high cost of learning from circularly shifted examples of the foreground target, thus it produces the unwanted boundary effects. This dilemma has been alleviated to some extent with additional pre-defined spatial constraints on filter coefficients. For example, Danelljan *et al.* [7] introduced the Spatially Regularized Discriminative Correlation Filters (SRDCF) to mitigate boundary effects. With the coefficient spatially penalized according to their distance to the center, the tracker is expected to focus on information near the center. Galoogahi *et al.* [8] multiplied the filter directly with a binary matrix to generate real positive and negative samples for model training. The aforementioned two spatial constraints are widely used in the subsequent research works [9,10,11].

The appearance model of most DCF-based trackers is updated via a linear interpolation approach and it cannot adapt to ubiquitous appearance change, leading to filter degradation inevitably. Some attempts are made to tackle the problem of filter degradation, *e.g.*, training set management [12,13,14], temporal restriction [15,16], tracking confidence verification [17,18] and over-fitting alleviation [19]. Among them, the temporal regularization is proven to be an effective way.

In this work, the problems of boundary effect and filter degradation are solved by the proposed adaptive spatio-temporal regularized correlation filters (ASTR-CF). Meanwhile, the ASTR-CF is applied to real-time UAV target tracking. We compared our approach with state-of-the-art trackers on DTB70 and UAV123-@10fps benchmarks. The results demonstrate that ASTR-CF outperforms state-of-the-art trackers in terms of accuracy and computational speed.

2 Related work

Despite the great success of DCF in visual tracking, it remains a challenge to achieve high performance tracking for an arbitrary object in unconstrained scenarios due to the intrinsic problems of spatial boundary effect and temporal filter degradation [15]. To solve these problems, spatial regularization and temporal regularization are introduced to the DCF framework successively.

2.1 Spatial regularization

Learning DCF in the frequency domain produces unwanted boundary effects which reduce the tracking performance. [7,15]. To alleviate the boundary effect problem, SRDCF [7] stimulates the interest in spatial regularization which allocates more energy for the central region of a filter using a predefined spatial weighting function. A similar idea has been pursued through pruning the training

samples or learned filters with a predefined mask [20,21,22]. A common characterization of the approaches above is that they are all based on a fixed spatial regularization pattern, decreasing the ambiguity emanating from the background and resulting in a relatively large search window for tracking. Different from the aforementioned approaches, Dai *et al.* [11] proposed an Adaptive Spatially Regularized Correlation Filters (ASRCF) model, which could estimate an object-aware spatial regularization and obtain more reliable filter coefficients during the tracking process.

2.2 Spatio-temporal regularization

Most of DCF-based trackers updated the appearance model via a linear interpolation approach, but it cannot adapt to ubiquitous appearance changes. To address this problem, Li *et al.* [15] introduced a temporal regularization module to SRDCF and incorporated both spatial and temporal regularization into DCF. The improved version, named as STRCF, is a rational approximation of the full SRDCF formulation on multiple training images, and can also be exploited for simultaneous DCF learning and model updating. Although STRCF has achieved competent performance, it remains two limitations. **(i)** The fixed spatial regularization failing to address appearance variation in the unforeseeable aerial tracking scenarios. **(ii)** The unchanged temporal penalty strength μ (set as 15 in [15]) is not general in all kinds of situations.

In this work, a novel Adaptive Spatio-Temporal Regularized Correlation Filters (ASTR-CF) model was proposed to estimate an object-aware spatial regularization and context-aware temporal regularization. The overall procedure of the tracking process is shown in Fig. 1. Meanwhile, the ADMM algorithm is directly introduced to solve the ASTR-CF model making it more generic.

3 ASTR-CF

3.1 Objective function of ASTR-CF

CF: The original multi-channel CF model in the spatial domain aims to minimize the following objective function [4],

$$E(\mathbf{H}) = \frac{1}{2}\left\|\mathbf{y} - \sum_{k=1}^{K} \mathbf{x}^k * \mathbf{h}^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{h}^k\right\|_2^2. \tag{1}$$

Here, $\mathbf{x}^k \in \mathbb{R}^{T\times 1}(k = 1, 2, 3, \ldots, K)$ and $\mathbf{h}^k \in \mathbb{R}^{T\times 1}(k = 1, 2, 3, \ldots, K)$ denote the extracted feature with length T in the t-th frame and filter trained in the t-th frame respectively. The vector $\mathbf{y} \in \mathbb{R}^{T\times 1}$ is the desired response (*i.e.*, the Gaussian-shaped ground truth) and $*$ denotes the convolution operator. $\mathbf{H} = \left[\mathrm{h}^1, \mathrm{h}^2, \ldots, \mathrm{h}^K\right]$ is the matrix representing the filters from all the K channels.

The original CF model suffers from periodic repetitions on boundary positions caused by circulant shifted samples, which inevitably degrades the tracking

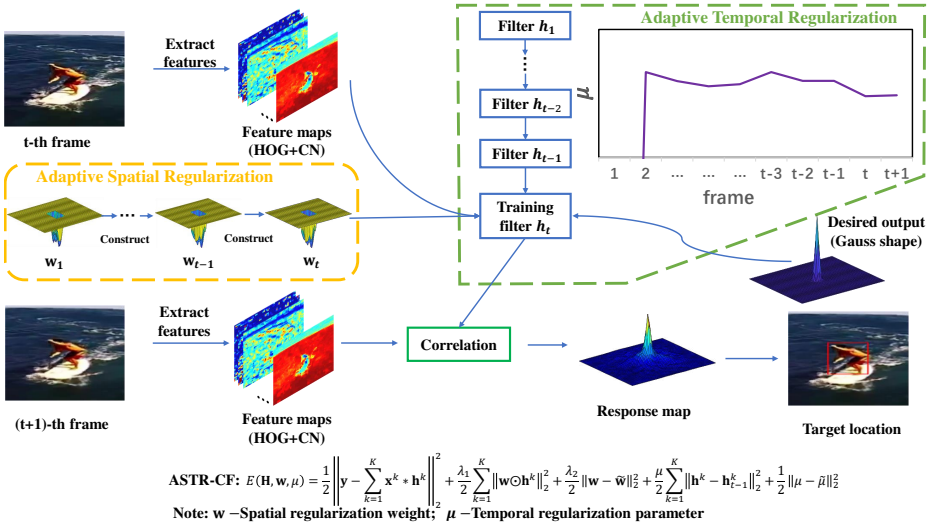

Fig. 1. Tracking framework based on the proposed ASTR-CF. In the training stage, a training patch is cropped at the estimated location of the target at the t-th frame. We extract the feature (HOG [23] and Color Names [24]) maps combined with prior filters and $\mathbf{w}$ to train the current filter. At the $(t+1)$-th frame, the trained filter is used to produce a response map, based on which the target is located.

performance. To solve this problem, several spatial constraints have been introduced to alleviate unexpected boundary effects. The representative methods are SRDCF [7] and STRCF [15].

SRDCF: The SRDCF method [7] introduces a spatial regularization to penalize the filter coefficients with respect to their spatial locations and the objective function is formulated as,

$$E(\mathbf{H}) = \frac{1}{2}\left\|\mathbf{y} - \sum_{k=1}^{K}\mathbf{x}^k * \mathbf{h}^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\widetilde{\mathbf{w}} \odot \mathbf{h}^k\right\|_2^2. \tag{2}$$

Here, $\widetilde{\mathbf{w}}$ is a negative Gaussian-shaped spatial weight vector to make the learned filters have a high response around the center of the tracked object. However, although SRDCF is effective in suppressing the adverse boundary effects, it also increases the computational burden due to the following two reasons. **(i)** The failure of exploiting circulant matrix structure. **(ii)** The large linear equations and Gauss-Seidel solver. More implementation details are refer to [7].

STRCF: The STRCF model [15] introduces a spatial-temporal regularized module to CF and the objective function is formulated as,

$$E(\mathbf{H}) = \frac{1}{2}\left\|\mathbf{y} - \sum_{k=1}^{K}\mathbf{x}^k * \mathbf{h}^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\widetilde{\mathbf{w}} \odot \mathbf{h}^k\right\|_2^2 + \frac{\mu}{2}\sum_{k=1}^{K}\left\|\mathbf{h}^k - \mathbf{h}_{t-1}^k\right\|_2^2. \tag{3}$$

Here, $\mathbf{x}^k \in \mathbb{R}^{T\times 1}(k = 1, 2, 3, \ldots, K)$ is the extracted feature with length T in frame t. $\mathbf{h}^k, \mathbf{h}^k_{t-1} \in \mathbb{R}^{T\times 1}(k = 1, 2, 3, \ldots, K)$ denote the filter of the t-th channel trained in the k-th and (t-1)-th frame respectively. As for regularization, the spatial regularization parameter $\widetilde{\mathbf{w}}$ is imitated from SRDCF [7] to decrease boundary effect, and temporal regularization(the third term in Eq. (3), is firstly proposed to restrict filter's variation by penalizing the difference between the current and previous filters.

However, as aforementioned, the spatial regularization and temporal penalty strength of STRCF [15] are fixed. Therefore, it fails to address the appearance variation in the unforeseeable aerial tracking scenarios.

Our Objective Function: Motivated by the discussions above, we propose a novel ASTR-CF method to learn effective multi-channel CFs, and our objective function is defined as follows,

$$\begin{aligned} E(\mathbf{H}, \mathbf{w}, \mu) =& \frac{1}{2}\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{x}^k * \mathbf{h}^k \right\|_2^2 + (\frac{\lambda_1}{2}\sum_{k=1}^{K}\left\|\mathbf{w}\odot\mathbf{h}^k\right\|_2^2 + \frac{\lambda_2}{2}\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2^2) + \\ & \left(\frac{\mu}{2}\sum_{k=1}^{K}\left\|\mathbf{h}^k - \mathbf{h}^k_{t-1}\right\|_2^2 + \frac{1}{2}\|\mu-\tilde{\mu}\|_2^2\right). \end{aligned} \tag{4}$$

Here, the first term is the ridge regression term that convolves the training data $\mathbf{X} = \left[\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K\right]$ with the filter $\mathbf{H} = \left[\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^K\right]$ to fit the Gaussian-distributed ground truth $\mathbf{y}$. The second term introduces an adaptive spatial regularization on the filter $\mathbf{H}$. The spatial weight $\mathbf{w}$ requires to be optimized to approximate a reference weight $\widetilde{\mathbf{w}}$. This constraint introduces prior information on $\mathbf{w}$ and avoids model degradation. λ_1 and λ_2 are the regularization parameters of the second terms. The third term introduces an adaptive temporal regularization, where $\tilde{\mu}$ and μ denote the reference and optimized context-aware temporal regularization parameter respectively [6]. $\tilde{\mu}$ is denoted as ,

$$\tilde{\mu} = \frac{\zeta}{1 + \log\left(\nu||\mathbf{\Pi}||_2 + 1\right)}, \quad \|\mathbf{\Pi}\|_2 \leq \phi. \tag{5}$$

Here, $\mathbf{\Pi} = \left[\left|\Pi^1\right|, |\Pi^2|, \cdots \left|\Pi^T\right|\right]$ denotes the response variations. ζ and ν denote hyper parameters.

3.2 Optimization of ASTR-CF

We express the objective function *i.e.*, Eq. (4), in the frequency domain using Parseval's theorem, and convert it into the equality constrained optimization

form,

$$\begin{aligned} E(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu) =& \frac{1}{2}\left\|\widehat{\mathbf{y}} - \sum_{k=1}^{K} \widehat{\mathbf{x}}^k \odot \widehat{\mathbf{g}}^k\right\|_2^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K}\left\|\mathbf{w}\odot\mathbf{h}^k\right\|_2^2 + \\ & \frac{\lambda_2}{2}\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2^2 + \frac{\mu}{2}\sum_{k=1}^{K}\left\|\widehat{\mathbf{g}}^k - \widehat{\mathbf{g}}_{t-1}^k\right\|_2^2 + \frac{1}{2}\|\mu-\tilde{\mu}\|_2^2, \\ s.t., \widehat{\mathbf{g}}^k =& \sqrt{T}\mathbf{F}\mathbf{h}^k,\ k=1,\ldots,K. \end{aligned} \tag{6}$$

Here, $\widehat{\mathbf{G}} = \left[\widehat{\mathbf{g}}^1, \widehat{\mathbf{g}}^2, \ldots, \widehat{\mathbf{g}}^K\right]$ $\left(\widehat{\mathbf{g}}^k = \sqrt{T}\mathbf{F}\mathbf{h}^k, k = 1, 2, \ldots, K\right)$ is an auxiliary variable matrix. The symbol $\widehat{}$ denotes the discrete Fourier transform form of a given signal, and $\mathbf{F}$ is the orthonormal $T \times T$ matrix of complex basis vectors to map any T dimensional vectorized signal into the Fourier domain. The model in Eq. (6) is bi-convex, and can be minimized to obtain a local optimal solution using ADMM [25]. The augmented Lagrangian form of Eq. (6) can be formulated as,

$$\begin{aligned} L(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu, \widehat{\mathbf{V}}) =& E(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu) + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k\right\|_2^2 + \\ & \sum_{k=1}^{K}\left(\widehat{\mathbf{v}}^k\right)^{\mathrm{T}}\left(\widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k\right). \end{aligned} \tag{7}$$

Here, $\mathbf{V}$ is the Lagrange multiplier, and $\widehat{\mathbf{V}}$ is the corresponding Fourier transform. By introducing $\mathrm{s}^k = \frac{1}{\gamma}\mathrm{v}^k$, the optimization of Eq. (7) is equivalent to solving,

$$L(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu, \widehat{\mathbf{S}}) = E(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu) + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k + \widehat{\mathbf{s}}^k\right\|_2^2. \tag{8}$$

Then, the ADMM algorithm is adopted by alternately solving the following subproblems.

Subproblem H: If $\widehat{\mathbf{G}}, \mathbf{w}, \mu$ and $\widehat{\mathbf{S}}$ are given, the optimal $\mathbf{H}^*$ can be obtained as,

$$\begin{aligned} \mathbf{h}^{k^*} &= \underset{\mathbf{h}^k}{\operatorname{argmin}}\left\{\frac{\lambda_1}{2}\left\|\mathbf{w}\odot\mathbf{h}^k\right\|_2^2 + \frac{\gamma}{2}\left\|\widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k + \widehat{\mathbf{s}}^k\right\|_2^2\right\} \\ &= \left[\lambda_1\mathbf{W}^{\mathrm{T}}\mathbf{W} + \gamma T\mathbf{I}\right]^{-1}\gamma T\left(\mathbf{g}^k + \mathbf{s}^k\right) \\ &= \frac{\gamma T\left(\mathbf{g}^k + \mathbf{s}^k\right)}{\lambda_1(\mathbf{w}\odot\mathbf{w}) + \gamma T}. \end{aligned} \tag{9}$$

Here, $\mathbf{W} = \mathrm{diag}\,(\mathbf{w}) \in \mathbb{R}^{T\times T}$. Eq. (9) shows that the solution of $\mathbf{h}^k$ merely requires the element-wise multiplication and the inverse fast Fourier transform (*i.e.*,$\mathbf{g}^k = \frac{1}{\sqrt{T}}\mathbf{F}^{\mathrm{T}}\widehat{\mathbf{g}}^k$ and $\mathbf{s}^k = \frac{1}{\sqrt{T}}\mathbf{F}^{\mathrm{T}}\widehat{\mathbf{s}}^k$).

Subproblem $\widehat{\mathbf{G}}$: If $\mathbf{H}$, $\mathbf{w}$, μ, and $\widehat{\mathbf{S}}$ are given, the optimal $\widehat{\mathbf{G}}^*$ can be estimated by solving the optimization problem as,

$$\begin{aligned}\widehat{\mathbf{G}}^* = \underset{\widehat{\mathbf{G}}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \mathbf{y} - \sum_{k=1}^{K} \widehat{\mathbf{x}}^k \odot \widehat{\mathbf{g}}^k \right\|_2^2 + \frac{\mu}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \widehat{\mathbf{g}}_{t-1}^k \right\|_2^2 + \right. \\ \left. \frac{\gamma}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k + \widehat{\mathbf{s}}^k \right\|_2^2 \right\}.\end{aligned} \tag{10}$$

However, it is difficult to optimize Eq. (10) due to its high computation complexity. Thus, we consider processing on all channels of each pixel to simplify our formulation written by,

$$\begin{aligned}\mathcal{V}_j^*(\widehat{\mathbf{G}}) = \underset{\mathcal{V}_j(\widehat{\mathbf{G}})}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \widehat{\mathbf{y}}_j - \mathcal{V}_j(\widehat{\mathbf{X}})^{\mathrm{T}} \mathcal{V}_j(\widehat{\mathbf{G}}) \right\|_2^2 + \frac{\mu}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{G}}) - \mathcal{V}_j\left(\widehat{\mathbf{G}}_{t-1}\right) \right\|_2^2 + \right. \\ \left. \frac{\gamma}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{G}}) + \mathcal{V}_j(\widehat{\mathbf{S}}) - \mathcal{V}_j(\sqrt{T}\mathbf{F}\mathbf{H}) \right\|_2^2 \right\}.\end{aligned} \tag{11}$$

Here, $\mathcal{V}_j(\widehat{\mathbf{X}}) \in \mathbb{C}^{K\times 1}$ denotes the values of all K channels of $\widehat{\mathbf{X}}$ on pixel j, $(j = 1, 2, \ldots, \mathrm{T})$. Then, the analytical solution of Eq. (11) can be obtained as,

$$\mathcal{V}^*(\widehat{\mathbf{G}}) = \frac{1}{\mu + \gamma} \left[\mathbf{I} - \frac{\mathcal{V}_j(\widehat{\mathbf{X}}) \mathcal{V}_j(\widehat{\mathbf{X}})^{\mathrm{T}}}{\mu + \gamma + \mathcal{V}_j(\widehat{\mathbf{X}})^{\mathrm{T}} \mathcal{V}_j(\widehat{\mathbf{X}})} \right] \rho, \tag{12}$$

here,

$$\rho = \mathcal{V}_j(\widehat{\mathbf{X}}) \widehat{\mathbf{y}}_j + \mu \left[\mathcal{V}_j\left(\widehat{\mathbf{G}}_{t-1}\right) \right] + \gamma \left[\mathcal{V}_j(\sqrt{T}\mathbf{F}\mathbf{H}) - \mathcal{V}_j(\widehat{\mathbf{S}}) \right]. \tag{13}$$

The derivation of Eq. (12) uses the Sherman Morrsion formula,

$$\left(\mathbf{A} + \mathbf{u}\mathbf{v}^{\mathrm{T}}\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}}{1 + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{u}}. \tag{14}$$

Here, $\mathbf{u}$ and $\mathbf{v}$ are two column vectors and $\mathbf{u}\mathbf{v}^{\mathrm{T}}$ is a rank-one matrix.

Solving $\mathbf{w}$: If $\mathbf{H}$, $\widehat{\mathbf{G}}$, μ and $\widehat{\mathbf{S}}$ are fixed, the closed-form solution regrading $\mathbf{w}$ can be determined as,

$$\begin{aligned}\mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \frac{\lambda_1}{2} \sum_{k=1}^{K} \left\| \mathbf{w} \odot \mathbf{h}^k \right\|_2^2 + \frac{\lambda_2}{2} \|\mathbf{w} - \widetilde{\mathbf{w}}\|_2^2 \right\} \\ &= \left[\lambda_1 libinxu_sd@outlook.com \sum_{k=1}^{K} \left(\mathbf{N}^k\right)^{\mathrm{T}} \mathbf{N}^k + \lambda_2 \mathbf{I} \right]^{-1} \lambda_2 \widetilde{\mathbf{w}} \\ &= \frac{\lambda_2 \widetilde{\mathbf{w}}}{\lambda_1 \sum_{k=1}^{K} \mathbf{h}^k \odot \mathbf{h}^k + \lambda_2 \mathbf{I}}.\end{aligned} \tag{15}$$

Here, $\mathbf{N}^k = \operatorname{diag}\left(\mathbf{h}^k\right) \in \mathbb{R}^{T\times T}$.

Solving μ: Given other variables $\mathbf{H}$, $\widehat{\mathbf{G}}$, $\mathbf{w}$, and $\widehat{\mathbf{S}}$, the optimal solution of μ can be determined as,

$$\begin{aligned}\mu^* &= \underset{\mu}{\operatorname{argmin}} \left\{ \frac{\mu}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \widehat{\mathbf{g}}_{t-1}^k \right\|_2^2 + \frac{1}{2} \|\mu - \tilde{\mu}\|_2^2 \right\} \\ &= \tilde{\mu} - \frac{1}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \widehat{\mathbf{g}}_{t-1}^k \right\|_2^2 .\end{aligned} \tag{16}$$

Lagrangian Multiplier Update: We update Lagrangian multipliers as,

$$\widehat{\mathbf{S}}^{i+1} = \widehat{\mathbf{S}}^i + \gamma^i \left(\widehat{\mathbf{G}}^{i+1} - \widehat{\mathbf{H}}^{i+1} \right). \tag{17}$$

Here, i and $i+1$ denote the iteration index. The step size regularization constant γ takes the form of $\gamma^{i+1} = \min\left(\gamma_{\max}, \beta\gamma^i\right)$ $(\beta = 10, \gamma_{\max} = 10000)$. By iteratively solving the five subproblems above, we can optimize our objective function effectively and obtain the optimal filter $\widehat{\mathbf{G}}$, object-aware spatial regularization weight $\mathbf{w}$ and context-aware temporal regularization parameter μ in frame t. Then $\widehat{\mathbf{G}}$ is used for detection in frame $t+1$.

3.3 Target localization

The location of the target can be determined in the Fourier domain as,

$$\widehat{\mathcal{R}}_t = \sum_{k=1}^{K} \widehat{\mathbf{x}}^k \odot \widehat{\mathbf{g}}_{t-1}^k. \tag{18}$$

Here, $\mathcal{R}_t$ and $\widehat{\mathcal{R}}_t$ denote the response map and its Fourier transform. After obtaining the response map, the optimal location can be obtained based on the maximum response.

4 Experimental Results

In this section, we demonstrate the effectiveness of our tracker on DTB70 [26] and UAV123@10fps [27] datasets among the current state-of-the-art trackers. We use the same evaluation criteria on the two benchmarks.

The experiments of tracking performance evaluation are conducted using MATLAB R2017a on a PC with an i7-8700K processor (3.7GHz), 32GB RAM, and an NVIDIA GTX 1080Ti GPU. For the hyper-parameters of our tracker, we set $\lambda_1 = 1, \lambda_2 = 0.001, \nu = 2 \times 10^{-5}$, and $\zeta = 13$. The threshold of ϕ is 3000, and the ADMM iteration is set to 4.

4.1 Quantitative evaluation

DTB70: The DTB70 dataset [26] contains 70 difficult UAV image sequences, primarily addressing the problem of severe UAV motion. In addition, various cluttered scenes and objects with different sizes as well as aspect ratios are included. We compare our tracker with 12 state-of-the-art trackers, including trackers using hand-crafted features (*i.e.*, AutoTrack [6], BACF [8], DAT [28], DSST [29], ECO-HC [12], KCF [4], SRDCF [7], and STRCF [15]), using deep feature-based or pretrained deep architecture-based trackers (*i.e.*, ASRCF [11], IBCCF [30], UDT+ [31], and MDNet [32]). To make a fair comparison, the publicly available codes or results provided by the original authors are employed.

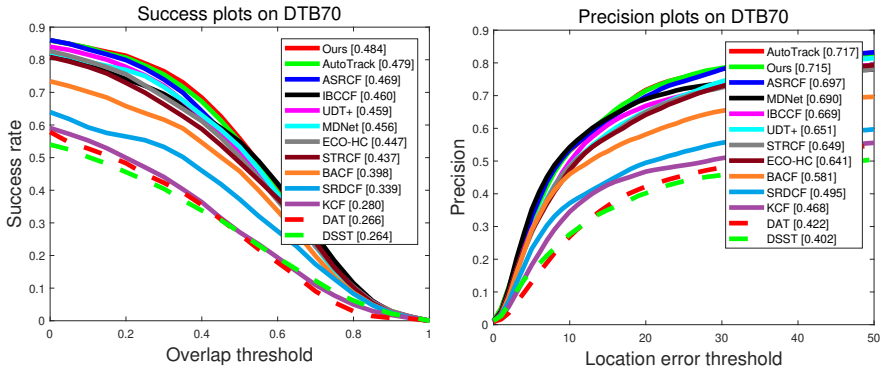

Fig. 2. Comparison of the success rate and precision plots with the state-of-the-art trackers on DTB70 dataset [26]. The numbers in the legend indicate the representative AUC for success plots and precisions at 20 pixels for precision plots.

We evaluate the trackers based on One Pass Evaluation (OPE) rule [33], and two measures are used for evaluation, namely success rate and precision. The success rate can display the percentage of situations when the overlap between the estimated bounding box and the ground truth is greater than different thresholds, and Area under the curve (AUC) is utilized for ranking. Precision can demonstrate the percentage of scenarios when the distance between the estimated bounding box and ground truth one is smaller than different thresholds, and the score at 20 pixels is used for ranking. Fig. 2 depicts both the success rate and precision of different trackers. Overall, the proposed tracker achieves the best result with an AUC score of 0.484 among all the other trackers. For the distance precision, the proposed ASTR-CF outperforms most of the competing trackers except for AutoTrack [6]. It is noteworthy that the proposed tracker surpasses its counterpart SRDCF [7] and STRCF [15] by 22% and 6.6%, respectively. What's more, only with hand-crafted features, our tracker outperforms deep feature-based trackers (ASRCF [11], and IBCCF [30]) and pre-trained deep architecture-based trackers (MDNet [32], and UDT+ [31]).

UAV123@10fps: The UAV123@10fps dataset [27] is a temporarily down-sampled version of the UAV123 [27]. It increases the tracing challenge compared with the original UAV123 [27] because the displacements of the moving objects become bigger. Nine state-of-the-art trackers, *i.e.*, AutoTrack [6], BACF [8], DSST [29], ECO-HC [12], MEEM [34], SRDCF [7], STRCF [15], Struck [35], MUSTer [36] are implemented for comparison. The comparative results are depicted in Fig. 3. One can see that the proposed tracker outperforms all the other state-of-the-art trackers in terms of both success rate and precision.

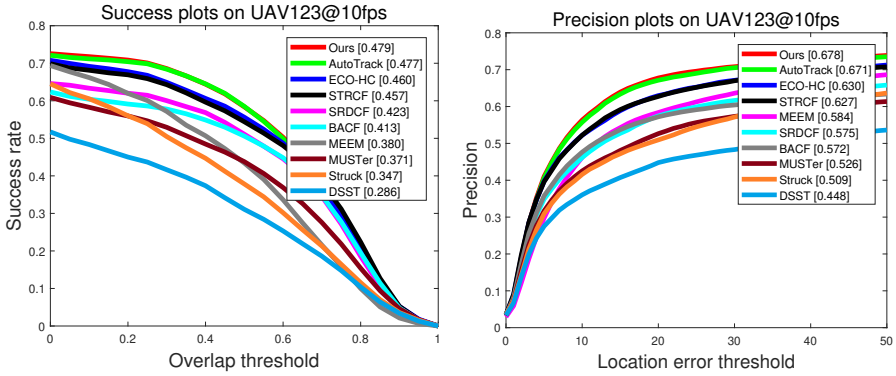

Fig. 3. Performance evaluation on UAV123@10fps dataset [27] in terms of success plots and precision plots.

Fig. 4 shows the overlap success plots of different trackers on 6 attributes, *e.g.*, illumination variation, partial occlusion, viewpoint changes, fast motion, scale variation, and low resolution. Our tracker achieves the best performance in all these attributes. This is mainly attributed to the proposed adaptive spatio-temporal regularization, in which the adaptive temporal regularization enables the learned filter to perform more robust to occlusion while adapting well to large appearance variation. Meanwhile, the learned filters focus on the reliable features of the tracked object, it can alleviate the effects of unexpected noises within the object region by introducing adaptive spatial regularization. *More attribute-based evaluations can be seen in the supplementary material.*

Finally, we perform qualitative evaluations of different trackers on several video sequences. For a clearer visualization, we exhibit the results of ASTR-CF and 4 state-of-the-art trackers, *i.e.*, AutoTrack [6], STRCF [15] ECO-HC [12], and BACF[8]. The tracking results on 6 video sequences are shown in Fig. 5. One can note that the proposed ASTR-CF performs favorably against the state-of-the-art hand-crafted trackers.

Overall performance evaluation: Average performance of the top-5 CPU-based trackers on DTB70 dataset [26] and UAV123@10fps dataset [27] are sum-

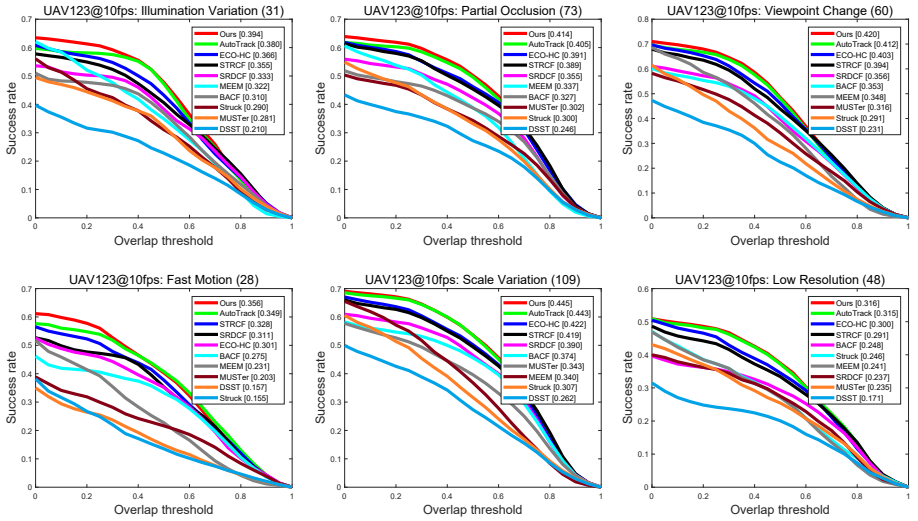

Fig. 4. Evaluation of different trackers with 6 attributes on the UAV123@10fps dataset [27]. Success plot can display the percentage of situations when the overlap between estimated bounding boxes and ground truth one is greater than different thresholds. Area under curve (AUC) is utilized for ranking.

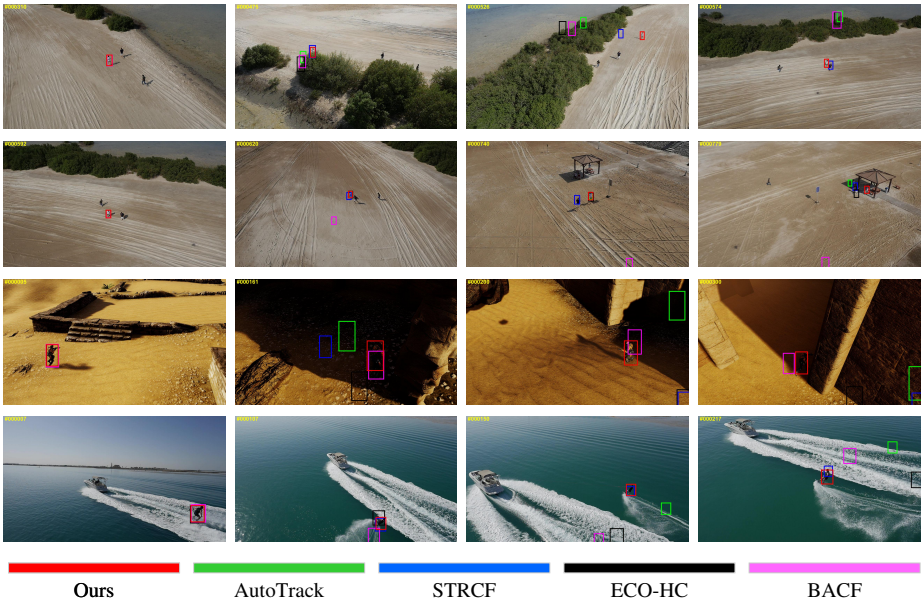

Fig. 5. Qualitative comparison of our approach with state-of-the-art trackers on the group2_2, group2_3, person1_s, and wakeboard5 sequences.

marized in Table 1. One can see that the proposed tracker performs the best in terms of both success rate and precision. Meanwhile, it has a second fast computational speed of 55.5fps, only slower than AutoTrack (56.5fps). However, it is two times faster than STRCF (25.3fps). This is attributed to the adaptive temporal regularization which can reduce meaningless and detrimental training on contaminated samples. *More detailed results can be found in supplementary materials.*

Table 1. Average accuracy and computational speed comparisons of top-5 CPU-based trackers on DTB70 [26] and UAV123@10fps [27]. The best three results are shown in red, green, and blue fonts, respectively.

Tracker	Ours	AutoTrack	ECO-HC	STRCF	BACF
Success rate	0.481	0.478	0.455	0.450	0.408
Precision	0.691	0.687	0.634	0.635	0.575
FPS	55.5	56.5	47.7	25.3	48.1

4.2 Ablation study

To validate the effectiveness, our tracker is compared to itself with different modules enabled. The overall evaluation is presented in Table 2. With ASR (adaptive spatial regularization) module and ATR (adaptive temporal regularization) module being added to the baseline (STRCF [15]), the performance is improved smoothly. Besides, our final tracker improves the baseline method by 4.7% and 6.6% in terms of success rate and precision criterion, respectively.

.

Table 2. Ablation analysis on the DTB70 dataset.

Tracker	Ours	Baseline+ATR	Baseline+ASR	Baseline
Success rate	0.484	0.480	0.476	0.437
Precision	0.715	0.701	0.704	0.649

5 Conclusion

In this study, a novel Adaptive Spatio-Temporal Regularized Correlation Filter (ASTR-CF) is proposed to solve the problems of boundary effect and filter degradation in the application of UAV-based tracking. An alternating direction method of multipliers (ADMM) algorithm is developed to solve the ASTR-CF

model efficiently. Comparative experiments are carried out to validate the ASTR-CF model. Experimental results demonstrate that the proposed ASTR-CF outperforms the state-of-the-art trackers in terms of both accuracy and computational speed.

References

1. Zhang, W., Song, K., Rong, X., Li, Y.: Coarse-to-fine uav target tracking with deep reinforcement learning. IEEE Transactions on Automation Science and Engineering **16** (2019) 1522–1530
2. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Discriminative scale space tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2016) 1561–1575
3. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: CVPR. (2010)
4. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015) 583–596
5. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.: Staple: Complementary learners for real-time tracking. In: CVPR. (2016)
6. Li, Y., Fu, C., Ding, F., Huang, Z., Lu, G.: Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In: CVPR. (2020)
7. Danelljan, M., Hager, G., Khan, F.S., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: ICCV. (2015)
8. Galoogahi, H.K., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: ICCV. (2017)
9. Zhou, Y., Han, J., Yang, F., Zhang, K., Hong, R.: Efficient correlation tracking via center-biased spatial regularization. IEEE Transactions on Image Processing **27** (2018) 6159–6173
10. Guo, Q., Han, R., Feng, W., Chen, Z., Wan, L.: Selective spatial regularization by reinforcement learned decision making for object tracking. IEEE Transactions on Image Processing **29** (2020) 2999–3013
11. Dai, K., Wang, D., Lu, H., Sun, C., Li, J.: Visual tracking via adaptive spatially-regularized correlation filters. In: CVPR. (2019)
12. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: CVPR. (2017)
13. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In: CVPR. (2016)
14. Lukezic, A., Vojir, T., CehovinZajc, L., Matas, J., Kristan, M.: Discriminative correlation filter tracker with channel and spatial reliability. International Journal of Computer Vision **126** (2018) 671–688
15. Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.: Learning spatial-temporal regularized correlation filters for visual tracking. In: CVPR. (2018)
16. Li, Y., Fu, C., Huang, Z., Zhang, Y., Pan, J.: Keyfilter-aware real-time uav object tracking. arXiv preprint arXiv:2003.05218 (2020)
17. Wang, M., Liu, Y., Huang, Z.: Large margin object tracking with circulant feature maps. CVPR (2017)

18. Fu, C., Huang, Z., Li, Y., Duan, R., Lu, P.: Boundary effect-aware visual tracking for uav with online enhanced background learning and multi-frame consensus verification. arXiv preprint arXiv:1908.03701 (2019)
19. Sun, Y., Sun, C., Wang, D., He, Y., Lu, H.: Roi pooled correlation filters for visual tracking. CVPR (2019)
20. Gu, X., Xu, X.: Accurate mask-based spatially regularized correlation filter for visual tracking. Journal of Electronic Imaging **26** (2017) 013002–013002
21. Kang, B., Chen, G., Zhou, Q., Yan, J., Lin, M.: Visual tracking via multi-layer factorized correlation filter. IEEE Signal Processing Letters **26** (2019) 1763–1767
22. Sun, C., Wang, D., Lu, H., Yang, M.H.: Learning spatial-aware regressions for visual tracking. CVPR (2018)
23. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
24. Danelljan, M., Khan, F.S., Felsberg, M., v. d. Weijer, J.: Adaptive color attributes for real-time visual tracking. In: CVPR. (2014)
25. Boyd, S., Parikh, N., Chu, E.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc (2011)
26. Li, S., Yeung, D.Y.: Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In: AAAI. (2017)
27. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: ECCV. (2016)
28. Possegger, H., Mauthner, T., Bischof, H.: In defense of color-based model-free tracking. In: CVPR. (2015)
29. Danelljan, M., Häger, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: BMVC. (2014)
30. Li, F., Yao, Y., Li, P., Zhang, D., Zuo, W., Yang, M.: Integrating boundary and center correlation filters for visual tracking with aspect ratio variation. In: ICCVW. (2017)
31. Wang, N., Song, Y., Ma, C., Zhou, W., Liu, W., Li, H.: Unsupervised deep tracking. In: CVPR. (2019)
32. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: CVPR. (2016)
33. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015) 1834–1848
34. Zhang, J., Ma, S., Sclaroff, S.: Meem: Robust tracking via multiple experts using entropy minimization. In: ECCV. (2014)
35. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: ICCV. (2011)
36. Hong, Z., Zhe Chen, Wang, C., Mei, X., Prokhorov, D., Tao, D.: Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In: CVPR. (2015)