Learning Soft-Consistent Correlation Filters for RGB-T Object Tracking

Yulong Wang[1], Chenglong Li[1,2(✉)], and Jin Tang[1]

[1] School of Computer Science and Technology, Anhui University, Hefei, China
wylemail@qq.com, lcl1314@foxmail.com, tj@ahu.edu.cn
[2] Center for Research on Intelligent Perception and Computing,
NLPR, CASIA, Beijing, China

Abstract. To track objects efficiently and effectively in adverse illumination conditions even in dark environment, this paper presents a novel soft-consistent correlation filters (SCCF) using RGB and thermal infrared (RGB-T) data for visual tracking. The proposed SCCF uses soft consistency to take both collaboration and heterogeneity into account for joint learning of the correlation filters of RGB and thermal spectra, while the computational time is reduced significantly by employing the Fast Fourier Transform (FFT). Moreover, a novel weighted fusion mechanism is proposed to compute the final response map in the detection phase. Extensive experiments on the benchmark dataset show that the proposed approach performs favorably against state-of-the-art methods, while runs at 50 frames per second.

Keywords: Visual tracking · RGB and thermal fusion
Correlation filter · Soft-consistent

1 Introduction

Visual tracking is an active research area in the computer vision community, since it is an essential and significant task in various applications, such as visual surveillance, robotics, human-computer interaction, and self-driving systems, to name a few [8,21,22]. Despite of many breakthroughs recently [16,23,29], the visual tracking mainly relies on traditional RGB sensors and tracks target objects in case of cluttered background and low visibility at night and in bad weather, and is thus still regarded as a challenging problem.

The adoption of thermal infrared sensors has provided new opportunities to advance the state-of-the-art trackers by handle the aforementioned challenges [13,15,17–20,26]. However, how to perform efficient and effective fusion of different modalities for boosting tracking performance is an open issue.

In recent years, many methods [13,18–20,26] have been proposed to fuse different spectra for improving tracking performance. Some trackers [13,20,26] focus on the sparse representation in Bayesian filtering framework because of its capability of suppressing noises and errors. Some trackers [13,19] introduce

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11259, pp. 295–306, 2018.
https://doi.org/10.1007/978-3-030-03341-5_25

spectral weights to fuse RGB and thermal information. Despite all these significant progress, these methods [13,19] still have some limitations. These methods only consider the collaboration of different source data. However, different spectra are usually heterogeneous (e.g., RGB and thermal), and thus direct fusion that only employs the collaboration might be ineffective. On the other hand, the method [13] based on collaborative sparse representation in Bayesian filtering framework is time-consuming. However, most applications demand real-time tracking.

To deal with these issues, we present a novel multi-spectral approach based on correlation filters [10] to perform efficient object tracking. Specifically, we propose a novel scheme to deploy the inter-spectral information by imposing soft consistency in the correlation filters. Our method take both the collaboration and the heterogeneity of different spectral information into account for more effective fusion. For the collaboration, we observe that the learned filters should select similar circular shifts such that they have similar motion. While for the heterogeneity, we intend to allow filters have sparse different elements to each other. Moreover, we design a novel mechanism to fuse RGB and thermal information for robust visual tracking. We calculate the spectral weights according to the response map in the detection phase, and the final response map is obtained by weighted fusion of each spectral response map.

We validate the effectiveness and efficiency of the proposed method on the benchmark dataset, i.e., GTOT [13], and the results show that our approach achieves big superiority in terms of accuracy and comparable performance in terms of efficiency.

To summarize, the main contributions of this work are three-fold.

- A novel soft-consistent correlation filters for RGB-T object tracking is proposed. In order to take both collaboration and the heterogeneity of RGB and thermal spectra into account, the correlation filters of multi-spectral are learned jointly by imposing soft consistency. And the computational time is reduced significantly by employing the Fast Fourier Transform (FFT).
- A spectral fusion mechanism is designed. The spectral weights are obtained according to the response map in the detection phase, and the final response map is obtained by weighted fusion of different spectra.
- It performs favorably against a number of state-of-the-art trackers with the running speed over 50 frames per second. To facilitate further studies, our source code will be made available to the public.

2 Related Work

We review the related work to us from two research streams, i.e., RGB-T object tracking and Correlation filter tracking.

2.1 RGB-T Object Tracking

RGB-T object tracking has drawn a lot of attentions in the computer vision community with the popularity of thermal infrared sensors [3,13,14,18–20,26].

Cvejic et al. [3] investigate the impact of pixel-level fusion of videos from RGB-T surveillance cameras, and accomplish their tracker by means of a particle filter with the fusion of a color cue and the structural similarity measure. Wu et al. [26] and Liu and Sun [20] directly employ the sparse representation to calculate the likelihood score using reconstruction residues or coefficients in Bayesian filtering framework. They ignore modality reliabilities in fusion, which may limit the tracking performance when facing malfunction or occasional perturbation of individual sources. Li et al. [13] and Li et al. [19] introduce modality weights to handle this problem, and propose sparse representation based algorithms to fuse RGB and thermal information. Different from these methods, we take both collaboration and the heterogeneity of RGB and thermal spectrums into account by imposing soft consistency in the correlation filter tracking framework to perform efficient and effective multispectral tracking.

2.2 Correlation Filter Tracking

Correlation filters have achieved great breakthroughs in visual tracking due to its accuracy and computational efficiency [1,4–7,10,11,29]. Bolme et al. [1] first introduce correlation filters into visual tracking, named MOSSE, and achieve hundreds of frames per second, and high tracking accuracy. Recently, many researchers further improve MOSSE from different aspects. For example, Henriques et al. [10,11] extend MOSSE to non-linear one with kernel trick, and incorporate multiple channel features efficiently by summing all channels in kernel space. To handle scale variations, Danelljan et al. [4] learn correlation filters for translation and scale estimation separately by using a scale pyramid representation. Dong et al. [7] propose a sparse correlation filter for combining the robustness of sparse representation and the efficiency of correlation filter. Zhang et al. [29] integrate multiple parts and multiple features into a unified correlation particle filter framework to perform effective object tracking.

3 Proposed SCCF Tracker

In this section, we first present the technical details of the proposed algorithm and then describe the optimization process of the model.

3.1 SCCF Formulation

For a typical correlation filter, many negative samples are used to improve the discriminability of the track-by-detector scheme. In this work, denote $\mathbf{x}_k$ as the feature vector of $M \times N \times D$ of k-th spectrum, where M, N, and D indicates the width, height, and the number of channels, respectively. We consider all the circular shifts of $\mathbf{x}_k$ along the M and N dimensions as training samples of k-th spectrum. Each shifted sample $\mathbf{x}^k_{m,n}$, $(m,n) \in \{0,1,...,M-1\} \times \{0,1,...,N-1\}$, has a Gaussian function label $y(m,n) = e^{-\frac{(m-M/2)^2+(n-N/2)^2}{2\sigma^2}}$, where σ is the kernel width. Let $\mathbf{X}_k = [\mathbf{x}_{0,0}, ..., \mathbf{x}_{m,n}, ...\mathbf{x}_{M-1,N-1}]^{\mathrm{T}}$ denote all training samples of the k-th spectrum $(k = 1, ..., K)$. The purpose is to find the optimal correlation filters $\mathbf{w}_k$ for K different spectra,

$$\min_{\mathbf{w}_k} \sum_{k=1}^{K} \frac{1}{2} ||\mathbf{X}_k \mathbf{w}_k - \mathbf{y}||_2^2 + \lambda_1 ||\mathbf{w}_k||_2^2, \tag{1}$$

where λ_1 is a regularization parameter. The objective function (1) can equivalently be expressed in its dual form,

$$\min_{\mathbf{z}_k} \sum_{k=1}^{K} \frac{1}{4\lambda_1} \mathbf{z}_k^{\mathrm{T}} \mathbf{G}_k \mathbf{z}_k + \frac{1}{4} \mathbf{z}_k^{\mathrm{T}} \mathbf{z}_k - \mathbf{z}_k^{\mathrm{T}} \mathbf{y}. \tag{2}$$

Here, the vector $\mathbf{z}_k$ contains $M \times N$ dual optimization variables $\mathbf{z}_{m,n}^k$, and $\mathbf{G}_k = \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}}$. The two solutions are related by $\mathbf{w}_k = \frac{\mathbf{X}_k^{\mathrm{T}} \mathbf{z}_k}{2\lambda_1}$. The discriminative training samples $\mathbf{x}_{m,n}^k$ are selected by the learned $\mathbf{z}_{m,n}^k$ to distinguish the target object from the background. Obviously, the training samples $\mathbf{x}_{m,n}^k$, $(m, n) \in \{0, 1, ..., M-1\} \times \{0, 1, ..., N-1\}$ are the all possible circular shifts, which denote the possible locations of the target object.

Most of existing works only consider the collaboration of different source data [13,19]. However, different spectra are usually heterogeneous (e.g., RGB and thermal), and thus direct fusion that only employs the collaboration might be ineffective. Therefore, in this paper, we propose a novel scheme to take both the collaboration and the heterogeneity of different spectral information into account for more effective fusion. For the collaboration, we observe that the learned $\{\mathbf{z}_k\}$ should select similar circular shifts such that they have similar motion. While for the heterogeneity, we intend to allow $\{\mathbf{z}_k\}$ have sparse different elements to each other. Taking the above considerations together, we propose a soft-consistent constraint on $\{\mathbf{z}_k\}$ that makes them consistent while allowing the sparse inconsistency exists, and formulated as a l_1-optimization based sparse learning problem. Finally, we obtain the soft-consistent correlation filter(SCCF) for multi-spectral tracking as

$$\min_{\mathbf{z}_k} \sum_{k=1}^{K} \frac{1}{4\lambda_1} \mathbf{z}_k^{\mathrm{T}} \mathbf{G}_k \mathbf{z}_k + \frac{1}{4} \mathbf{z}_k^{\mathrm{T}} \mathbf{z}_k - \mathbf{z}_k^{\mathrm{T}} \mathbf{y} + \lambda_2 \sum_{k=2}^{K} ||\mathbf{z}_k - \mathbf{z}_{k-1}||_1, \tag{3}$$

where λ_1 and λ_2 are regularization parameters.

3.2 Optimization Algorithm

In this section, we present algorithmic details on how to efficiently solve the optimization problem (3). Two auxiliary variables $\mathbf{P}$ and $\mathbf{q}_k$ are introduced to make Eq. (3) separable:

$$\begin{aligned} \min_{\mathbf{z}_k, \mathbf{P}, \mathbf{q}_k} \sum_{k=1}^{K} \frac{1}{4\lambda_1} \mathbf{q}_k^{\mathrm{T}} \mathbf{G}_k \mathbf{q}_k + \frac{1}{4} \mathbf{q}_k^{\mathrm{T}} \mathbf{q}_k - \mathbf{q}_k^{\mathrm{T}} \mathbf{y} + \lambda_2 ||\mathbf{P}||_1 \\ s.t. \mathbf{P} = \mathbf{C}\mathbf{Z}, \mathbf{z}_k = \mathbf{q}_k, \end{aligned} \tag{4}$$

where $\mathbf{Z} = [\mathbf{z}_1; \mathbf{z}_2; ...; \mathbf{z}_K]$, $\mathbf{C}$ is the consistency matrix, which is defined as: $\mathbf{C} = \begin{bmatrix} -\mathbf{I}^1 & \mathbf{I}^2 & & & \\ & -\mathbf{I}^2 & \mathbf{I}^3 & & \\ & & ... & ... & \\ & & & -\mathbf{I}^{K-1} & \mathbf{I}^K \end{bmatrix}$. $\mathbf{I}$ is the identity matrix.

We use the fast first-order Alternating Direction Method of Multipliers (ADMM) to efficiently solve the optimization problem (4). By introducing augmented Lagrange multipliers to incorporate the equality constraints into the objective function, we obtain a Lagrangian function that can be optimized through a sequence of simple closed form update operations in (5).

$$\begin{aligned} \min_{\mathbf{z}_k, \mathbf{P}, \mathbf{q}_k} & \sum_{k=1}^{K} \frac{1}{4\lambda_1} \mathbf{q}_k^{\mathrm{T}} \mathbf{G}_k \mathbf{q}_k + \frac{1}{4} \mathbf{q}_k^{\mathrm{T}} \mathbf{q}_k - \mathbf{q}_k^{\mathrm{T}} \mathbf{y} + \langle \mathbf{Y}_{2,k}, \mathbf{q}_k - \mathbf{z}_k \rangle + \frac{\mu}{2} ||\mathbf{q}_k - \mathbf{z}_k||_2^2 \\ & + \lambda_2 ||\mathbf{P}||_1 + \langle \mathbf{Y}_1, \mathbf{P} - \mathbf{C}\mathbf{Z} \rangle + \frac{\mu}{2} ||\mathbf{P} - \mathbf{C}\mathbf{Z}||_F^2 \\ = & \sum_{k=1}^{K} \frac{1}{4\lambda_1} \mathbf{q}_k^{\mathrm{T}} \mathbf{G}_k \mathbf{q}_k + \frac{1}{4} \mathbf{q}_k^{\mathrm{T}} \mathbf{q}_k - \mathbf{q}_k^{\mathrm{T}} \mathbf{y} + \frac{\mu}{2} ||\mathbf{q}_k - \mathbf{z}_k + \frac{\mathbf{Y}_{2,k}}{\mu}||_2^2 - \frac{1}{2\mu} ||\mathbf{Y}_{2,k}||_2^2 \\ & + \lambda_2 ||\mathbf{P}||_1 + \frac{\mu}{2} ||\mathbf{P} - \mathbf{C}\mathbf{Z} + \frac{\mathbf{Y}_1}{\mu}||_F^2 - \frac{1}{2\mu} ||\mathbf{Y}_1||_F^2 \end{aligned} \tag{5}$$

Here, $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^{\mathrm{T}} \mathbf{B})$ denotes the matrix inner product. $\mathbf{Y}_1$ and $\mathbf{Y}_{2,k}$ are Lagrangian multipliers. We then alternatively update one variable by minimizing (5) with fixing other variables. Besides the Lagrangian multipliers, there are three variables, including $\mathbf{q}_k$, $\mathbf{Z}$ and $\mathbf{P}$, to solve. The solutions of the subproblems are as follows:

q-subproblem. Given fixed $\mathbf{P}$ and $\mathbf{Z}$, $\mathbf{q}_k$ is updated by solving the optimization problem (6) with the solution (7)

$$\min_{\mathbf{q}_k} \sum_{k=1}^{K} \frac{1}{4\lambda_1} \mathbf{q}_k^{\mathrm{T}} \mathbf{G}_k \mathbf{q}_k + \frac{1}{4} \mathbf{q}_k^{\mathrm{T}} \mathbf{q}_k - \mathbf{q}_k^{\mathrm{T}} \mathbf{y} + \frac{\mu}{2} ||\mathbf{q}_k - \mathbf{z}_k + \frac{\mathbf{Y}_{2,k}}{\mu}||_2^2, \tag{6}$$

$$\mathbf{q}_k = (\frac{1}{2\lambda_1} \mathbf{G}_k + \frac{1}{2} \mathbf{I} + \mu \mathbf{I})^{-1} (\mathbf{y} + \mu \mathbf{z}_k - \mathbf{Y}_{2,k}). \tag{7}$$

Here, $\mathbf{G}_k = \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}}$. $\mathbf{I}$ is an identity matrix. Note that, all circulant matrices are made diagonal by the Discrete Fourier Transform (DFT), regardless of the generating vector. If $\mathbf{X}_k$ is a circulant matrix, it can be expressed with its base sample $\mathbf{x}_k$ as

$$\mathbf{X}_k = F diag(\hat{\mathbf{x}}_k) F^{\mathrm{H}}, \tag{8}$$

where $\hat{\mathbf{x}}_k$ denotes the DFT of the generating vector, $\hat{\mathbf{x}}_k = \mathcal{F}(\mathbf{x}_k)$, and F is a constant matrix that does not depend on $\mathbf{x}_k$. The constant matrix F is known as the DFT matrix. $\mathbf{X}_k^{\mathrm{H}}$ is the Hermitian transpose, i.e., $\mathbf{X}_k^{\mathrm{H}} = (\mathbf{X}_k^*)^{\mathrm{T}}$, and $\mathbf{X}_k^*$ is the complex-conjugate of $\mathbf{X}_k$. For real numbers, $\mathbf{X}_k^{\mathrm{H}} = \mathbf{X}_k^{\mathrm{T}}$. It (Eq. (7)) can

Algorithm 1. Optimization Procedure to Eq. (4).

```
Input: The spectra feature matrix $\mathbf{X}_k (k = 1, 2..., K)$ and Gaussian function label $\mathbf{y}$,
  the parameters $\lambda_1$ and $\lambda_2$;
  Set $\mathbf{q}_k = \mathbf{Y}_{2,k} = 0, \mathbf{P} = \mathbf{Y}_1 = 0, \mathbf{Z} = 0, \mu_0 = 0.1, \mu_{max} = 10^{10}, \rho = 1.2, \epsilon = 10^{-15}, maxIter = 10$ and $t = 0$.
Output: The filter $\mathbf{z}_k$.
  while not converged do
    Update $\mathbf{q}_{k,t+1}$ by Eq. (9);
    Update $\mathbf{P}_{t+1}$ by Eq. (11);
    Update $\mathbf{Z}_{t+1}$ by Eq. (13);
    Update Lagrange multipliers as follows:
      $\mathbf{Y}_{1,t+1} = \mathbf{Y}_{1,t} + \mu_t(\mathbf{P} - \mathbf{CZ})$;
      $\mathbf{Y}_{2,k,t+1} = \mathbf{Y}_{2,k,t} + \mu_t(\mathbf{q}_k - \mathbf{z}_k)$;
    Update $\mu_{t+1}$ by $\mu_{t+1} = \min\ (\mu_{max}, \rho\mu_t)$;
    Update $t$ by $t = t + 1$;
    Check the convergence condition, i.e. the maximum element changes of $\mathbf{q}_k, \mathbf{P}$ and
    $\mathbf{Z}$ between two consecutive iterations are less than $\epsilon$ or the maximum number of
    iterations reaches maxIter.
  end while
```

be calculated very efficiently in the Fourier domain by considering the circulant structure property of $\mathbf{X}_k$,

$$\mathbf{q}_k = \mathcal{F}^{-1}[\frac{2\lambda_1(\hat{\mathbf{y}} + \mu\hat{\mathbf{z}}_k - \hat{\mathbf{Y}}_{2,k})}{\hat{\mathbf{x}}_k^* \odot \hat{\mathbf{x}}_k + \lambda_1 + 2\lambda_1\mu}]. \tag{9}$$

Here, $\mathcal{F}^{-1}$ denotes the inverse DFT, while $\odot$ as well as the fraction denote the element-wise product and division, respectively. The $\mathbf{x}_k$ is the base sample of circulant matrix $\mathbf{X}_k$.

P-subproblem. Given fixed $\mathbf{Z}$ and $\mathbf{q}_k$, Eq. (5) can be rewritten as

$$\min_{\mathbf{P}} \lambda_2||\mathbf{P}||_1 + \frac{\mu}{2}||\mathbf{P} - \mathbf{CZ} + \frac{\mathbf{Y}_1}{\mu}||_F^2. \tag{10}$$

According to (Lin et al. 2009), an efficient closed-form solution can be computed by the soft-thresholding (or shrinkage) method:

$$\mathbf{P} = S_{\frac{\lambda_2}{\mu}}(\mathbf{CZ} - \frac{\mathbf{Y}_1}{\mu}), \tag{11}$$

where the definition of $S_\lambda(a)$ is $S_\lambda(a) = \mathrm{sign}(a)\max(0, |a| - \lambda)$.

Z-subproblem. Given fixed $\mathbf{q}_k$ and $\mathbf{P}$, Eq. (5) can be rewritten as

$$\min_{\mathbf{Z}} \frac{\mu}{2}(||\mathbf{P} - \mathbf{CZ} + \frac{\mathbf{Y}_1}{\mu}||_F^2 + ||\mathbf{Q} - \mathbf{Z} + \frac{\mathbf{Y}_2}{\mu}||_F^2). \tag{12}$$

where $\mathbf{Q} = [\mathbf{q}_1; \mathbf{q}_2; ...; \mathbf{q}_K]$. The solution of Eq. (12) is:

$$\mathbf{Z} = (\mu\mathbf{C}^{\mathrm{T}}\mathbf{C} + \mu\mathbf{I})^{-1}(\mu\mathbf{C}^{\mathrm{T}}\mathbf{P} + \mathbf{C}^{\mathrm{T}}\mathbf{Y}_1 + \mu\mathbf{Q} + \mathbf{Y}_2) \tag{13}$$

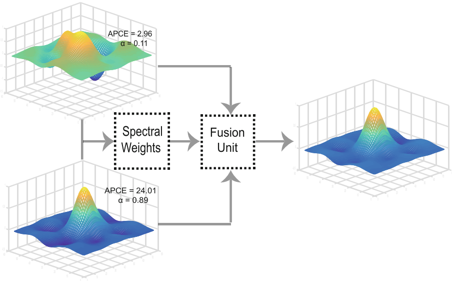

Fig. 1. Pipeline of the proposed spectral fusion mechanism. The spectral weights are obtained according to the response map in the detection phase, and the final response map is obtained by weighted fusion of different spectra response maps.

Since each subproblem of Eq. (4) is convex, we can guarantee that the limit point by our algorithm satisfies the Nash equilibrium conditions [27]. And the main steps of the optimization procedure are summarized in Algorithm 1.

3.3 Tracking

Target Position Estimation. After solving this optimization problem, we obtain the correlation filter $\mathbf{z}_k$ for each type of spectrum. Given an image patch in the next frame, the feature vector on the k-th spectrum is denoted by $\mathbf{s}_k$ and of size $M \times N \times D$. We first transform it to the Fourier domain $\hat{\mathbf{s}}_k = \mathcal{F}(\mathbf{s}_k)$, and then the k-th correlation response map is computed by

$$\mathbf{R}_k = \mathcal{F}^{-1}(\hat{\mathbf{s}}_k \odot \hat{\mathbf{x}}_k^* \odot \hat{\mathbf{z}}_k). \tag{14}$$

Some existed trackers [13] learn spectral weights in a single unified algorithm. Actually, this may increase the complexity of the proposed model. In this work, we use a novel criterion called average peak-to-correlation energy ($APCE$) measure, as proposed in [25], to calculate the priori influence factor. The definition of $APCE$ is

$$APCE = \frac{|R_{max} - R_{min}|^2}{mean(\sum_{m,n}(R_{m,n} - R_{min})^2)}, \tag{15}$$

where R_{max}, R_{min} and $R_{m,n}$ denote the maximum,minimum and the m-th row n-th column entry of the response map $\mathbf{R}$, respectively. $APCE$ indicates the degree of fluctuation of the response maps. For sharper peaks and fewer noise, i.e., the target apparently appearing in the detection region, $APCE$ will become larger and the response map will become smooth except for only one sharp peak. Otherwise, $APCE$ will significantly decrease if the response map is multi-peaks. Based on the nature of the $APCE$, we design a new method to calculate the weights of different spectra as follow:

$$\alpha_k = \frac{APCE_k}{\sum_{k=1}^{K} APCE_k}, \tag{16}$$

where $APCE_k$ denotes the value of $APCE$ of the k-th spectrum. As illustrated in Fig. 1, the weight of reliable spectrum is larger than unreliable spectrum because the $APCE$ of reliable spectrum is much larger than unreliable spectrum. Then the final correlation response map is computed by

$$\mathbf{R} = \sum_{k=1}^{K} \alpha_k \mathbf{R}_k. \tag{17}$$

The target location then can be estimated by searching for the position of maximum value of the correlation response map $\mathbf{R}$ of size $M \times N$.

Model Update. Similar to other CF trackers [10,23,24,29]. To improve our robustness to pose, scale and illumination changes, we adopt an incremental strategy, which only uses new samples $\mathbf{x}_k$ in the current frame to update models as shown in (18), where t is the frame index and η is a learning rate parameter.

$$\begin{aligned} \mathcal{F}(\mathbf{x}_k^t) &= (1-\eta)\mathcal{F}(\mathbf{x}_k^{t-1}) + \eta\mathcal{F}(\mathbf{x}_k^t), \\ \mathcal{F}(\mathbf{z}_k^t) &= (1-\eta)\mathcal{F}(\mathbf{z}_k^{t-1}) + \eta\mathcal{F}(\mathbf{z}_k^t). \end{aligned} \tag{18}$$

4 Experiments

In this section, we present extensive experimental evaluations on the proposed soft-consistent correlation filters (SCCF) tracker. We first introduce the experimental setups, and then extensive experiments are conducted to evaluate the SCCF tracker against plenty of state-of-the-art trackers on GTOT benchmark.

4.1 Experimental Setups

Implementation Details. We set the regularization parameters of (3) to λ_1 = 0.038 and λ_2 = 0.012, and use a kernel width of 0.1 for generating the Gaussian function labels. Their learning rate η in (18) is set to 0.025. To remove the boundary discontinuities, the extracted feature are weighted by a cosine window. In addition, we utilize an adaptive multi-scale strategy to adapt to the scale variations. We implement our tracker in MATLAB on an Intel I7-6700K 4.00 GHz CPU with 32 GB RAM. Furthermore, all the parameter settings are available in the source code to be released for accessible reproducible research.

Dataset. Our algorithm is evaluated on a large visual tracking benchmark dataset: GTOT [13]. GTOT includes 50 aligned RGB-T video pairs with about 12 K frames in total. They are annotated with ground truth bounding boxes and various visual attributes.

Evaluation Protocol. All trackers are evaluated according to widely used metrics, precision rate (PR) and success rate (SR), as defined in GTOT [13]. PR is the percentage of frames whose output location is within the given threshold distance of ground truth. SR is the ratio of the number of successful frames whose overlap is larger than a threshold. By changing the threshold, the SR plot can be obtained, and we employ the area under curve of SR plot to define the representative SR.

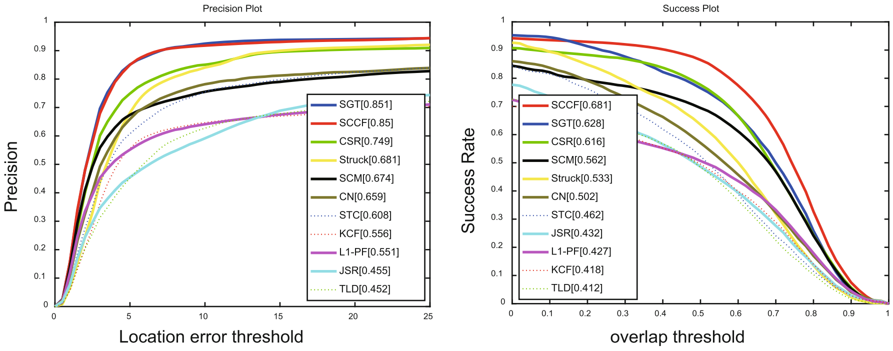

Fig. 2. The evaluation results on public GTOT benchmark. The representative score of PR/SR is presented in the legend.

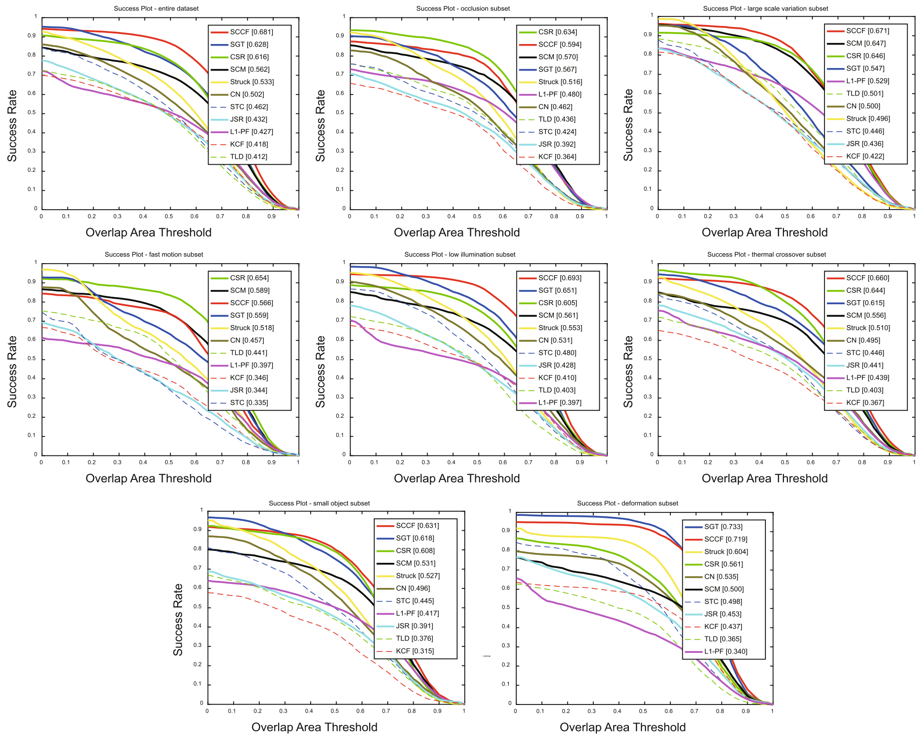

Fig. 3. Attribute-based evaluation on 50 sequences. We also put the overall performance here (the first one) for comparison convenience facing a single challenge and their combination.

4.2 Performance Evaluation

We evaluate our SCCF algorithm with 10 trackers on GTOT, including CSR [13], SGT [19], Struck [9], SCM [30], CN [6], STC [28], KCF [10], L1-PF [26], JSR [20] and TLD [12].

Quantitative Evaluation. As shown in Fig. 2, we report the PR/SR score for each tracker in the figure legend. Among all the trackers, our SCCF method occupies the best one in terms of SR. Compared with CSR, SCCF achieves about 6.5% improvement with SR. Furthermore, compared with SGT, SCCF achieves much better performance with about 5.3% improvement. Although SGT tracker performs the best against the other trackers in PR score, its model is more complex than ours. Moreover, the proposed tracker performs at about 50 FPS (frames per second) which is much faster than SGT (about 5 FPS).

Attribute-Based Evaluation. We further analyze the robustness of the proposed tracker performance in various scenes (e.g., thermal crossover, low illumination, fast motion) annotated in the benchmark. Our tracker performs well against other methods in most tracking challenges as shown in Fig. 3. In particular, SCCF outperforms other methods by a huge margin in handling low illumination and thermal crossover, which can be attributed to the use of soft consistency. However, our method does not perform as well in the presence of occlusion and deformation, as SCCF does not adopt a delayed update strategy [2,25] in order to reduce the computational load.

More qualitative results are given in Fig. 4.

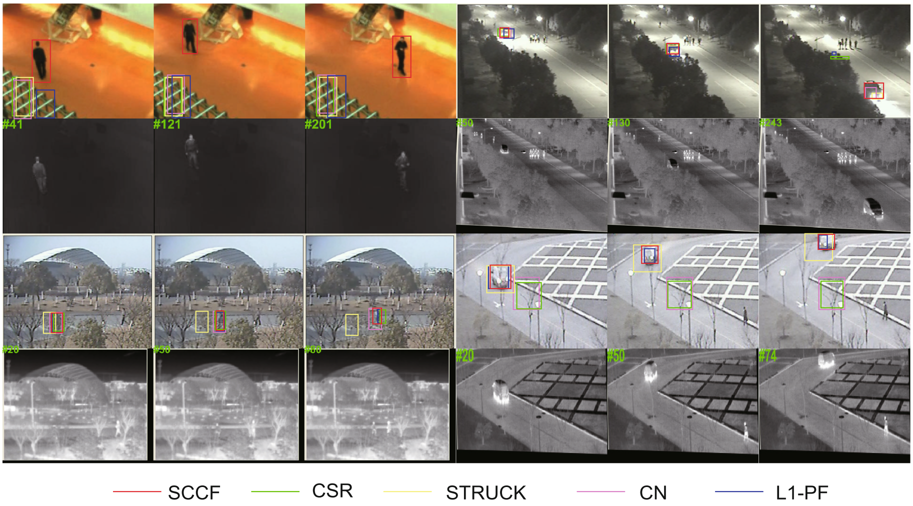

Fig. 4. Sample results of our method against other tracking methods, including L1-PF, CSR, Struck, and CN.

5 Conclusion

In this paper, we propose a novel learning soft-consistent correlation filters for RGB-T object tracking. The proposed tracking algorithm can effectively exploit collaboration and heterogeneity among different spectra to learn their correlation filters jointly. Moreover, we design a novel mechanism to fuse RGB and thermal information for robust visual tracking. Experimental results compared with several state-of-the-art methods on visual tracking benchmark demonstrate the effectiveness and robustness of the proposed algorithm. In the future, we will investigate the performance of multi-channel features (such as HOG) and design a new algorithm based on this work to calculate the correlation filters and spectral weights simultaneously.

Acknowledgment. This work is jointly supported by National Natural Science Foundation of China (61702002, 61472002), China Postdoctoral Science Foundation, Natural Science Foundation of Anhui Province (1808085QF187), Natural Science Foundation of Anhui Higher Education Institution of China (KJ2017A017), and Co-Innovation Center for Information Supply & Assurance Technology, Anhui University.

References

1. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: CVPR, pp. 2544–2550 (2010)
2. Choi, J., Chang, H.J., Yun, S., Fischer, T., Demiris, Y., Jin, Y.C.: Attentional correlation filter network for adaptive visual tracking. In: IEEE Conference on CVPR, pp. 4828–4837 (2017)
3. Cvejic, N., et al.: The effect of pixel-level fusion on object tracking in multi-sensor surveillance video. In: Proceedings of IEEE Conference on CVPR (2007)
4. Danelljan, M., Häger, G., Khan, F.S.: Accurate scale estimation for robust visual tracking. In: BMVC, pp. 65.1–65.11 (2014)
5. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ECO: efficient convolution operators for tracking. In: Proceedings of IEEE Conference on CVPR (2017)
6. Danelljan, M., Khan, F.S., Felsberg, M., Weijer, J.V.D.: Adaptive color attributes for real-time visual tracking. In: Proceedings of IEEE Conference on CVPR, pp. 1090–1097 (2014)
7. Dong, Y., Yang, M., Pei, M.: Visual tracking with sparse correlation filters. In: IEEE ICIP, pp. 439–443 (2016)
8. Emami, A., Dadgostar, F., Bigdeli, A., Lovell, B.C.: Role of spatiotemporal oriented energy features for robust visual tracking in video surveillance. In: IEEE Conference on AVSS, pp. 349–354 (2012)
9. Hare, S., Saffari, A., Torr, P.H.S.: Struck: structured output tracking with kernels. In: ICCV, pp. 263–270 (2011)
10. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE TPAMI **37**(3), 583–596 (2015)
11. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 702–715. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_50

12. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE TPAMI **34**(7), 1409–1422 (2012)
13. Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE TIP **25**(12), 5743–5756 (2016)
14. Li, C., Hu, S., Gao, S., Tang, J.: Real-Time Grayscale-Thermal Tracking via Laplacian Sparse Representation. Springer, Cham (2016)
15. Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: RGB-T object tracking: benchmark and baseline. arXiv:1805.08982 (2018)
16. Li, C., Lin, L., Zuo, W., Tang, J., Yang, M.H.: Visual tracking via dynamic graph learning. IEEE TPAMI (2018). https://doi.org/10.1109/TPAMI.2018.2864965
17. Li, C., Wang, X., Zhang, L., Tang, J., Wu, H., Lin, L.: Weighted low-rank decomposition for robust grayscale-thermal foreground detection. IEEE TCSVT **27**(4), 725–738 (2017)
18. Li, C., Wu, X., Zhao, N., Cao, X., Tang, J.: Fusing two-stream convolutional neural networks for RGB-T object tracking. Neurocomputing **281**, 78–85 (2018)
19. Li, C., Zhao, N., Lu, Y., Zhu, C., Tang, J.: Weighted sparse representation regularized graph learning for RGB-T object tracking. In: Proceedings of ACM MM (2017)
20. Liu, H.P., Sun, F.C.: Fusion tracking in color and infrared images using joint sparse representation. Inf. Sci. **55**(3), 590–599 (2012)
21. Liu, L., Xing, J., Ai, H.: Multi-view vehicle detection and tracking in crossroads. In: ACPR, pp. 608–612 (2011)
22. Liu, L., Xing, J., Ai, H., Xiang, R.: Hand posture recognition using finger geometric feature. In: ICPR, pp. 565–568 (2013)
23. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: IEEE ICCV, pp. 3074–3082 (2016)
24. Qi, Y., et al.: Hedged deep tracking. In: CVPR (2016)
25. Wang, M., Liu, Y., Huang, Z.: Large margin object tracking with circulant feature maps. In: CVPR, pp. 4800–4808 (2017)
26. Wu, Y., Blasch, E., Chen, G., Bai, L., Ling, H.: Multiple source data fusion via sparse representation for robust visual tracking. In: ICIF, pp. 1–8 (2011)
27. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIIMS **6**(3), 1758–1789 (2015)
28. Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.-H.: Fast visual tracking via dense spatio-temporal context learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 127–141. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_9
29. Zhang, T., Xu, C., Yang, M.H.: Multi-task correlation particle filter for robust object tracking. In: IEEE Conference on CVPR, pp. 4819–4827 (2017)
30. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparse collaborative appearance model. IEEE TIP **23**(5), 2356 (2014)