Cross-Modal Ranking with Soft Consistency and Noisy Labels for Robust RGB-T Tracking

Chenglong Li[1,2][0000−0002−7233−2739], Chengli Zhu[2][0000−0001−8714−6755], Yan Huang[1][0000−0002−8239−7229], Jin Tang[2][0000−0002−4123−268X], and Liang Wang[1][0000−0001−5224−8647]

[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Labotary of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]School of Computer Science and Technology, Anhui Univeristy
{lcl1314,zcl912,jtang99029}@foxmail.com, {yhuang,wangliang}@nlpr.ia.ac.cn

Abstract. Due to the complementary benefits of visible (RGB) and thermal infrared (T) data, RGB-T object tracking attracts more and more attention recently for boosting the performance under adverse illumination conditions. Existing RGB-T tracking methods usually localize a target object with a bounding box, in which the trackers or detectors is often affected by the inclusion of background clutter. To address this problem, this paper presents a novel approach to suppress background effects for RGB-T tracking. Our approach relies on a novel cross-modal manifold ranking algorithm. First, we integrate the *soft cross-modality consistency* into the ranking model which allows the sparse inconsistency to account for the different properties between these two modalities. Second, we propose an *optimal query learning* method to handle label noises of queries. In particular, we introduce an intermediate variable to represent the optimal labels, and formulate it as a l_1-optimization based sparse learning problem. Moreover, we propose a single unified optimization algorithm to solve the proposed model with stable and efficient convergence behavior. Finally, the ranking results are incorporated into the patch-based object features to address the background effects, and the structured SVM is then adopted to perform RGB-T tracking. Extensive experiments suggest that the proposed approach performs well against the state-of-the-art methods on large-scale benchmark datasets.

Keywords: Visual tracking, Information fusion, Manifold ranking, Soft cross-modality consistency, Label optimization

1 Introduction

The goal of RGB-T tracking is to estimate the states of the target object in videos by fusing RGB and thermal (corresponds the visible and thermal infrared spectrum data, respectively) information, given the initial ground truth bounding box. Recently, researchers pay more and more attention on RGB-T tracking [1,2,3,4,5] partly due to the following reasons. i) The imaging quality of

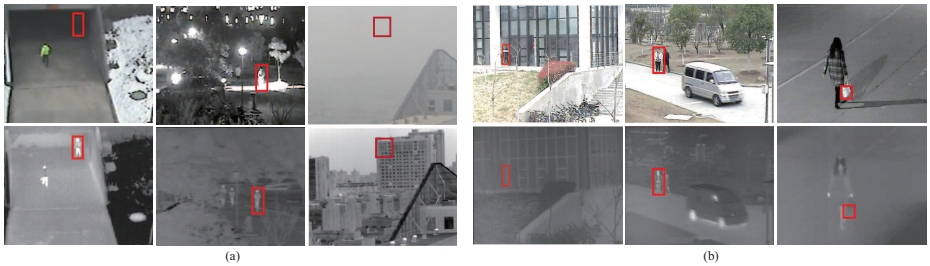

Fig. 1. Typically complementary benefits of RGB and thermal data [5]. (a) Benefits of thermal sources over RGB ones, where visible spectrum is disturbed by low illumination, high illumination and fog. (b) Benefits of RGB sources over thermal ones, where thermal spectrum is disturbed by glass and thermal crossover.

visible spectrum is limited under bad environmental conditions (e.g., low illumination, rain, haze and smog, etc.). ii) The thermal information can provide the complementary benefits for visible spectrum, especially in adverse illumination conditions. iii) The thermal sensors have many advantages over others, such as the long-range imaging ability, the insensitivity to lighting conditions and the strong ability to penetrate haze and smog. Fig. 1 shows some examples.

Most of RGB-T tracking methods focus on the sparse representation because of its capability of suppressing noises and errors [2,3,4]. These approaches, however, only adopt pixel intensities as feature representation, and thus be difficult to handle complex scenarios. Li et al. [5] extend the spatially ordered and weighted patch descriptor [6] to a RGB-T one, but this approach may be affected by the inaccurate initialization to their model. Deep learning based trackers [7,8,9] adopt powerful deep features or networks to improve tracking performance, but extending them to multi-modal ones has the following issues: i) Regarding thermal as one channel of RGB or directly concatenating their features might not make the best use of the complementary benefits from multiple modalities [4]. For example, if one modality is malfunction, fusing it equals to adding noises, which might disturb tracking performance [4]. ii) Designing multi-modal networks usually leads to the time-consuming procedures of network training and testing, especially for multiple input videos.

In this paper, we propose a novel cross-modal ranking algorithm for robust RGB-T tracking. Given one bounding box of the target object, we first partition it into non-overlapping patches, which are characterized by RGB and thermal features (such as color and gradient histograms). The bounding box can thus be represented with a graph with image patches as nodes. Motivated by [6,5], we assign each patch with a weight to suppress background information, and propose a cross-modal ranking algorithm to compute the patch weights. The patch weights are then incorporated into the RGB-T patch features, and the object location is finally predicted by applying the structured SVM [10]. Fig. 2 shows the pipeline of our approach. In particular, our cross-modal ranking algorithm advances existing ones in the following aspects.

First, we propose a general scheme for effective multimodal fusion. The RGB and thermal modalities are heterogeneous with different properties, and the hard consistency [11,4] between these two modalities may be difficult to perform effective fusion. Therefore, we propose a *soft cross-modality consistency* to enforce ranking consistency between modalities while allowing sparse inconsistency exists.

Second, we propose a novel method to mitigate the effects of ranking noises. In conventional manifold ranking models, the query quality is very important for ranking accuracy, and thus how to set good queries need to be designed manually [12,13,14]. In visual tracking, the setting of initial patch weights (i.e., queries) is not always reasonable due to noises of tracking results and irregular object shapes [6]. To handle this problem, we introduce an intermediate variable to represent the optimal labels of initial patches, and optimize it in a semi-supervised way based on the observation that *visually similar patches tend to have same labels or weights*. We formulate it as a l_1-optimization based sparse learning problem to promote sparsity of the inconsistency between inferred queries and initial ones (because most of the initial queries should be correct and the remaining ones are noises). We call this process as *optimal query learning* in this paper.

Finally, we present an efficient solver for the objective. Instead of individual consideration for each problem, we propose a single unified optimization framework to learn the patch weights and the optimal queries at a same time, which can be beneficial to boosting their respective performance. In particular, an efficient ADMM (alternating direction method of multipliers) [15] is adopted, and a linearized operation [16] is also employed to avoid matrix inversion for efficiency. By this way, our algorithm has a stable convergence behavior, and each iteration has small computational complexity.

In summary, we make the following contributions to RGB-T tracking and related applications. i) We integrate a soft consistency into the cross-modal ranking process to model the interdependency between two modalities while allowing sparse inconsistency exists to account for their heterogeneous properties. The proposed cross-modality consistency is general, and can be applied to other multimodal fusion problems. ii) To mitigate noise effects of initial patches, we introduce an intermediate variable to represent the optimal labels of the initial patches, and formulate it as a l_1-optimization based sparse learning problem. It is also general and applicable to other semi-supervized tasks, such as saliency detection and interactive object segmentation. iii) We present a unified ADMM-based optimization framework to solve the objective with stable and efficient convergence behavior, which makes our tracker very efficient. iv) To demonstrate the efficiency and superior performance of the proposed approach over the state-of-the-art methods, we conduct extensive experiments on two large-scale benchmark datasets, i.e., GTOT [4] and RGBT210 [5].

2 Related Work

The methods of visual tracking are vast, we only discuss the most related to us.

RGB-T tracking has drawn much attention in the computer vision community with the popularity and affordability of thermal infrared sensors [17]. Works on RGB-T tracking mainly focus on sparse representation because of its capability of suppressing noises and errors [2,3,18,4]. Wu et al. [2] concatenate the intensity features of image patches from RGB and thermal sources into a one-dimensional vector, which is sparsely represented in the target template space. The RGB-T tracking is performed in Bayesian filtering framework by defining reconstruction residues as the likelihood. Liu et al. [3] perform joint sparse representation on both RGB and thermal modalities, and fuse the resultant tracking results using min operation on the sparse representation coefficients. A Laplacian sparse representation is proposed to learn a multi-modal features using the reconstruction coefficients that encode both the spatial local information and occlusion handling [18]. Li et al. [4] propose a collaborative sparse representation based trackers to adaptively fuse RGB and thermal modalities by assigning each modality with a reliability weight. These approaches, however, only adopt pixel intensities as feature representation, and thus be difficult to handle complex scenarios. Kim et al. [6] propose a Spatially Ordered and Weighted Patch (SOWP) descriptor for target object based on the random walk algorithm, and achieve excellent performance for tracking. Li et al. [19] extend SOWP by optimizing a dynamic graph, and an another extension is further proposed to integrate multimodal information adaptively for RGB-T tracking [5].

Different from these works, we propose a novel cross-modal ranking algorithm for RGB-T tracking from a new perspective. In particular, our approach has the following advantages. i) **Generality**. The proposed model and schemes are general and applicable, including soft cross-modality consistency and optimal query learning, and can be easily extended to other vision problems. ii) **Effectiveness**. Our approach performs well against the state-of-the-art RGB and RGB-T trackers on two large-scale benchmark datasets. iii) **Efficiency**. The proposed optimization algorithm is with a fast and stable convergence behavior, which makes our tracker very efficient.

3 Cross-Modal Ranking Algorithm

Our cross-modal ranking algorithm aims to compute patch weights to suppress background effects in the bounding box description of target object. This section will introduce the details of our cross-modal ranking model and the associated optimization algorithm. The weighted patch feature construction and object tracking will be described in detail in the next section. For clarity, we present the pipeline of our tracking approach in Fig. 2.

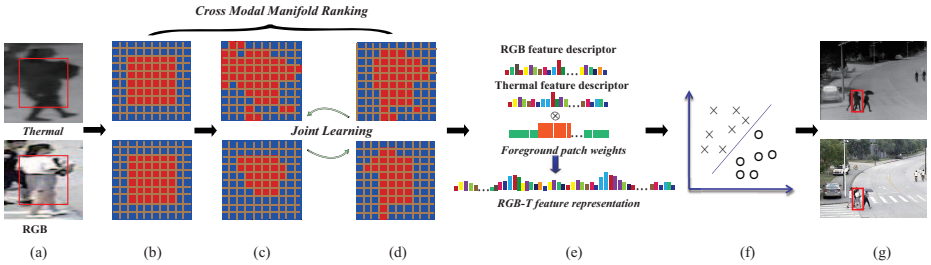

Fig. 2. Pipeline of our approach. (a) Cropped regions, where the red bounding box represents the region of initial patches. (b) Patch initilization indicated by red color. (c) Optimized results from initial patches. (d) Ranking results with the soft cross-modality consistency. (e) RGB-T feature representation. (f) Structured SVM. (g) Tracking results.

3.1 Model Formulation

The graph-based manifold ranking problem is described as follows: given a graph and a node in this graph as query, the remaining nodes are ranked based on their affinities to the given query. The goal is to learn a ranking function that defines the relevance between unlabelled nodes and queries [12]. We employ the graph-based manifold ranking model to solve our problem.

Given the target bounding box, we first partition it into a set of non-overlapping patches, which are described with RGB and thermal features (e.g., color, thermal and gradient histograms). To mitigate the effects of background information, we assign each patch with a weight that describes its importance belonging to target, and compute these weights via the cross-modal ranking algorithm. Given a patch feature set $\mathbf{X}^m = \{\mathbf{x}_1^m, ..., \mathbf{x}_n^m\}$, some patches are labelled as queries and the rest need to be ranked according to their affinities to the queries. Here, $m \in \{1, 2, ..., M\}$ indicates the m-th modality, and M denotes the number of modalities. Note that RGB-T data is the special case with $M = 2$, and we discuss its general form from the applicable perspective. Let $\mathbf{s}^m : \mathbf{X}^m \rightarrow \mathbb{R}^n$ denotes a ranking function which assigns a ranking value $\mathbf{s}_i^m$ to each patch $\mathbf{x}_i^m$ in the m-th modality, and $\mathbf{s}^m$ can be viewed as a vector $\mathbf{s}^m = [\mathbf{s}_1^m, ..., \mathbf{s}_n^m]^T$. In this work, we regard the initial patch weights as query labels, and $\mathbf{s}^m$ is thus a patch weight vector.

Let $\mathbf{q}^m = [\mathbf{q}_1^m, ..., \mathbf{q}_n^m]^T$ denote an indication vector, in which $\mathbf{q}_i^m = 1$ if $\mathbf{x}_i^m$ is target object patch, and $\mathbf{q}_i^m = 0$ if $\mathbf{x}_i^m$ is the background patch. $\mathbf{q}^m$ is computed by the initial ground truth (for the first frame) or tracking results (for the subsequent frames) as follows. For i-th patch, if it belongs to the shrunk region of the bounding box then $\mathbf{q}_i^m = 1$, and if it belongs to the expanded region of the bounding box then $\mathbf{q}_i^m = 0$, as shown in Fig. 3 (a). The remaining patches are non-determined, and will be diffused by other patches. In general, the ranking is performed in a two-stage way to account for background and objects, respectively [13], but we aim to integrate them in a unified model. To this end, we define an indication vector $\mathbf{\Gamma}$ that $\mathbf{\Gamma}_i = 1$ indicates that the

i-th patch is foreground or background patch, and $\mathbf{\Gamma}_i = 0$ denotes that the i-th patch is non-determined patch. Given the graph G^m of the m-th modality, through extending traditional manifold ranking model [12], the optimal ranking of queries are computed by solving the following optimization problem:

$$\min_{\{\mathbf{s}^m\}} \frac{1}{2} \sum_{m=1}^{M} \sum_{i,j=1}^{n} \mathbf{W}_{ij}^m || \frac{\mathbf{s}_i^m}{\sqrt{\mathbf{D}_{ii}^m}} - \frac{\mathbf{s}_j^m}{\sqrt{\mathbf{D}_{jj}^m}} ||^2 + \lambda ||\mathbf{\Gamma} \circ (\mathbf{s}^m - \mathbf{q}^m)||_F^2 + \frac{\lambda_2}{2} \|\mathbf{s}^m\|_F^2, \tag{1}$$

where λ is a parameter to balance the smoothness term and fitting term, and λ_2 is a regularization parameter. $\circ$ indicates the element-wise product. $\mathbf{D}^m$ is the degree matrix of the graph affinity matrix $\mathbf{W}^m$, whose computation is as follows. In the m-th modality, if graph nodes v_i and v_j are adjacent with 8-neighbors, they are connected by an edge e_{ij}, which is assigned a weight $\mathbf{W}_{ij}^m = \exp^{(-\gamma \|\mathbf{x}_i^m - \mathbf{x}_j^m\|)}$, where γ is the scaling parameter, which is set to 5 in this paper.

In (1), it inherently indicates that the available modalities are independent, which may significantly limit the performance in dealing with occasional perturbation or malfunction of individual sources. In addition, the settings of initial patch weights (i.e., queries) are not always reasonable due to noises of tracking results and irregular object shapes, as shown in Fig. 3 (a). In this paper, we integrate the *soft cross-modality consistency* and the *optimal query learning* into (1) to handle above problems, respectively.

Soft cross-modality consistency. To take advantage of the complementary benefits of RGB and thermal data, we need impose the modality consistency on the ranking process. Wang et al. [11] propose a multi-graphs regularized manifold ranking method to integrate different protein domains using hard constraints, i.e., employing multiple graphs to regularize the same ranking score. It is not suitable for our problem, as RGB and thermal sources are heterogeneous with different properties. Therefore, we introduce a *soft cross-modality consistency* to enforce ranking consistency between modalities while allowing sparse inconsistency exists to account for their heterogeneous properties. To this end, we propose the *soft cross-modality consistency* as a l_1-optimization based sparse learning problem as follows:

$$\min_{\{\mathbf{s}^m\}} \lambda_1 \sum_{m=2}^{M} ||\mathbf{s}^m - \mathbf{s}^{m-1}||_1 = \min_{\mathbf{s}^m} \lambda_1 ||\mathbf{C}\mathbf{S}||_1, \tag{2}$$

where λ_1 is a regularization parameter, and $\mathbf{S} = [\mathbf{s}^1; \mathbf{s}^2; ...; \mathbf{s}^M]$. $\mathbf{C}$ is the cross-modal consistency matrix, which is defined as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}^1 & -\mathbf{I}^2 & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{I}^2 & -\mathbf{I}^3 & & \\ & & \cdots & \cdots & \\ \mathbf{0} & & & \mathbf{I}^{M-1} & -\mathbf{I}^M \end{bmatrix}$$

where $\mathbf{I}$ is the identity matrix.

Optimal query learning. To mitigate noise effects of initial patch weights, we introduce an intermediate variable to represent the optimal ones, and optimize it in a semi-supervised way. The details are presented below.

Denoting the intermediate variable as $\hat{\mathbf{q}}^m = [\hat{\mathbf{q}}_1^m, ..., \hat{\mathbf{q}}_n^m]^T$, we first introduce two constraints for inferring $\hat{\mathbf{q}}^m$, i.e., *visual similarity constraint* and *inconsistency sparsity constraint*. The first constraint assumes that visually similar patches should have same labels and weights, and vice versa. Therefore, we add a smoothness term $\sum_{i,j=1}^{n} \mathbf{W}_{ij}^m (\hat{\mathbf{q}}_i^m - \hat{\mathbf{q}}_j^m)^2$ that can make visual similarity become a graph smoothness constraint. The second constraint aiming to compel sparsity in $\hat{\mathbf{q}}^m - \mathbf{q}^m$ is enlightened by the common use of l_1-norm sparsity regularization term in data noise, which has been proven to be effective even when the data noise is not sparse [20,21]. Therefore, we formulate it as $||\hat{\mathbf{q}}^m - \mathbf{q}^m||_1$, where l_1-norm is used to promote sparsity on the inconsistency between inferred labels and initial ones (because most of the initial labels should be correct and the remaining ones are noises). Fig. 3 shows the superiority of the l_1 norm over the l_2 norm. By combining these two constraints, the proposed l_1-optimization problem is formulated as follows:

$$\min_{\{\hat{\mathbf{q}}^m\}} \alpha \sum_{i,j=1}^{n} \mathbf{W}_{ij}^m (\hat{\mathbf{q}}_i^m - \hat{\mathbf{q}}_j^m)^2 + \beta ||\hat{\mathbf{q}}^m - \mathbf{q}^m||_1, \tag{3}$$

where α and β are the balance parameters. Integrating the *soft cross-modality consistency* (2) and the *optimal query learning* (3) into (1), the final cross-modal ranking model is written as:

$$\begin{aligned} \min_{\{\mathbf{s}^m\},\{\hat{\mathbf{q}}^m\}} & \frac{1}{2} \sum_{m=1}^{M} \Big(\sum_{i,j=1}^{n} \mathbf{W}_{ij}^m || \frac{\mathbf{s}_i^m}{\sqrt{\mathbf{D}_{ii}^m}} - \frac{\mathbf{s}_j^m}{\sqrt{\mathbf{D}_{jj}^m}} ||^2 + \lambda ||\mathbf{\Gamma} \circ (\mathbf{s}^m - \hat{\mathbf{q}}^m)||_F^2 \\ & + \frac{\lambda_2}{2} \|\mathbf{s}^m\|_F^2 + \alpha \sum_{i,j=1}^{n} \mathbf{W}_{ij}^m (\hat{\mathbf{q}}_i^m - \hat{\mathbf{q}}_j^m)^2 + \beta ||\hat{\mathbf{q}}^m - \mathbf{q}^m||_1 \Big) + \lambda_1 ||\mathbf{CS}||_1. \end{aligned} \tag{4}$$

Although (4) seems complex, as demonstrated in the experiments, the tracking performance is insensitive to parameter variations.

3.2 Optimization Algorithm

Although the variables of (4) are not joint convex, the subproblem to each variable with fixing others is convex and has a closed-form solution. The ADMM (alternating direction method of multipliers) algorithm [15] is efficient and effective solver for the problems like (4). To apply ADMM to our problem, we introduce two auxiliary variables $\mathbf{P} = \mathbf{CS}$ and $\mathbf{f}^m = \hat{\mathbf{q}}^m$ to make (4) separable. With some algebra, we have

$$\begin{aligned} \min_{\{\mathbf{s}^m\},\{\hat{\mathbf{q}}^m\},\mathbf{P},\{\mathbf{f}^m\}} & \sum_{m=1}^{M} ((\mathbf{s}^m)^T \mathbf{L}^m \mathbf{s}^m + \lambda ||\mathbf{\Gamma} \circ (\mathbf{s}^m - \hat{\mathbf{q}}^m)||_F^2 + \frac{\lambda_2}{2} \|\mathbf{s}^m\|_F^2 \\ & + 2\alpha (\mathbf{f}^m)^T (\mathbf{D}^m - \mathbf{W}^m) \mathbf{f}^m + \beta \|\hat{\mathbf{q}}^m - \mathbf{q}^m\|_1) + \lambda_1 \|\mathbf{P}\|_1, \\ s.t. \quad & \mathbf{P} = \mathbf{CS}, \mathbf{f}^m = \hat{\mathbf{q}}^m, \end{aligned} \tag{5}$$

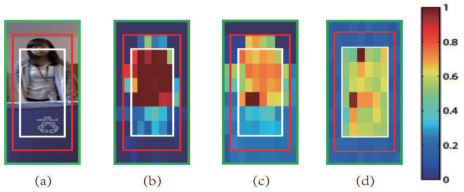

Fig. 3. Comparison of l_1-norm and l_2-norm in learning the optimal queries. (a) Target bounding box (red color), shrink bounding box (white color) and expand bounding box (green color). (b) Heatmap optimized by l_1-norm. (c) Heatmap optimized by l_2-norm. (d) Heatmap without optimal query learning. Herein, the heatmap represents the ranking results.

where $\mathbf{L}^m = \mathbf{I} - (\mathbf{D}^m)^{-\frac{1}{2}}\mathbf{W}^m(\mathbf{D}^m)^{-\frac{1}{2}}$ is the normalized Laplacian matrix of m-th modality. The augmented Lagrange function of (5) is:

$$\begin{aligned}
&\mathbb{L}(\{\mathbf{s}^m\},\{\hat{\mathbf{q}}^m\},\mathbf{P},\{\mathbf{f}^m\},\mathbf{Y}_1,\mathbf{Y}_2)\\
&=\sum_{m=1}^{M}((\mathbf{s}^m)^T\mathbf{L}^m\mathbf{s}^m+\lambda||\mathbf{\Gamma}\circ(\mathbf{s}^m-\hat{\mathbf{q}}^m)||_F^2+\frac{\lambda_2}{2}\|\mathbf{s}^m\|_F^2\\
&+2\alpha(\mathbf{f}^m)^T(\mathbf{D}^m-\mathbf{W}^m)\mathbf{f}^m+\beta\|\hat{\mathbf{q}}^m-\mathbf{q}^m\|_1)+\lambda_1\|\mathbf{P}\|_1\\
&+\frac{\mu}{2}(\|\mathbf{P}-\mathbf{CS}+\frac{\mathbf{Y}_1}{\mu}\|_F^2+\sum_{m=1}^{M}\|\hat{\mathbf{q}}^m-\mathbf{f}^m+\frac{\mathbf{y}_2^m}{\mu}\|_F^2)\\
&-\frac{1}{2\mu}(\|\mathbf{Y}_1\|_F^2+\|\mathbf{Y}_2\|_F^2),
\end{aligned}\tag{6}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2=[\mathbf{y}_2^1,\mathbf{y}_2^2,...,\mathbf{y}_2^M]$ are the Lagrangian multipliers, and μ is the Lagrangian parameter. Due to space limitation, we present the detailed derivations in the **supplementary file**. ADMM alternatively updates one variable by minimizing (6) with fixing other variables. Besides the Lagrangian multipliers, there are four variables, including $\mathbf{S}$, $\hat{\mathbf{q}}^m$, $\mathbf{P}$ and $\mathbf{f}^m$ to solve. Note that the $\mathbf{S}$-subproblem includes the inversion operation of a matrix with size of $Mn\times Mn$, which is time consuming. To handle this problem, we adopt a linearized operation [16] to avoid matrix inversion for efficiency. Due to space limitation, we only present the solutions of these subproblems as follows:

$$\begin{aligned}
&\mathbf{f}^m=(4\alpha(\mathbf{D}^m-\mathbf{W}^m)+\mu\mathbf{I})^{-1}(\mu\hat{\mathbf{q}}^m+\mathbf{y}_2^m)\\
&\hat{\mathbf{q}}^m=soft_thr_1(\mathbf{s}^m,\mathbf{f}^m-\frac{\mathbf{y}_2^m}{\mu},\mathbf{q}^m,\lambda\circ\mathbf{\Gamma}\circ\mathbf{\Gamma},\frac{\mu}{2},\beta)\\
&\mathbf{P}=soft_thr(\mathbf{CS}-\frac{\mathbf{Y}_1}{\mu},\frac{\lambda_1}{\mu})\\
&\mathbf{S}_{k+1}=\mathbf{S}_k-\frac{1}{\eta\mu}\nabla_{\mathbf{S}_k}J_k,
\end{aligned}\tag{7}$$

where $soft\_thr$ is a soft thresholding operator and $soft\_thr_1$ is also a soft thresholding operator with different inputs to $soft\_thr$, see the **supplementary file**

for detailed definitions. k indicates the k-th iteration, and J_k is the abbreviation of $J(\mathbf{S}_k, \hat{\mathbf{Q}}_k^m, \mathbf{P}_k, \mathbf{Y}_{1,k}, \mu_k) = \mathbf{S}_k^T\mathbf{L}\mathbf{S}_k + \lambda\|\mathbf{\Gamma} \circ (\mathbf{S}_k - \hat{\mathbf{Q}}_k)\|_F^2 + \frac{\mu_k}{2}\|\mathbf{P}_k - \mathbf{C}\mathbf{S}_k + \frac{\mathbf{Y}_{1,k}}{\mu_k}\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{S}_k\|_F^2$, where $\hat{\mathbf{Q}} = [\hat{\mathbf{q}}^1; \hat{\mathbf{q}}^2; ...; \hat{\mathbf{q}}^M]$, and

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}^1 & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{L}^2 & & \\ & & ... & \\ \mathbf{0} & & & \mathbf{L}^M \end{bmatrix}$$

$\nabla_{\mathbf{S}} J$ is the partial differential of J with respect to $\mathbf{S}$, and $\eta = \frac{1}{M}\sum_{m=1}^{M}\|\mathbf{X}^m\|_F^2$. Please refer to the **supplementary file** for the detailed derivations.

4 RGB-T Object Tracking

This section first imposes the optimized patch weights on the extracted multi-spectral features for more robust feature representation, and then present the tracker's details.

4.1 Feature Representation

We perform cross-modal ranking to obtain the patch weights, i.e., $\mathbf{s}^1, \mathbf{s}^2, ..., \mathbf{s}^M$. Let $\mathbf{x}_i = [\mathbf{x}_i^1; ...; \mathbf{x}_i^M] \in \mathbb{R}^{dM\times 1}$ be the RGB-T feature vector of i-th patch. Then, we construct the final collaborative feature representation by incorporating the patch weights. Specifically, for the i-th patch, we compute its final weight $\hat{\mathbf{s}}_i$ by combining all modal weights as follows:

$$\hat{\mathbf{s}}_i = \frac{1}{1 + \exp(-\sigma \frac{\sum_{m=1}^{M} \mathbf{s}_i^m}{M})}, \tag{8}$$

where σ is a scaling parameter fixed to 35 in this work. The collaborative feature representation is thus obtained by $\hat{\mathbf{x}} = [\hat{\mathbf{s}}_1\mathbf{x}_1; ...; \hat{\mathbf{s}}_n\mathbf{x}_n] \in \mathbb{R}^{dMn\times 1}$.

4.2 Tracking

We adopt the structured SVM (S-SVM) [10] to perform object tracking in this paper, and other tracking algorithm, such as correlation filters [22], can also be utilized.

Instead of using binary-labeled samples, S-SVM employs the structured sample that consists of a target bounding box and nearby boxes in the same frame to prevent the labelling ambiguity in training the classifier. Specifically, it constrains that the confidence score of an target bounding box y_t is larger than that of nearby box y by a margin determined by the intersection over union overlap ratio (denoted as $IoU(y_t, y)$) between two boxes:

$$\mathbf{h}^* = \arg\min_{\mathbf{h}} \ \xi||\mathbf{h}||^2 + \sum_{\mathbf{y}} \max\{0, \triangle(y_t, y) - \mathbf{h}^T\epsilon(y_t, y)\}, \tag{9}$$

where $\triangle(y_t, y) = 1 - IoU(y_t, y)$, $\epsilon(y_t, y) = \Psi(y_t) - \Psi(y)$, and $\xi = 0.0001$ is a regularization parameter. $\Psi(y_t)$ denotes the object descriptor representing a bounding box y_t at the t-th frame, and $\mathbf{h}$ is the normal vector of a decision plane. In this paper, we employ the stochastic variance reduced gradient (SVRG) technique [23] to optimize (9). By this way, S-SVM can reduce adverse effects of false labelling.

Given the bounding box of the target object in previous frame $(t-1)$, we first set a searching window in current frame t, and sample a set of candidates within the searching window. S-SVM selects the optimal target bounding box y_t^* in the t-th frame by maximizing a classification score:

$$y_t^* = \arg\max_{y_t} \ (\omega \mathbf{h}_{t-1}^T \Psi(y_t) + (1-\omega)\mathbf{h}_0^T \Psi(y_t)), \tag{10}$$

where ω is a balancing parameter, and $\mathbf{h}_{t-1}$ is the normal vector of a decision plane of $(t-1)$-th frame. $\mathbf{h}_0$ is learnt in the initial frame, which can prevent it from learning drastic appearance changes. To prevent the effects of unreliable tracking results, we update the classifier only when the confidence score of tracking result is larger than a threshold θ, where the confidence score of tracking result in t-th frame is defined as the average similarity between the weighted descriptor of the tracked bounding box and the positive support vectors: $\frac{1}{|\mathbb{V}_t|}\sum_{\mathbf{v}\in\mathbb{V}_t} \mathbf{v}^T \Psi(y_t^*)$, where $\mathbb{V}_t$ is the set of the positive support vectors at time t. In addition, we update object scales with three frames interval using the method from [24].

5 Performance Evaluation

5.1 Evaluation Settings

Data. There are only two large RGB-T tracking datasets, i.e., GTOT [4] and RGBT210 [5]. They are large and challenging enough, and we evaluate our approach on them for comprehensive validations. GTOT includes 50 RGB-T video clips with ground truth object locations under different scenarios and conditions. RGBT210 is another larger dataset for RGB-T tracking evaluation. It is highly-aligned, and contains 210 video clips with both RGB and thermal data. This dataset takes many challenges into consideration, such as camera moving, different occlusion levels, large scale variations and environmental challenges. The precision rate (PR) and success rate (SR) are employed to measure quantitative performance of various trackers.

Parameters. We fix all parameters and other settings in our experiments. We partition all bounding box into 64 non-overlapping patches to balance accuracy-efficiency trade-off [6], and extract RGB-T features for each patch, including color, thermal and gradient histograms, where the dimensions of gradient and each color channel are set to be 8. To improve the efficiency, each frame is scaled so that the minimum side length of a bounding box is 32 pixels, and the side length of a searching window is fixed to be $2\sqrt{WH}$, where W and H are the

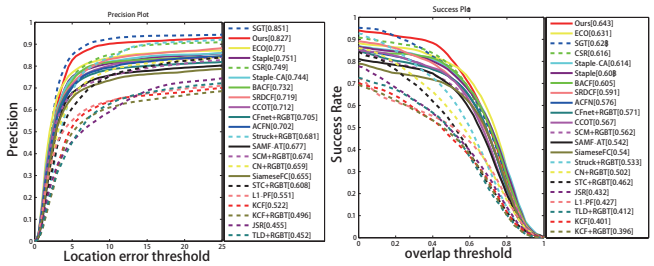

Fig. 4. Success Rate (SR) on the public GTOT benchmark dataset.

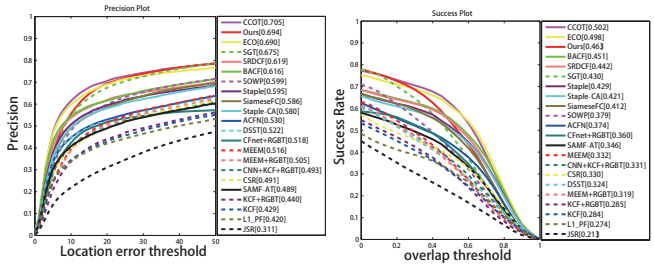

Fig. 5. The evaluation results on the public RGBT210 benchmark dataset. The representative score of PR/SR is presented in the legend.

width and height of the scaled bounding box, respectively. We shrink and expand the tracked bounding box (lx, ly, W, H) as $(lx+0.1W, ly+0.1H, 0.8W, 0.8H)$ and $(lx - W', ly - H', W + 2W', H + 2H')$, respectively, where (lx, ly) denotes the top-left coordinate of the tracked bounding box, and W' and H' indicate the patch width and height, respectively.

The proposed model involves several parameters in (6), including α, β, λ, λ_1 and λ_2, and the tracking sensitivity with different parameters are shown in Table 1. The results show the robustness of the proposed model to parameters' variations, and we set α, β λ, λ_1 and λ_2 to be 0.65, 0.002, 0.56, 0.3 and 0.4, respectively. In S-SVM, we empirically set $\{\omega, \theta\} = \{0.598, 0.3\}$, and employ a linear kernel.

Baselines. For comprehensive evaluation, we compare ours method with 23 popular trackers, some of which are from GTOT and RGBT210 benchmarks. Since there are few RGB-T trackers [2,3,4,18,5], we extend some RGB tracking methods to RGB-T ones by concatenating RGB and thermal features into a single vector or regarding the thermal as an extra channel, such as KCF [22], Struck [25], SCM [26] and CFnet [27]. In addition, we also select recently proposed state-of-the-art trackers for comparison, such as C-COT [9], ECO [28], ACFnet [29], SiameseFC [30] and Staple-CA [31], see Fig. 4 and Fig. 5 for details.

Table 1. Success Rate (SR) of the proposed method with different parameters on the GTOT dataset.

Param	Setting	SR	Param	Setting	SR	Param	Setting	SR
	0.325	0.624		0.0002	0.615			
α	0.65	0.643	β	0.002	0.643			
	1.3	0.59		0.02	0.633			
	0.28	0.62		0.15	0.628		0.2	0.602
λ	0.56	0.643	λ_1	0.3	0.643	λ_2	0.4	0.643
	1.12	0.605		0.6	0.628		0.8	0.632

Table 2. Attribute-based Precision Rate and Success Rate (PR/SR %) on RGBT210 dataset with 9 trackers, including CSR [4], DSST [32], MEEM [33], CNN [22], SOWP [6], KCF [22], SGT [5], CFnet [27] and ECO [28]. The best and second results are in red and green colors, respectively.

	ECO	SOWP	DSST	CSR	SGT	KCF+ RGBT	MEEM+ RGBT	CNN+KCF+ RGBT	CFnet+ RGBT	Ours
NO	87.7/64.3	75.0/46.1	70.2/41.4	68.1/45.2	82.4/50.7	56.6/36.3	64.7/41.2	63.7/42.9	69.7/52.2	86.1/59.4
PO	72.2/52.5	61.3/39.5	57.0/35.1	52.7/36.6	75.4/48.3	49.6/31.6	57.4/35.5	56.0/36.4	57.2/38.4	77.1/52.2
HO	58.3/41.3	52.0/32.8	39.4/25.7	37.1/24.3	53.1/34.1	33.0/22.2	37.2/24.2	36.6/25.9	39.3/27.3	54.3/34.6
LI	66.6/45.6	48.3/30.7	47.8/29.0	47.3/31.1	71.6/44.7	48.3/30.4	39.2/25.6	52.8/34.5	49.8/33.6	71.4/46.4
LR	64.1/38.1	51.0/29.1	52.8/29.1	46.0/23.1	65.8/37.5	42.6/26.2	44.9/23.4	54.6/32.5	45.2/27.7	64.8/37.4
TC	82.1/58.8	70.0/44.9	50.9/32.2	43.2/29.3	64.9/40.7	39.0/24.1	58.2/35.6	49.6/33.2	42.8/29.4	65.8/43.0
DEF	61.2/45.0	61.4/41.7	46.5/33.0	44.7/33.0	65.3/45.9	40.6/29.5	48.7/33.5	44.8/34.4	48.9/35.2	65.2/45.8
FM	58.2/39.2	56.0/32.3	34.4/21.2	42.6/25.0	58.0/33.1	33.3/19.1	43.5/26.8	37.1/24.1	36.5/23.0	58.8/34.9
SV	74.5/55.4	62.8/37.7	58.7/33.5	53.3/37.5	67.4/41.7	42.4/27.5	52.8/33.0	50.3/32.6	56.7/40.6	72.5/49.2
MB	67.8/49.9	55.2/38.3	32.3/23.2	34.7/23.8	58.6/39.6	29.1/20.7	46.2/31.4	30.4/22.0	30.3/22.4	58.4/40.5
CM	61.7/45.0	55.8/36.9	38.7/26.9	38.9/27.4	59.0/40.7	37.5/26.0	48.7/31.9	36.2/27.0	37.2/27.9	59.7/41.8
BC	52.9/35.2	47.2/28.6	43.8/26.3	38.4/23.7	58.6/35.5	41.0/25.6	40.5/23.4	42.3/28.4	43.7/28.1	57.9/35.2
ALL	69.0/49.8	59.9/37.9	52.2/32.4	49.1/33.0	67.5/43.0	44.0/28.5	50.5/31.9	49.3/33.1	51.8/36.0	69.4/46.3

5.2 Comparison Results

GTOT Evaluation. We present the evaluation results on the GTOT dataset in Fig. 4. Overall, the proposed algorithm performs favorably against the state-of-the-art methods. In particular, our approach outperforms the state-of-the-art methods using deep features with a clear margin, e.g., 5.0%/1.2% over ECO [28] and 11.5%/7.6% over C-COT [9] in PR/SR score. It is beneficial to the effective fusion of visible and thermal information in our method. Note that the methods based on deep features have weak performance on GTOT, including ECO and C-COT. It may be partly due to the weakness of deep features in representing the target objects with low resolution (many targets are small in GTOT). Our approach can handle this challenging factor. Fig. 4 shows that our tracker performs well against the state-of-the-art RGB-T methods, which suggest that the proposed fusion approach is effective. SGT [5] is better than our tracker in PR mainly due to adaptive fusion of different modalities by introducing modality weights, but performs weaker than ours in SR.

RGBT210 Evaluation. We further evaluate our method on the RGBT210 dataset in Fig. 5 and Table 2. The comparison curves show that our tracker also performs well against the state-of-the-art methods on RGBT210. In particular, our approach outperforms the state-of-the-art RGB-T tracking methods, e.g., 1.9%/3.3% over SGT [5] and 20.3%/13.3% over CSR [4] in PR/SR score. It justifies the effectiveness of the proposed method in fusing multimodal information

Table 3. PR/SR (%) of the proposed method with the different versions on the GTOT dataset.

	Ours-noC	Ours-no$\hat{q}$	Ours-noS	Ours
PR	78.7	78.0	71.1	82.7
SR	61.2	63.1	57.6	64.3

for visual tracking. For the state-of-the-art methods using deep features, the proposed tracker performs well against the SiameseFC [30] and CFnet [27] methods in all aspects. The proposed tracker performs equally well against the C-COT [9] and ECO [28] schemes in terms of PR and slightly worse in terms of SR. Furthermore, the proposed algorithm advances the C-COT and ECO methods in several aspects.

- It does not require laborious pre-training or a large training set, and also does not need to save a large pre-trained deep model. We initialize the proposed model using the ground truth bounding box in the first frame, and update it in subsequent frames.
- It is easy to implement as each subproblem of the proposed model has a closed-form solution.
- It performs favorably against the state-of-the-art deep tracking methods in terms of efficiency on a cheaper hardware setup (Ours: 8 FPS on 4.0GHz CPU, ECO: 8 FPS on 3.4GHz CPU and NVIDIA Tesla K40m GPU, C-COT: 1 FPS).
- It performs more robustly than the ECO and C-COT methods in some situations. In particular, it outperforms the ECO method on sequences with partial occlusion, low illumination, object deformation and background clutters in terms of PR and SR, which suggests the effectiveness of our approach in fusing the multimodal information and suppressing the background effects during tracking.

In addition, the example visual results on RGBT210 and GTOT are presented in the **supplementary file**, which further qualitatively verify the effectiveness of our method.

5.3 Ablation Study

To justify the significance of the main components, we implement 3 versions of our approach for empirical analysis on GTOT. The 3 versions are: 1) Ours-noC, that computes the patch weights without the constraint of cross-modal consistency. 2) Ours-no$\hat{q}$, that removes the optimal query learning operation in ranking model. 3) Ours-noS, that removes the patch weights in the feature presentation.

From the evaluation results reported in Table 3, we can draw the following conclusions. 1) The patch weights in collaborative object representation plays

critical roles in RGB-T tracking by observing that Ours outperforms Ours-noS. 2) The improvements of Ours over Ours-no$\hat{q}$ demonstrate the effectiveness of the introduced optimal query learning. 3) The soft consistency is important for cross-modal ranking from the observation that Ours-noC is much lower than Ours.

5.4 Runtime Performance

The experiments are carried out on a PC with an Intel i7 4.0GHz CPU and 32GB RAM, and implemented in C++. The proposed tracker performs at about 8 frames per second. In particular, our ranking algorithm converges within 30 iterations, and costs about 20 ms per frame (tested on all datasets). Note that our codes do not include any optimization and parallel operation, and the feature extraction and the structured SVM take most of time per frame (above 80%).

6 Conclusion

In this paper, we propose a graph-based cross-modal ranking algorithm to learn robust RGB-T object features for visual tracking. In the ranking process, we introduce the soft cross-modality consistency between modalities and the optimal query learning to improve the robustness. The solver to the proposed model is fast makes the tracker efficient. Extensive experiments on two large-scale benchmark datasets demonstrate the effectiveness and efficiency of the proposed approach against the state-of-the-art trackers.

However, our approach has the following two major limitations. First, the tracking performance is affected by the imaging limitation of some individual source, as shown in Table 2 (TC). Second, the runtime does not meet the demand of real-time applications. In future work, we will introduce the modality weights [4,5] in our model to address the first limitation, and implement our approach using parallel computation to improve the efficiency, such as multi-thread based multimodal feature extraction and GPU based structured SVM [34].

Acknowledgment

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61702002, 61472002, 61525306, 61633021, 61721004, 61420102015), Beijing Natural Science Foundation (4162058), Capital Science and Technology Leading Talent Training Project (Z181100006318030), 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. Cvejic, N., Nikolov, S.G., Knowles, H.D., Loza, A., Achim, A., Bull, D.R., Canagarajah, C.N.: The effect of pixel-level fusion on object tracking in multi-sensor surveillance video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2007)
2. Wu, Y., Blasch, E., Chen, G., Bai, L., Ling, H.: Multiple source data fusion via sparse representation for robust visual tracking. In: Proceedings of International Conference on Information Fusion. (2011)
3. Liu, H., Sun, F.: Fusion tracking in color and infrared images using joint sparse representation. Information Sciences **55**(3) (2012) 590–599
4. Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing **25**(12) (2016) 5743–5756
5. 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 International Conference on Multimedia. (2017)
6. Kim, H.U., Lee, D.Y., Sim, J.Y., Kim, C.S.: Sowp: Spatially ordered and weighted patch descriptor for visual tracking. In: Proceedings of IEEE International Conference on Computer Vision. (2015)
7. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: Proceedings of IEEE International Conference on Computer Vision. (2015)
8. Wang, L., Ouyang, W., Wang, X., Lu, H.: Visual tracking with fully convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision. (2015)
9. Danelljan, M., Robinson, A., Khan, F.S., Felsberg, M.: Beyond filters: Learning continuous convolution operators for visual tracking. In: Proceedings of European Conference on Computer Vision. (2016)
10. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research **6** (2005) 1453–1484
11. Wang, J., Bensmail, H., Gao, X.: Multiple graph regularized protein domain ranking. In: BMC Bioinformatics. (2012)
12. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Scholkopf, B.: Ranking on data manifolds. In: Proceedings of Neural Information Processing Systems. (2004)
13. Zhang, L., Yang, C., Lu, H., Ruan, X., Yang, M.H.: Ranking saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(9) (2017) 1892–1904
14. Wang, L., Lu, H., Yang, M.H.: Constrained superpixel tracking. IEEE Transactions on Cybernetics **48**(3) (2017) 1030–1041
15. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3** (2011) 1–122
16. Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low rank representation. In: Proceedings of Annual Conference on Neural Information Processing Systems. (2011)
17. Gade, R., Moeslund, T.B.: Thermal cameras and applications: a survey. Machine Vision and Applications **25** (2014) 245–262

18. Li, C., Hu, S., Gao, S., Tang, J.: Real-time grayscale-thermal tracking via laplacian sparse representation. In: Proceedings of International Conference on Multimedia Modeling. (2016)
19. Li, C., Lin, L., Zuo, W., Tang, J.: Learning patch-based dynamic graph for visual tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2017) 4126–4132
20. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(2) (2009) 210–227
21. Fu, Y., Hospedales, T.M., Xiang, T., Xiong, J., Gong, S., Wang, Y., Yao, Y.: Robust subjective visual property prediction from crowdsourced pairwise labels. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(3) (2016) 563–577
22. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized filters. IEEE Transactions on Pattern Analysis and Machine Intelligence (2015)
23. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Proceedings of Annual Conference on Neural Information Processing Systems. (2013)
24. Ma, C., Yang, X., Zhang, C., Yang, M.H.: Long-term tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2015)
25. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: Proceedings of IEEE International Conference on Computer Vision. (2011)
26. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2012)
27. Valmadre, J., et al.: End-to-end representation learning for correlation filter based tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2017)
28. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2017)
29. Choi, J., Chang, H.J., Yun, S., Fischer, T., Demiris, Y., Choi, J.Y., et al.: Attentional filter network for adaptive visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
30. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. arXiv preprint arXiv:1606.09549 (2016)
31. Mueller, M., Smith, N., Ghanem, B.: Context-aware correlation filter tracking. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
32. Danelljan, M., Hager, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: Proceedings of British Machine Vision Conference. (2014)
33. Zhang, J., Ma, S., Sclaroff, S.: MEEM: robust tracking via multiple experts using entropy minimization. In: Proceedings of European Conference on Computer Vision. (2014)
34. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(10) (2016) 2096–2109