SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

April 2022, Vol. 65 142101:1–142101:13
https://doi.org/10.1007/s11432-020-3160-5

RGBT tracking via reliable feature configuration

Zhengzheng TU[1,2], Wenli PAN[1,2], Yunsheng DUAN[3], Jin TANG[1,2] & Chenglong LI[1,2*]

[1]*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei* 230601*, China;*
[2]*School of Computer Science and Technology, Anhui University, Hefei* 230601*, China;*
[3]*Network Information Center, Anhui University, Hefei* 230601*, China*

Received 12 July 2020/Revised 24 October 2020/Accepted 2 December 2020/Published online 17 March 2022

Abstract There is an increasing interest in RGBT tracking recently because of the complementary benefits of RGB and thermal infrared data. However, the reliability of each modality will change over time quite possibly, and the modality with bad reliability will disturb tracking performance. Thus, we propose a novel reliability-based feature configuration approach in the correlation filter framework for robust RGBT tracking. Specifically, we configure a feature set based on RGB, thermal, and RGBT data. To measure the reliabilities of different feature configurations, we equip each feature configuration with a tracker and design a guideline judging whether the tracker is reliable. We use the tracker with the best reliability for tracking. Experimental results show that the proposed tracker achieves promising performance against other RGBT tracking methods.

Keywords RGBT tracking, multi-modal, reliability guideline, feature configuration, correlation filter

Citation Tu Z Z, Pan W L, Duan Y S, et al. RGBT tracking via reliable feature configuration. Sci China Inf Sci, 2022, 65(4): 142101, https://doi.org/10.1007/s11432-020-3160-5

1 Introduction

In computer vision, visual tracking has become a hot topic, which tracks an arbitrary object where the initial state is given. Many efficient and effective tracking methods, such as deep learning-based methods [1–6] and correlation filtering-based methods [7–10], have been proposed. Numerous visual tracking methods belong to discriminative models, which take object tracking as a binary classification task in distinguishing the target from the background. As an important branch of discriminative models and owing to its good accuracy and efficiency, correlation filter-based trackers, for example, MOSSE [7], KCF [8], and DSST [11] trackers, are receiving a lot of attention. Visual tracking has been applied to many fields, such as video surveillance, navigation, human-computer interaction, and medical diagnosis.

Visual trackers based on RGB images are easily drifted and lose targets because of the influence of various complex circumstances, such as bad weather, low illumination, and cluttered background. To alleviate these problems, many researchers proposed multi-modal object tracking algorithms [12–17] by fusing thermal infrared data with RGB data, as thermal infrared imaging has a strong penetration force to overcome the effects of the aforementioned complex circumstances. Most existing RGBT tracking methods usually adopt different strategies to fuse different modalities, where all modalities are used in tracking all the time. For example, some RGBT methods [14, 16] pursue sparse representation for each modality and then optimize them in a joint learning framework. Some other methods [13, 15, 17] introduce modality weights to achieve adaptive fusion of different source data. However, the robustness of weight computation affects tracking performance much. Deep RGBT trackers [17, 18] use powerful feature representations to improve tracking performance greatly, but the tracking speed is generally slow. However, always adopting both modalities in existing methods might not be optimal for RGBT tracking as the modality with low reliability on some occasions will disturb tracking performance.

* Corresponding author (email: lcl1314@foxmail.com)

Figure 1 (Color online) Comparison of different trackers based on different feature configurations. Trackers 1–3 are configured with RGB, thermal infrared, and RGBT data, which are used in our tracking framework, and we update their predictions by the result with the most reliability, denoting them as Trackers 1′–3′, respectively.

To handle this problem, this paper presents a novel reliability-based feature configuration approach in the correlation filter framework for robust RGBT tracking. In the tracking process, we do not know which modality is good or bad, and the reliabilities of different modalities might also change over time. Therefore, we configure a feature set based on RGB, thermal, and RGBT data, aiming to dynamically select the best feature representation over time to exploit more reliable cues in tracking. Thus, we design a guideline judging the reliability of a feature in the feature set by evaluating the tracker with this feature as input. We adopt the correlation filter tracker as the baseline because of its good balance in robustness and efficiency, and other trackers can also be used in our framework. First, consistency indicates whether the tracker is stable. The higher the consistency between different trackers, the more reliable the algorithm is. Hence, we calculate the similarity between pairs of the tracker with the current feature configuration and other trackers. Second, the smoothness represents whether the tracker itself is reliable, evaluating the tracker's trajectory continuity. The trajectory continuity can represent the reliability of the tracker's results to some extent. Third, the robustness expresses whether the tracker is anti-jamming, which adds backward tracking to punish the damaged forward tracking. The damage degree of the tracker can be obtained by calculating the residual between forward tracking and backward tracking.

Comparison of the proposed method against trackers with different feature configurations is shown in Figure 1, in which it shows that single feature configuration is not robust to challenging environments while selecting the most reliable tracker in each frame to predict the tracking result by our guideline makes the best use of complementary benefits of different feature configurations.

The main contributions of our work are as follows.

• We propose a novel approach based on reliable feature configuration for robust RGBT tracking, handling the effects of noisy modalities by dynamically selecting the most reliable feature configuration in RGBT tracking.

• We propose a guideline consisting of consistency, smoothness, and robustness for evaluating the trackers with different feature configurations over time.

• We perform extensive experiments on two large-scale benchmark datasets, suggesting that our method achieves superior performance over compared methods.

2 Related work

2.1 RGB trackers

Popular RGB tracking methods are mainly classified into deep learning-based methods [1–6] and correlation filtering-based methods [8, 10, 11, 19, 20]. As a classic deep learning-based method, MDNet [1] proposed a CNN-based multi-domain learning model separating domain-specific information from the target. SiamFC [2] proposed a new tracking method based on a fully convolutional siamese network,

in which its speed was very fast. Based on MOSSE [7], KCF [8] introduced a cycle structure and a kernel function to increase the speed. CNT [19] extended the CSK [8] by using the color name (CN [21]) instead of targets' original grayscale features. SAMF [22] proposed scale adaptation and fused multiple features that are grayscale, CN [21] and HOG [23] for tracking. DSST [11] used the HOG [23] feature only and designed two independent trackers in estimating the target's position and scale, respectively, thus achieving adaptive updating of the target. Many methods [24–26] combined deep features and correlation filters for visual tracking because of the excellent feature representation performance of the deep neural network. C-COT [24] and ECO [25] used three interpolations to fuse deep feature maps with different resolutions and continuous domain learning to achieve accurate sub-pixel localization. ASRCF [26] used the fusion of deep features and HOG [23] feature to estimate the location and used HOG [23] feature to estimate the scale, thus improving location accuracy and reducing computation. According to the time setting, RGB tracking can also be divided into long-term tracking [4,5,6] and short-term tracking methods [1–3, 8, 10, 11, 19, 20]. As the recent work of long-term tracking, LTMU [4] proposed a meta-updater in controlling the tracker's update to solve inaccurate updates in long-term tracking. Globaltrack [5] proposed a new long-term target tracking method based on global search ideas to eliminate local location assumption. It enabled the tracker to search the target at any location and scale, thus avoiding cumulative error during long-term tracking. SPLT [6] proposed a long-term tracking method based on skimming and perusal modules, which can accurately capture the tracked target in the local search area and select the most likely candidate regions. In contrast to the aforementioned methods, the proposed method is a short-tracking method.

2.2 Multi-model tracking methods

To improve tracking performance, many researchers proposed to fuse multiple models [27–29]. Tang et al. [27] used different features to represent the target, adopted the online support vector machine (SVM) classifier for each kind of feature, and adaptively fused the classifiers. Kwon et al. [29] proposed a visual tracker sampler to achieve robust tracking by searching different components for the best trackers in each frame. The above methods indeed boost tracking performance. However, a single tracker easily drifts when the scenario is challenging. Some tracking frameworks based on multi-model selection [20, 30–32] have been proposed. MTA [20] proposed multihypothesis trajectory analysis to select the most robust one from multiple trackers. Also, MCCT [32] proposed to select the best expert for tracking from multiple experts that can learn multiple appearance models.

2.3 RGBT trackers

Many RGBT tracking algorithms [12–18, 33] adopted different techniques to fuse modalities recently. Some algorithms [12–15] adopted sparse representation to fuse different modalities. CMRT [16] proposed a cross-modal ranking algorithm to suppress the background and then used the structured SVM for RGBT tracking. Lan et al. [33] proposed an optimal discriminative learning framework to collaboratively learn a classifier for each modality and the reliability weight of each modality to realize modality fusion. Combining RGB and thermal modality with deep convolution neural networks is very popular in achieving high performance. For example, Zhu et al. [17] proposed a quality-aware feature aggregation network to integrate features from different modalities. Li et al. [18] proposed a multi-adapter convolutional network for RGBT tracking. Different from the aforementioned multi-modal fusion methods for RGBT tracking, we construct a reliability-based feature configuration by a guideline for RGBT tracking.

3 Proposed approach

3.1 Overview

There are many challenging scenarios in visual tracking, such as occlusion, haze, low illumination. Herein, we propose a new framework to select the most suitable tracker according to the reliability guideline, including robustness, smoothness, and consistency from multiple trackers, as shown in Figure 2. In the proposed framework, we adopt the HOG [23] and CN [21] features, and we form three kinds of feature configurations based on RGB, T (thermal infrared), and RGBT data. We use the correlation filter for each configuration to predict the tracking bounding box because of its good balance of accuracy and efficiency. Then, several predicted bounding boxes are obtained after the forward tracking and the

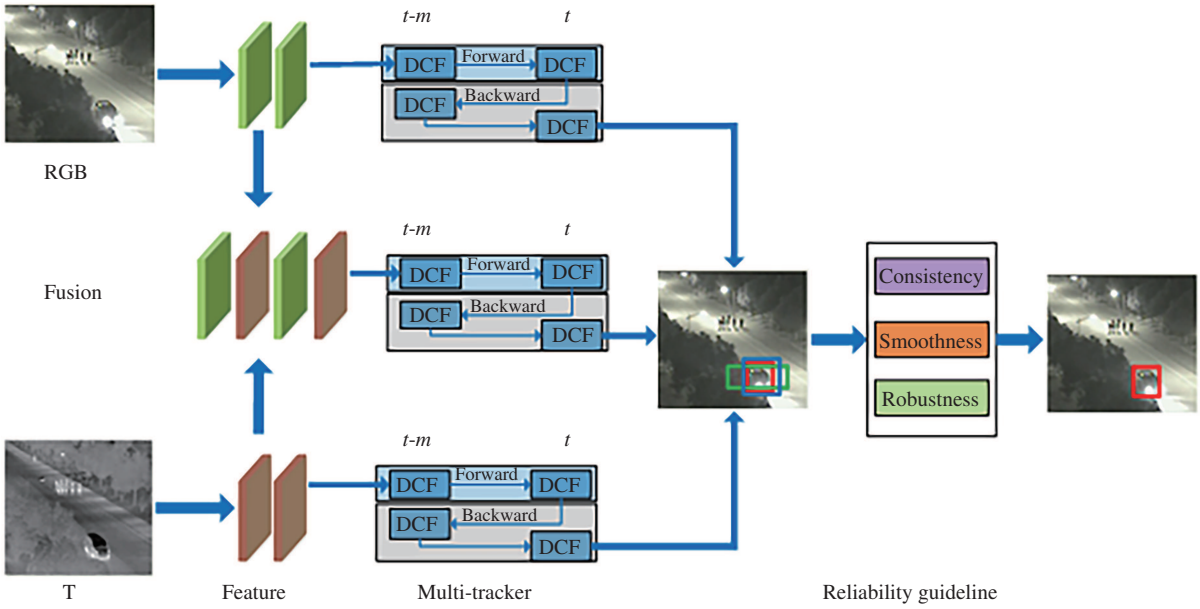

Figure 2 (Color online) Overview of our method. Three trackers with different inputs are obtained after configuring different features. The final result is predicted by the most reliable tracker evaluated by the reliability guideline, including consistency, smoothness, and robustness. Among them, the symbol t represents the t-th frame, and m represents the frame interval.

backward tracking. The forward trackers use the result of the best tracker from the previous frame every time, in other words, each tracker is initialized to the same before starting tracking at each frame to prevent the tracker from corruption. To select the most robust tracker, we employ backward tracking to define some guidelines to evaluate tracking reliabilities. We select the tracker with the highest reliability score and take the corresponding prediction as the final tracking result. Figure 2 shows an overview of the proposed method.

3.2 Feature configuration

As the first step, different feature configurations from RGB image, thermal infrared image, and their fusion are input into correlation filter-based tracker, whose procedure is described in Subsection 3.4. Specifically, the feature configuration input to the first tracker is from RGB images, and the calculation is given as follows:

$$\boldsymbol{X}^R = \boldsymbol{X}_h^R \oplus \boldsymbol{X}_c^R, \tag{1}$$

where $\boldsymbol{X}_h^R$ and $\boldsymbol{X}_c^R$ denote the matrices of HOG [23] and CN [21] features respectively from RGB images, and $\oplus$ is a concatenation operation.

The feature configuration to the second tracker is from thermal infrared images. The formula is similar to (1):

$$\boldsymbol{X}^T = \boldsymbol{X}_h^T \oplus \beta \boldsymbol{X}_c^T, \tag{2}$$

where $\boldsymbol{X}_h^T$ and $\boldsymbol{X}_c^T$ denote the matrices of HOG [23] and CN [21] features respectively from thermal infrared images. Since the role of color information is not very significant in thermal infrared images, it is assigned a low weight as β.

The feature configuration to the third tracker is from a fusion of RGB and thermal infrared image pair, combining advantages of RGB and thermal images. The formula is written as follows:

$$\boldsymbol{X}^F = \boldsymbol{X}^R \oplus \boldsymbol{X}^T. \tag{3}$$

Since the advantage of the CN [21] feature in thermal infrared images is not obvious, we design a parameter to balance the proportion of color features in RGB and thermal infrared images. The more detailed formula of (3) is as follows:

$$\boldsymbol{X}^F = \left(\boldsymbol{X}_h^R \oplus \boldsymbol{X}_h^T\right) \ \oplus \ \left((1-\sigma)\boldsymbol{X}_c^R \oplus \sigma \boldsymbol{X}_c^T\right), \tag{4}$$

where σ is a balance parameter. Other feature configurations, such as the combinations of different modalities with different weights obtained by learning, can also be used in the proposed framework, whereas simple combination is used to demonstrate the effectiveness of the proposed framework.

3.3 Reliability guideline

To deal challenges in various scenes, we design three trackers. However, we ultimately choose the most robust one by evaluating the reliability of each tracker with the following considerations.

Consistency. A more stable tracking is achieved when there is more consistency between trackers. As discussed previously, we design three kinds of feature configurations. We adopt specifically the same calculation way for consistency score as in the MCCT [32]. The consistency score reflects the similarity between trackers written as

$$\boldsymbol{S}^k_{\mathrm{cons}(t)} = \frac{\sum_{j=0}^{J} w_j \ \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{M}^k_{(n,[t-(J-j)+1])}}{\sum_{j=0}^{J} w_j \boldsymbol{O}^k_{(n,[t-(J-j)+1])} + \mu \sum_{j=0}^{J} w_j}, \tag{5}$$

where N and t denote the number of trackers and the current frame, respectively, and J indicates the period of calculation to increase the stability of the consistency score. $\boldsymbol{M}^k_{(n,[t-(J-j)+1])}$ and $\boldsymbol{O}^k_{(n,[t-(J-j)+1])}$ indicate the mean overlap ratio and standard variance of the overlap ratio of the k-th tracker and the n-th tracker in the $[t-(J-j)+1]$-th frame, respectively. $w_j = (1.1)^j$ represents the weight of $(j+1)$-th element in the period of calculation. $\boldsymbol{S}^k_{\mathrm{cons}(t)}$ is the consistency score of the k-th tracker in the t-th frame. As the frame is closer to the current frame, j and w_j become larger. μ is a constant for avoiding the zero as the denominator.

Smoothness. A big smoothness score means the trajectory has a small fluctuation during tracking, indicating that the tracking is more accurate. Following the MCCT [32], the smoothness score reflects the smoothness of the trajectory of the tracker during tracking. The formula is written as follows:

$$\boldsymbol{S}^k_{\mathrm{smoo}(t)} = \frac{\sum_{j=0}^{J} w_j \ \exp(-\frac{2||\boldsymbol{B}^k_{t-(J-j)} - \boldsymbol{B}^k_{t-(J-j)+1}||^2}{(\boldsymbol{W}^k_{t-(J-j)+1} + \boldsymbol{H}^k_{t-(J-j)+1})^2})}{\sum_{j=0}^{J} w_j}, \tag{6}$$

where $\boldsymbol{B}^k_{t-(J-j)}$ and $\boldsymbol{B}^k_{t-(J-j)+1}$ can be regarded as the center of the bounding box in the $[t-(J-j)]$-th frame and $[t-(J-j)+1]$-th frame of the k-th tracker, respectively. $\boldsymbol{W}^k_{t-(J-j)+1}$ and $\boldsymbol{H}^k_{t-(J-j)+1}$ are the width and height of the bounding box of the k-th tracker in the $[t-(J-j)+1]$-th frame, respectively. $\boldsymbol{S}^k_{\mathrm{smoo}(t)}$ is the smoothness score of the k-th ($k \in \{1, 2, 3\}$) tracker in the t-th frame.

Robustness. In addition to using the consistency and the smoothness to evaluate the stability of the trackers in specific scenarios, we also adopt the robust score to evaluate the robustness of the trackers themselves in the course of backward tracking [20]. The formula is as follows:

$$\boldsymbol{S}^k_{\mathrm{rob}(t)} = \exp\left(-\frac{||\overrightarrow{\boldsymbol{B}}^k_t - \overleftarrow{\boldsymbol{B}}^k_t||^2}{\delta^2}\right), \tag{7}$$

where $\overrightarrow{\boldsymbol{B}}^k_t$ and $\overleftarrow{\boldsymbol{B}}^k_t$ indicate the center of the bounding box for forward tracking and backward tracking in the k-th tracker of the t-th frame, respectively, and $\delta^2 = 500$. $\boldsymbol{S}^k_{\mathrm{rob}(t)}$ is the robust score, and it should be 1 without any deviation.

Overall reliability guideline. The overall reliability guideline of a tracker is derived from the aforementioned three parts, and the formula is as follows:

$$\boldsymbol{S}^k_{\mathrm{rel}(t)} = \xi \boldsymbol{S}^k_{\mathrm{cons}(t)} + (1-\xi)\ \boldsymbol{S}^k_{\mathrm{smoo}(t)} + \gamma \boldsymbol{S}^k_{\mathrm{rob}(t)}, \tag{8}$$

where ξ is a balanced parameter to control the sum of consistency score and smoothness score to be no more than 1. γ is a score weight that controls the ratio of three parts, preventing one of the scores becoming meaningless.

3.4 Tracking procedure

In this work, we adopt the standard discriminative correlation filter model, DCF [8], for each tracker, aiming to minimize the following problem:

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \left\| \boldsymbol{y} - \sum_{d=0}^{D} \boldsymbol{x}_d * \boldsymbol{w}_d \right\|^2 + \lambda \sum_{d=0}^{D} \|\boldsymbol{w}_d\|_2^2, \tag{9}$$

where $\boldsymbol{x}_d$ and $\boldsymbol{w}_d$ are the d-th ($d \in \{0, 1, \ldots, D-1\}$) channel of the feature data and filter, respectively, and $*$ is an element-wise multiplication operation. $\boldsymbol{y}$ is the sample's response obtained from the Gaussian distribution, and λ is a regularization parameter to overcome the overfitting. As same as the DCF-based methods [8, 22], we use the adaptive template online training to acquire features from the current frame. We crop the ROI (region of interest) of object position in the previous frame to obtain the feature map $\boldsymbol{z}_d^k$ and then obtain the response map $\boldsymbol{r}^k$ of multiple features as follows:

$$\boldsymbol{r}^k = \sum_{d=1}^{D} \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{w}_d^k) \odot (\mathcal{F}(\boldsymbol{z}_d^k))^{\mathrm{H}}), \tag{10}$$

where $\boldsymbol{w}_d^k$ represents the d-th ($d \in \{0, 1, \ldots, D-1\}$) channel of the filter in the k-th tracker. The target location is determined by finding the maximum value of $\boldsymbol{r}^k$. The position and scale estimation of the object is determined as the same as DSST [11]. In each frame, we first predict an optimal position and then compute the scale with the maximum response value by adjusting the scale of the bounding box.

Model update. After selection, we update the correlation filter online to adapt changes of target and background at any time. The adopted hand-crafted features (HOG [23] and CN [21]) have fast calculation speed and high resolution. The online adaptive updating model is described as follows:

$$\mathcal{F}^{\star}(\boldsymbol{w}_t) = (1-\eta)\mathcal{F}^{\star}\left(\boldsymbol{w}_{(t-1)}\right) + \eta\mathcal{F}(\boldsymbol{w}_t), \tag{11}$$

where $\mathcal{F}^{\star}(\boldsymbol{w}_t)$ is the updated model of the t-th frame, $\mathcal{F}^{\star}(\boldsymbol{w}_{(t-1)})$ is the model of the $(t-1)$-th frame, $\mathcal{F}(\boldsymbol{w}_t)$ is the current model, and η is the learning rate. However, it is inevitable that the model will drift as it learns background information when occlusion occurs. As the same as MCCT [32], we use the average reliability score as the threshold. When the current trackers average reliable score is higher than the threshold, the learning rate will not change, otherwise it will be updated. The learning rate is given as follows:

$$\eta = \begin{cases} \epsilon, & \text{if } R_t > \alpha \cdot P_t, \\ \epsilon \cdot [R_t / (\alpha \cdot P_t)]^{\theta}, & \text{otherwise}, \end{cases} \tag{12}$$

where $R_t = \frac{1}{K}\sum_{k=1}^{K} S_{\mathrm{rel}(t)}^k$ is the average reliability score of trackers, which becomes very small when there is an occlusion or severe deformation of the target. $P_t = \frac{1}{t}\sum_{i=1}^{t} R_i$ is the average of the total reliability score of the trackers from the first frame to the current frame. α and θ represent the threshold of reliability and the power exponent of the weight function, respectively. ϵ is the learning rate of the model. The sample will be considered unreliable when there is a low robustness score of R_t, and then this weight function will punish it and prevent corruption of the model.

4 Experiments

In this section, we present an extensive experimental evaluation for the proposed method. First, we describe the implementation details and evaluation setting. Then, we validate the effectiveness of each component of the proposed model through ablation studies.

4.1 Evaluation setting

We conduct experiments on the MATLAB 2016b (MathWorks, Natick, MA, USA) platform with Intel I5-7500K 3.40-GHz CPU of 16 GB RAM with a performance of 10 fps. There are three large RGBT tracking benchmark datasets: GTOT [13], RGBT210 [15], and RGBT234 [34]. Since the RGBT234 [34] dataset

Table 1 PR/SR (%) of the proposed method with different parameters on the GTOT dataset [13]

Parameter	Setting	PR	SR
β	0.01	84.5	69.2
	0.1	86.4	70.7
	1	84.8	69.4
σ	0	84.6	69.3
	0.1	86.4	70.7
	0.5	84.2	69.0

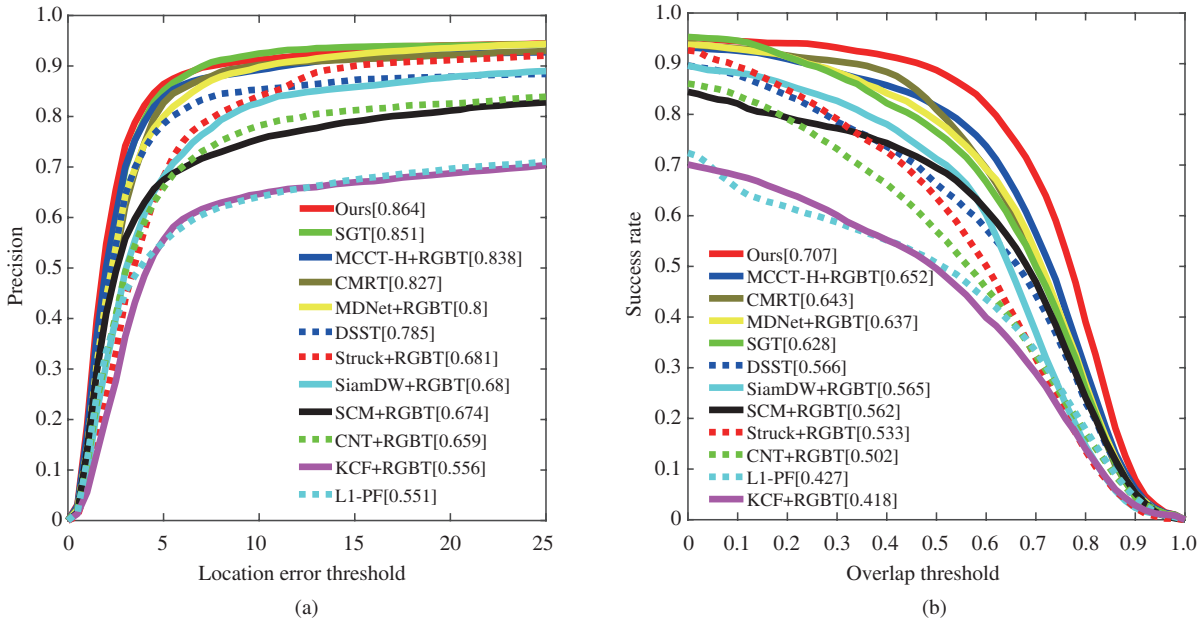

Figure 3 (Color online) PR/SR curves on GTOT [13], where the representative PR/SR scores are presented in the legend. (a) PR; (b) SR.

is a large RGBT tracking dataset extended from the RGBT210 [15] dataset, we validate the effectiveness of the proposed method on RGBT234 [34] and GTOT [13], covering various challenging scenarios for visual tracking. We use precision rate (PR) and success rate (SR) for quantitative evaluation. PR is the percentage of frames whose location of the tracked object is within the threshold distance of the ground truth, whereas SR is the percentage of successful frames whose overlap exceeds the given threshold.

Parameters. The parameter setting of the proposed method is as follows. From the extensive experiments, we set the score balanced parameter ξ to 0.2 and set $\alpha = 0.6$, $\theta = 3$, and the learning rate $\epsilon = 0.01$ for the adaptive model updating. β and σ are the feature configuration parameters. β is a weight reduction parameter for reducing the influence of the noise from the thermal infrared modality on the whole tracking. σ is a balance parameter to control the proportion of color features in different modalities. According to the results with different reduction weights shown in Table 1, we set the weight reduction parameter $\beta = 0.1$, and the color feature balance parameter $\sigma = 0.1$. To make a fair comparison, we fix all remaining parameters and other settings by changing the parameters within a certain range in all experiments so that all parameters are the most optimal in the end.

4.2 Evaluation on GTOT dataset

Overall performance. We perform evaluations on GTOT [13] with 50 RGBT challenging videos, and compare our method with 11 state-of-the-art trackers, including SGT [15], CMRT [16], DSST [11], L1-PF [14], MCCT-H [32]+RGBT, SiamDW [35]+RGBT, MDNet [1]+RGBT, KCF [8]+RGBT, SCM [36]+RGBT, Struck [37]+RGBT and CNT [19]+RGBT. We extend some RGB methods to RGBT methods by concatenating the images of RGBT because of the small number of RGBT trackers, including SCM [36]+RGBT, SiamDW [35]+RGBT, MDNet [1]+RGBT, CNT [19]+RGBT, Struck [37]+RGBT and MCCT-H [32]+RGBT. Figure 3 shows the evaluation results of our method against other state-of-

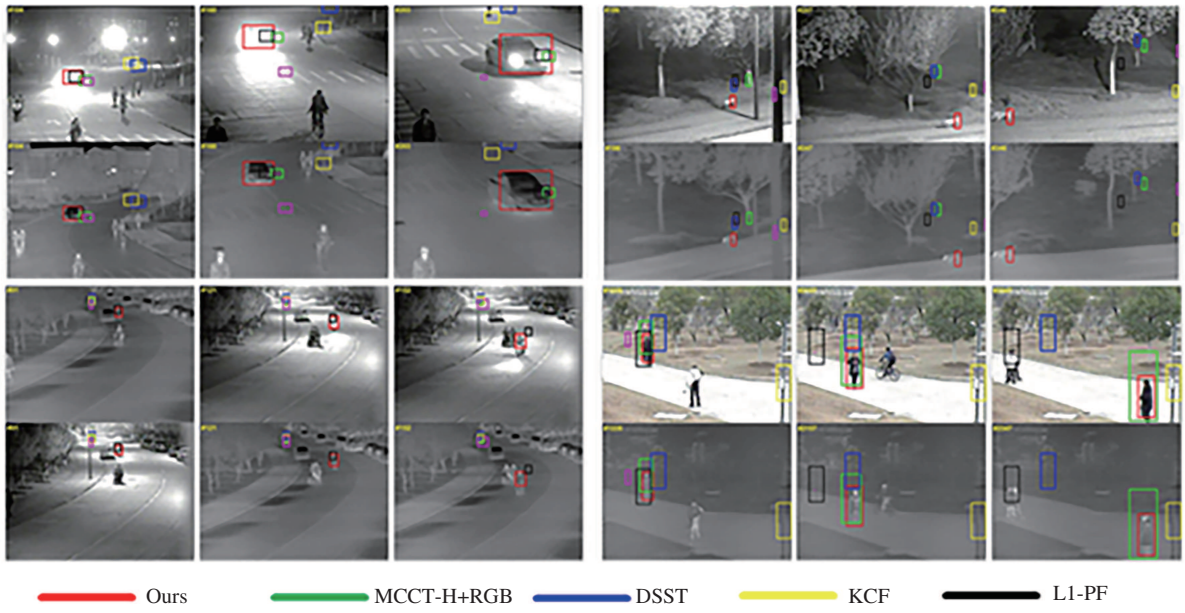

Figure 4 (Color online) Visual comparison of our method with four state-of-the-art trackers on GTOT [13].

Table 2 PR/SR (%) results with different RGBT trackers under different challenges on GTOT [13][a)]

	SGT	CMRT	SCM+RGBT	Struck+RGBT	MCCT-H+RGBT	SiamDW+RGBT	KCF+RGBT	MDNet+RGBT	Ours
DEF	91.9/73.3	84.7/65.2	59.7/50.0	75.6/60.4	89.8/70.2	69.1/58.2	55.3/43.7	81.7/68.8	**93.4/76.0**
LI	88.4/65.1	88.7/67.8	66.9/56.1	74.0/55.3	85.7/66.6	69.9/58.8	53.8/41.0	79.5/64.3	**90.1/73.7**
TC	84.8/61.5	81.1/62.2	68.8/55.6	67.9/51.0	82.2/62.1	63.5/51.7	49.9/36.7	79.5/60.9	**85.7/69.3**
SO	91.7/61.8	86.5/61.0	72.1/53.1	74.5/52.7	88.1/63.1	76.4/58.5	47.9/31.5	87.0/62.2	**92.1/70.3**
FM	79.8/55.9	**83.5/65.0**	69.1/58.9	63.9/51.8	71.7/56.9	71.1/57.6	48.1/34.6	80.5/59.8	79.6/64.9
LSV	84.1/54.7	**85.3**/66.7	79.6/64.7	66.0/49.6	82.5/63.4	68.9/56.5	60.5/42.2	77.0/57.3	83.6/**68.3**
OCC	80.9/56.7	82.5/62.6	49.2/50.7	71.7/51.6	74.8/58.8	67.5/53.6	53.8/36.4	**82.9/64.1**	79.1/63.5
All	85.1/62.8	82.7/64.3	67.4/56.2	68.1/53.3	83.8/65.2	68.0/56.5	55.6/41.8	80.0/63.7	**86.4/70.7**

a) The best results are shown in bold.

the-art trackers on GTOT [13]. Our algorithm adopts hand-crafted features only, but it is superior to some deep learning methods like SiamDW [35] and MDNet [1], as shown in Figure 3. Specifically, our method is 0.013/0.026 higher than SGT [15]/MCCT-H [32]+RGBT respectively in PR, and 0.079/0.055 higher than SGT [15]/MCCT-H [32]+RGBT in SR. Experimental performances prove that selecting the best tracker for the current scene is effective. More visual comparisons are given in Figure 4.

Challenge-based performance. GTOT [13] has 7 challenges: deformation (DEF), low illumination (LI), thermal crossover (TC), small object (SO), fast motion (FM), large scale variation (LSV), and occlusion (OCC). We compare our approach with other state-of-the-art methods on these challenges on GTOT [13], as shown in Table 2. The results show that our approach performs best under DEF, LI, SO, and TC, proving the effectiveness of our feature configuration and reliability guideline. The proposed method does not perform best when facing the challenges of FM, LSV, and OCC. OCC has always been a difficult problem for visual tracking (not only in RGBT tracking). The model based on standard DCF [8] in our method may not be able to perfectly deal with these problems when the partial target is occluded. Some comparative methods, including SGT [15], CMRT [16], and MDNet [1]+RGBT, design the reliable appearance model that can successfully track the target according to the local feature of the target. However, they do not make good use of the complementary properties of the modalities on LI and TC. For LSV, our method locates the target first and then refines the scale of the target. As location and scale refinement are locally optimal, the result may not be globally optimal when the scale varies greatly. Besides, FM often brings boundary effects to DCF [8] and thus produces wrong samples, which will weaken the discriminant ability of the classifier. However, the proposed method does not specifically deal with this problem. We will continue to improve the aforementioned shortcomings by learning the filter coefficients of different patches and improving the model's representation capability in future work.

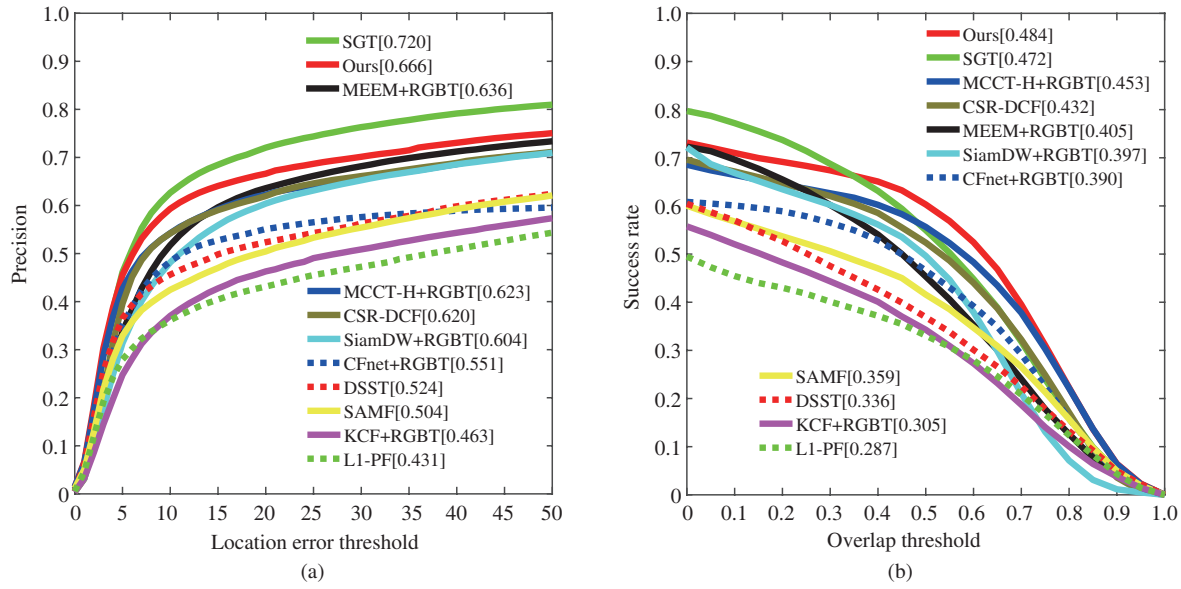

Figure 5 (Color online) PR/SR curves on RGBT234 [34], where the representative PR/SR scores are presented in the legend. (a) PR; (b) SR.

4.3 Evaluation on RGBT234 dataset

We also perform a comparison with RGBT234 [34] dataset, which contains 234 RGBT challenging videos with 12 attributes. We compare our method with 10 tracking algorithms, including SGT [15], L1-PF [14], MCCT-H [32]+RGBT, SiamDW [35]+RGBT, KCF [8]+RGBT, MEEM [30]+RGBT, CFNet [38]+RGBT, DSST [11], SAMF [22] and CSR-DCF [39], where DSST [11], SAMF [22] and CSR-DCF [39] are RGB methods. Figure 5 shows the results of our method and other state-of-the-art trackers on RGBT234 [34]. Compared with MCCT-H [32]+RGBT, the performance of our method increases by 0.043/0.031 in PR/SR, respectively. Although MCCT-H [32]+RGBT uses seven different trackers per frame, our method with three configurations has better tracking performance when we add backward tracking to three trackers, thus greatly improving the reliability. Besides, SGT [15] performs well in PR on account of using weighted sparse representation to fuse different modalities. However, their optimization process is too complicated. Our method performs better than SGT [15] in SR.

Challenge-based performance. RGBT234 [34] has 12 challenges: low resolution (LR), TC, no occlusion (NO), heavy occlusion (HO), LI, motion blur (MB), DEF, FM, scale variation (SV), camera moving (CM), partial occlusion (PO) and background clutter (BC). We compare our approach with other state-of-the-art methods on these challenges on RGBT234 [34] in Figure 6. The results show that our approach performs best in most challenges except for TC and MB. As thermal modality has the weakness of TC sometimes, it will invalidate the reliability guideline when thermal modality obtains large reliability. Hence, we will design some penalties to reduce modal noise in future work. For the challenge of MB caused by high-speed movement of camera or target, compared with CSR-DCF [39] and MCCT-H [32]+RGBT, the advantage of our method is not obvious. CSR-DCF [39] considers spatial reliability and channel reliability, and MCCT-H [32]+RGBT uses seven experts for tracking at the same time, taking full advantage of the consistency and difference between experts, both of them improve the adverse effect of motion blur. We will alleviate it by adding contextual information to improve the target's appearance in our future work.

4.4 Ablation study

To verify the validity of the selection system in our tracker, we implement and evaluate three variants of our method on GTOT [13]. Two variants and our method are constructed as follows. (1) "Ours-1F" means that we remove the two trackers for RGB and thermal data. (2) "Ours-3F" means that we fuse three trackers (RGB+thermal infrared, RGB, and thermal infrared) as one tracker. (3) "Ours" represents our tracker, which designs a selection system for three trackers (RGB+termal infrared, RGB, and thermal infrared). As shown in Table 3, the experimental results indicate that our method is superior to these

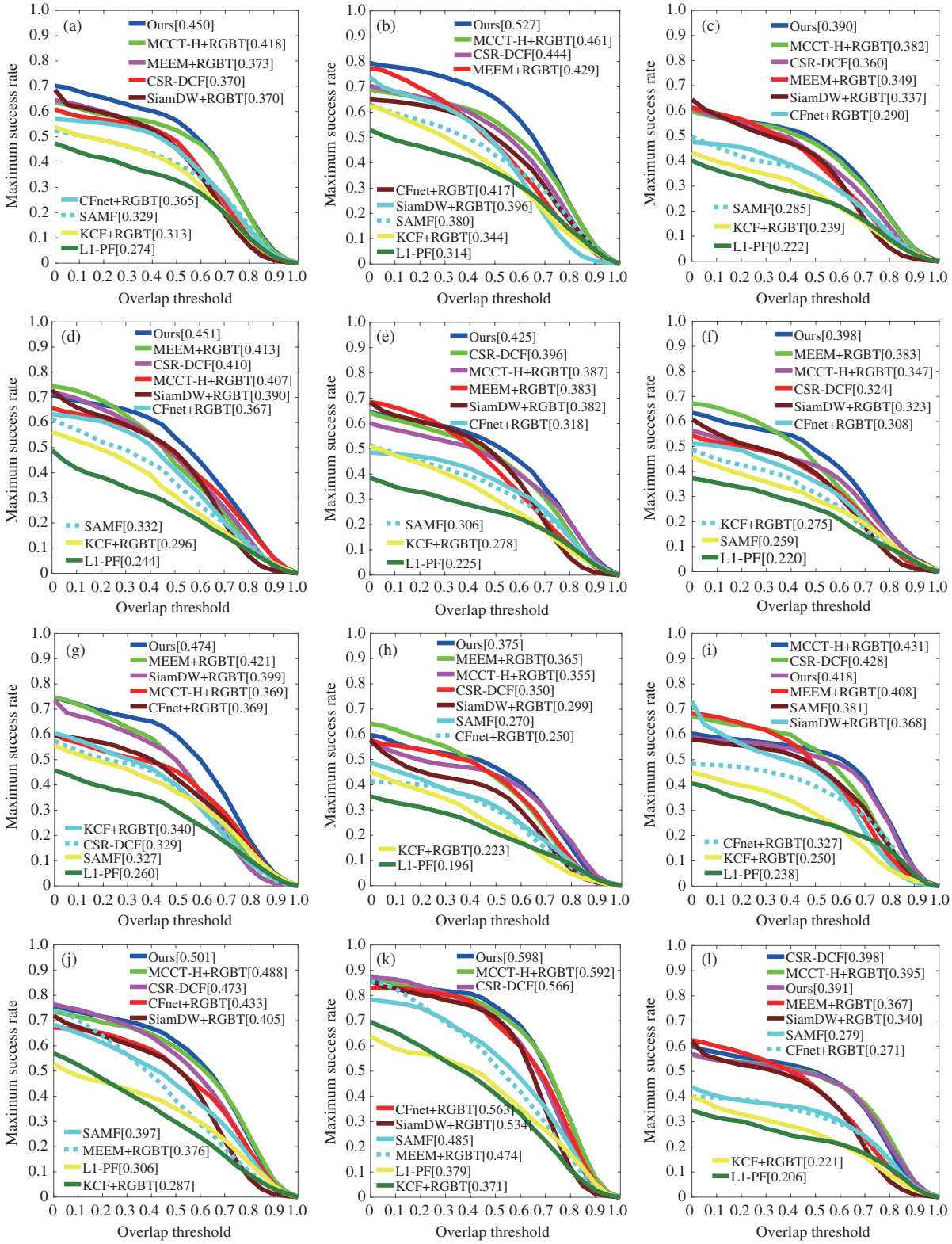

Figure 6 (Color online) SR curves under 12 challenges on RGBT234 [34], where the representative SR scores are presented in the legend. (a) Low resolution; (b) partial occlusion; (c) heavy occlusion; (d) deformation; (e) camera moving; (f) background clutter; (g) low illumination; (h) no occlusion; (i) motion blur; (j) scale variation; (k) fast motion; (l) thermal crossover.

variants, verifying that the selection strategy in our method is better than fusing multi-trackers directly.

To verify the validity of the reliability guideline in our tracker, we implement and evaluate two variants versions and our method on GTOT [13]. We construct three methods: "Baseline", "Ours-BA" and

Table 3 PR/SR (%) of the proposed method and the variants with different fusion ways on the GTOT dataset [13]

	Ours-1F	Ours-3F	Ours
PR	84.9	84.8	86.4
SR	69.3	69.1	70.7

Table 4 PR/SR (%) of the proposed method and the variants with different selection strategies on the GTOT dataset [13]

	Baseline	Ours-BA	Ours
PR	83.8	83.4	86.4
SR	65.2	68.6	70.7

Table 5 PR/SR (%) of the proposed method where the trackers are composed of different feature configurations on the GTOT dataset [13]

Feature configuration	Tracker	PR/SR
HOG, CN(RGB) HOG, CN(T) HOG, σCN(RGBT) HOG, σ_1CN(RGBT) HOG, σ_2CN(RGBT)	Ours+2	83.7/68.4
HOG, CN(RGB) HOG, CN(T) HOG, σCN(RGBT)	Ours	86.4/70.7
HOG, CN(RGB) HOG, CN(T)	Ours–1	73.3/61.1
HOG, CN(RGB)	Ours–2(1)	72.0/60.5
HOG, CN(T)	Ours–2(2)	71.5/59.6

"Ours". (1) "Baseline" means that we extend MCCT-H [32] to the RGBT method, which has seven trackers with a selection strategy. (2) "Ours-BA" indicates that we remove the reliability guideline and use a selection strategy as (1). (3) "Ours" represents our tracker, which uses the reliability guideline to choose the best one. As shown in Table 4, the experimental results indicate that our reliability guideline is superior to the selection strategy in MCCT-H [32].

To verify the optimality of the number of feature configurations adopted by our tracker, we implement a comparative analysis of three variant versions and our method: "Ours+2", "Ours–1", "Ours–2(1)", "Ours–2(2)" and "Ours". (1) "Ours+2" is a representation of five feature configurations. On the basis of our method, it adds two new feature configurations from RGBT fusion modality, one is the concatenation of HOG [23] and CN [21] features with σ_1 weight (0.7*CN(RGB), 0.3*CN(T)), and the other is the concatenation of HOG [23] and CN [21] features with σ_2 weight (0.5*CN(RGB), 0.5*CN(T)). (2) "Ours" is our method, as mentioned earlier, composed of three feature configurations. (3) "Ours–1" is the remaining feature configurations after the RGBT feature configuration is removed. It just uses the concatenation of HOG [23] and CN [21] features from two modalities. (4) "Ours–2(1)" is the remaining feature configuration after removing the RGBT feature and T feature configurations. It just uses the concatenation of HOG [23] and CN [21] features from the RGB modality. (5) "Ours–2(2)" is the remaining feature configuration after removing the RGBT feature and RGB feature configurations. It only uses the concatenation of HOG [23] and CN [21] features from T modality. Table 5 shows the detailed feature configuration. As shown in the table, we emphasize RGB data as more reliable than T data, and the adopted feature configuration achieves the best performance against other configurations. If fewer features are used, the representation capacity of these features will limit the performance of the tracker. If the number of feature configurations increases ("Ours+2" in Table 5), the performance might not be better because the consistency among trackers would make the evaluation of our reliability criteria not reliable in challenging scenarios. For this problem, we will research it in future work.

4.5 Limitation and future work

There are two major limitations in our method. (1) The speed of our tracking method is not real-time. (2) Performance is not improved when we use more feature configurations. Given these two limitations, we will make the following improvements in future work. First of all, we will use multi-thread and other

parallel calculations to improve the speed of the backward tracking because backward tracking is the most time-consuming part. We will then adopt more discriminative and powerful features such as the context feature and the deep feature to improve the consistency among trackers.

5 Conclusion

In this paper, we propose an efficient reliability-based feature configuration approach based on the correlation filter. Our method designs multiple trackers by various feature configurations. It then uses the new reliability guideline to select the best tracker with the least noise and the most robust appearance representation. Compared with many RGBT tracking algorithms, our method is simple, efficient, and superior to most compared trackers.

Acknowledgements The work was supported by Natural Science Foundation of Anhui Higher Education Institution of China (Grant Nos. KJ2020A0033, KJ2019A0005, KJ2019A0026), Major Project for New Generation of AI (Grant No. 2018AAA0100400), and National Natural Science Foundation of China (Grant No. 61976003), and NSFC Key Projects in International (Regional) Cooperation and Exchanges (Grant No. 61860206004).

References

1 Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 4293–4302
2 Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking. In: Proceedings of the European Conference on Computer Vision, Amsterdam, 2016. 850–865
3 Li B, Yan J, Wu W, et al. High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 8971–8980
4 Dai K, Zhang Y, Wang D, et al. High-performance long-term tracking with meta-updater. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, 2020. 6298–6307
5 Huang L, Zhao X, Huang K. GlobalTrack: a simple and strong baseline for long-term tracking. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020
6 Yan B, Zhao H, Wang D, et al. 'Skimming-Perusal' tracking: a framework for real-time and robust long-term tracking. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul, 2019. 2385–2393
7 Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 2544–2550
8 Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 583–596
9 Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of the European Conference on Computer Vision, Florence, 2012. 702–715
10 Danelljan M, Hager G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 4310–4318
11 Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference, Nottingham, 2014
12 Liu H P, Sun F C. Fusion tracking in color and infrared images using joint sparse representation. Sci China Inf Sci, 2012, 55: 590–599
13 Li C, Cheng H, Hu S, et al. Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Trans Image Process, 2016, 25: 5743–5756
14 Wu Y, Blasch E, Chen G, et al. Multiple source data fusion via sparse representation for robust visual tracking. In: Proceedings of the 14th International Conference on Information Fusion, Chicago, 2011. 1–8
15 Li C, Zhao N, Lu Y, et al. Weighted sparse representation regularized graph learning for RGB-T object tracking. In: Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, 2017. 1856–1864
16 Li C, Zhu C, Huang Y, et al. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In: Proceedings of the European Conference on Computer Vision, Munich, 2018. 808–823
17 Zhu Y, Li C, Tang J, et al. Quality-aware feature aggregation network for robust RGBT tracking. IEEE Trans Intell Veh, 2021, 6: 121–130
18 Li C L, Lu A D, Zheng A H, et al. Multi-adapter RGBT tracking. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, 2019
19 Danelljan M, Khan F S, Felsberg M, et al. Adaptive color attributes for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 1090–1097
20 Lee D Y, Sim J Y, Kim C S. Multihypothesis trajectory analysis for robust visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 5088–5096
21 van de Weijer J, Schmid C, Verbeek J, et al. Learning color names for real-world applications. IEEE Trans Image Process, 2009, 18: 1512–1523
22 Li Y, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration. In: Proceedings of the European Conference on Computer Vision, Zurich, 2014. 254–265
23 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, 2005. 886–893
24 Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking. In: Proceedings of the European Conference on Computer Vision, Amsterdam, 2016. 472–488
25 Danelljan M, Bhat G, Khan F S, et al. Eco: efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 6931–6939
26 Dai K, Wang D, Lu H, et al. Visual tracking via adaptive spatially-regularized correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 4670–4679

27 Tang F, Brennan S, Zhao Q, et al. Co-tracking using semi-supervised support vector machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Rio de Janeiro, 2007. 992–999
28 Gao Y, Ji R, Zhang L, et al. Symbiotic tracker ensemble toward a unified tracking framework. IEEE Trans Circ Syst Video Technol, 2014, 24: 1122–1131
29 Kwon J, Lee K M. Tracking by sampling trackers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Barcelona, 2011. 1195–1202
30 Zhang J, Ma S, Sclaroff S. MEEM: robust tracking via multiple experts using entropy minimization. In: Proceedings of the European Conference on Computer Vision, Zurich, 2014. 188–203
31 Li J, Deng C, Xu R Y D, et al. Robust object tracking with discrete graph-based multiple experts. IEEE Trans Image Process, 2017, 26: 2736–2750
32 Wang N, Zhou W, Tian Q, et al. Multi-cue correlation filters for robust visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 4844–4853
33 Lan X, Ye M, Zhang S, et al. Robust collaborative discriminative learning for RGB-infrared tracking. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018. 7008–7015
34 Li C, Liang X, Lu Y, et al. RGB-T object tracking: benchmark and baseline. Pattern Recogn, 2019, 96: 106977
35 Zhang Z, Peng H. Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 4591–4600
36 Zhong W, Lu H, Yang M H. Robust object tracking via sparsity-based collaborative model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 1838–1845
37 Hare S, Golodetz S, Saffari A, et al. Struck: structured output tracking with kernels. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 2096–2109
38 Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 5000–5008
39 Lukezic A, Vojir T, Zajc L C, et al. Discriminative correlation filter with channel and spatial reliability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 6309–6318