Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking

Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin, *Senior Member, IEEE*

***Abstract*— Integrating multiple different yet complementary feature representations has been proved to be an effective way for boosting tracking performance. This paper investigates how to perform robust object tracking in challenging scenarios by adaptively incorporating information from grayscale and thermal videos, and proposes a novel collaborative algorithm for online tracking. In particular, an adaptive fusion scheme is proposed based on collaborative sparse representation in Bayesian filtering framework. We jointly optimize sparse codes and the reliable weights of different modalities in an online way. In addition, this paper contributes a comprehensive video benchmark, which includes 50 grayscale-thermal sequences and their ground truth annotations for tracking purpose. The videos are with high diversity and the annotations were finished by one single person to guarantee consistency. Extensive experiments against other state-of-the-art trackers with both grayscale and grayscale-thermal inputs demonstrate the effectiveness of the proposed tracking approach. Through analyzing quantitative results, we also provide basic insights and potential future research directions in grayscale-thermal tracking.**

***Index Terms*— Collaborative sparse representation, multi-task modeling, grayscale-thermal tracking benchmark, adaptive tracking.**

I. INTRODUCTION

DESPITE significant progress, visual object tracking using visible spectrum camera remains a very challenging task in some complex scenarios, such as low illumination, background clutters, as well as bad weathers (rain, haze, smog, etc.). Fortunately, these factors can be addressed by leveraging the complementary benefits of fusing other modalities [1]. For example, the depth sensors can provide valuable additional depth information to substantially improve tracking results by robust occlusion and model drift handling. However, these sensors suffer from the limited range (e.g., 4-5 meters at most).

Thermal infrared cameras, as a kind of passive sensors, can capture infrared radiation emitted by subjects with a temperature above absolute zero. This type of cameras, originally developed for military use (e.g., surveillance during night), has recently become more economically affordable and thus applied to wide applications [2], [3]. On one hand, these sensors are more effective than visible spectrum cameras under poor lighting conditions, and can also overcome the above-mentioned limitation of depth sensors. On the other hand, visible spectrum cameras are more effective while separating two moving subjects are crossing or moving side (or called “thermal crossover” [4]). Therefore, grayscale and thermal data can complement information to each other to achieve robust tracking performance in challenging scenarios [5]–[7].

Focusing on a collaborative model and a comprehensive evaluation benchmark of grayscale-thermal tracking,[1] we address following remaining issues in this paper through existing works.

- How to achieve robust tracking by adaptively exploiting grayscale and thermal information based on their reliability. Previous works [1], [8] adopted simple weight schemes to achieve adaptive grayscale-thermal tracking, which might easily fail in challenging scenarios. Bunyak et al. [9] employed thermal information to assist grayscale tracking as it is less sensitive to illumination variations and shadows. When thermal information is unreliable, this kind of methods will lead to poor performance. Recent joint sparse representation methods [4], [10] presented promising results in grayscale-thermal tracking. However, these methods ignored the reliabilities of different modalities in sparse representation.
- How to create a comprehensive grayscale-thermal tracking benchmark to facilitate the research of this direction. The related research is presently limited by the lack of a comprehensive video benchmark. Existing grayscale-thermal video datasets, like OSU Color-Thermal [5] and LITIV [6], [7], contain small number of videos and induce significant bias [11]. In particular, OSU Color-Thermal [5] and LITIV [6], [7] datasets contain 6, 9 and 4 grayscale-thermal video sequences, respectively.

Manuscript received September 6, 2015; revised March 23, 2016 and June 22, 2016; accepted September 12, 2016. Date of publication September 27, 2016; date of current version October 18, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61472002 and Grant 61370185, in part by the Guangdong Natural Science Foundation under Grant S2013010013432 and Grant S2013010015940, in part by the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase), and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vishal Monga. *(Corresponding author: Jin Tang.)*

C. Li and J. Tang are with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com; tj@ahu.edu.cn).

H. Cheng, S. Hu, and L. Lin are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China, and also with the Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China (e-mail: chengh9@mail.sysu.edu.cn; hushiyi@mail2.sysu.edu.cn; linliang@ieee.org).

X. Liu is with the Department of Computer Science, San Diego State University, San Diego, CA 92182 USA (e-mail: xiaobai.liu@mail.sdsu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2614135

[1] The grayscale-thermal object tracking (GTOT) benchmark: http://hcp.sysu.edu.cn/resources/.

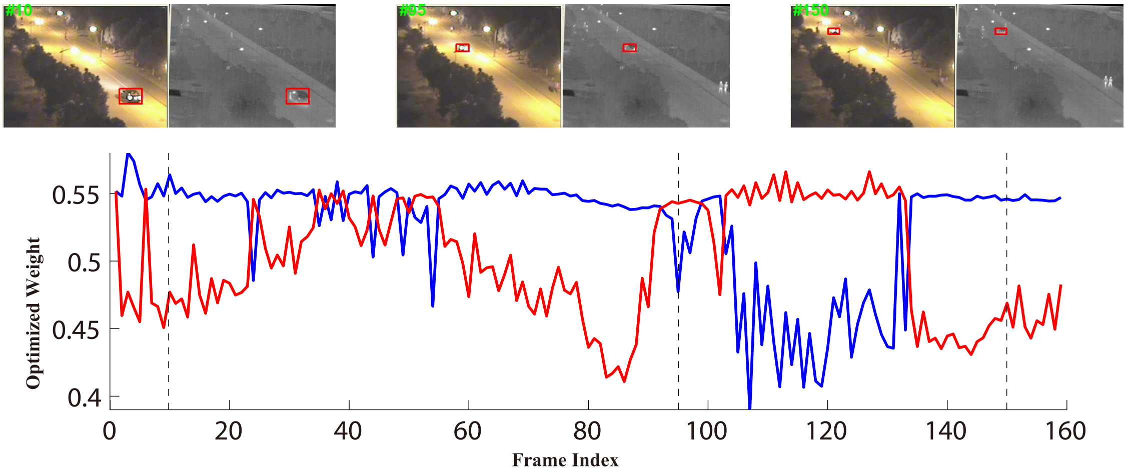


Fig. 1. Illustration of the optimized weights based on the modal reliabilities. The blue and red curves indicate the weights of grayscale and thermal sources, respectively. One can see that the optimized weights of grayscale-thermal sources are almost consistent with their reliabilities, and thus our method can achieve robust tracking even though one modality has occasional perturbation or malfunction.

Sparse representation, directly related to compressed sensing [12], has recently attracted much attention in visual tracking [13]–[16]. Liu and Sun [4] and Wu et al. [10] also employed the idea of sparse representation to grayscale-thermal tracking. More specifically, Wu et al. [10] concatenated the image patches from grayscale and thermal sources into a one-dimensional vector that is then sparsely represented in the target template space. Liu and Sun [4] performed joint sparse representation calculation on both grayscale and thermal modalities and fused the resultant tracking results using min operation on the sparse representation coefficients. These trackers have two main weaknesses: 1) They treated each modality equally, and thus may significantly limit the tracking performance in dealing with occasional perturbation or malfunction of individual sources. 2) They calculated the observation likelihood of each candidate in whole dictionary but ignored the discriminative information between object and background.

In this paper, we propose a novel adaptive tracking method based on collaborative sparse representation within the Bayesian filtering framework. For each modality, we employ the idea of sparse representation for the robustness against appearance contaminations inherited from the previously sparse trackers in grayscale or thermal videos [13], [17]. Unlike assuming that available modalities contribute equally in [4] and [10], we pursue a collaborative sparse representation for adaptive object tracking by introducing the weight variables to represent modal reliabilities, and jointly optimize the sparse codes and the reliable weights of different modalities in an online way. Collaborative sparse representation is capable of dealing with occasional perturbation or malfunction of individual sources, as shown in Fig. 1. Considering the sparse representation in each modality as an individual task, our method is essentially formulated as a multi-task learning problem. Finally, given the motion model, an object is located by maximizing the discriminative likelihood based on the proposed collaborative sparse representation in Bayesian filtering framework.

It is worth mentioning that our method has the following two characters: i) it is capable of collaboratively integrating grayscale and thermal information by jointly optimizing the sparse codes and reliable weights, and thus maintains the persistence of online tracking in challenging scenarios; ii) it derives discriminative likelihood from the proposed collaborative sparse representation for Bayesian filtering, which allows uncertainty reasoning over both grayscale data and thermal data.

In addition, we build a new grayscale-thermal object tracking (GTOT) benchmark and release it for evaluating tracking methods. The GTOT dataset includes 50 video pairs, each consisting of a grayscale video and a thermal video.[2] As a comprehensive platform, the benchmark provides statistics bias analysis, annotations of visual trajectories, evaluation metrics, and baseline methods with both grayscale and grayscale-thermal inputs.

This paper makes the following contributions to grayscale-thermal tracking and related applications.

- It proposes a novel adaptive tracking approach that can effectively integrate grayscale and thermal information based on the collaborative sparse representation in Bayesian filtering framework. In particular, we jointly optimize the sparse codes and the reliable weights of different modalities in the collaborative sparse representation.
- It creates one unified grayscale-thermal benchmark, including: 1) 50 video pairs with analysis of bias statistics, 2) ground truth annotations of every frames in both grayscale and thermal videos, which are finished by one person for consistency, 3) two evaluation metrics, and 4) two kinds of baseline methods. This benchmark will be available online for free academic usage.
- It presents extensive experiments against other state-of-the-art trackers with both grayscale and grayscale-thermal inputs. The evaluation results demonstrate the

[2] The GTOT webpage: http://hcp.sysu.edu.cn/resources/.

effectiveness of the proposed approach. Through analyzing quantitative results, we further provide basic insights and identify the potentials of thermal information in grayscale-thermal tracking.

The rest of this paper is organized as follows. In Section II, the relevant methods to our works are introduced. In Section III, we describe the details of the proposed approach. The new grayscale-thermal benchmark is presented and analyzed in Section IV, and the experimental results are shown in Section V. The final Section VI concludes this paper.

II. Related Work

This work is closely related to the advances in three research streams and the development of visual benchmarks.

Single-modality object tracking has been extensively studied in computer vision community. The most successful stories come from the application of the various machine learning techniques, including SVM [18], boosting [19], sparse representation [16], subspace learning [20], [21], metric learning [22] and correlation filter [23], [24]. Most of these methods, however, only work on grayscale data and thus suffer from the aforementioned challenges, *i.e.*, low-illumination, bad weathers, *etc*.

Multi-modality tracking has drawn a lot of attentions in the community with the popularity of various sensors, e.g., depth sensors [25], and thermal infrared sensors [1]. Conaire et al. [1] evaluated appearance model tracking performance of multiple different fusion schemes on manually annotated multi-modal surveillance videos. Bunyak et al. [9] presented a moving object detection and tracking system that fused grayscale and thermal videos within a level set framework. Conaire et al. [8] proposed a framework that can efficiently combine features for robust tracking, and instantiated the fusion of thermal infrared and visible spectrum features in this framework for automated surveillance applications. The impact of pixel-level fusion of videos from grayscale-thermal surveillance cameras was investigated by Cvejic et al. [26] to compare to tracking in single modality videos. This tracker had been accomplished by means of a particle filter which fuses a color cue and the structural similarity measure. A pedestrian tracker designed as a particle filter was introduced by Leykin and Hammoud [27] based on the proposed background model, in which each pixel was represented as a multi-modal distribution with the changing number of modalities for both grayscale and thermal input. Wu et al. [10] concatenated the image patches from grayscale and thermal sources into a one-dimensional vector that is then sparsely represented in the target template space. Liu and Sun [4] performed joint sparse representation calculation on both grayscale and thermal modalities and fused the resultant tracking results using min operation on the sparse representation coefficients. In contrast, we present a multi-task collaborative sparse representation for online tracking in this work.

Sparse representation was widely applied to visual tracking and achieved impressive results because of its capability of suppressing noises and errors [13]–[16], [28]. Mei and Ling [13] first proposed a sparse representation based tracker to handle the corrupted appearance, and recently it has been further improved [14]–[16]. Zhang et al. [14] constructed a multi-task sparse learning method denoted as multi-task tracking by employing the concept of sparse representation based on a particular framework. Bao et al. [15] proposed a fast real time l_1-tracker called the APG-l_1 tracker, which utilized the accelerated proximal gradient algorithm to improve the efficiency. Zhong et al. [16] developed a sparsity-based collaborative model for object tracking. The collaborative model combined a sparsity-based classifier learned from holistic templates and a sparsity-based generative model generated from local representations. We extend in this work the collaborative model with a Bayesian filter technique for robust seasoning.

There has been several *grayscale-thermal video datasets* for the various vision tasks. For example, OSU Color-Thermal dataset [5] contains six grayscale-thermal video sequence pairs recorded from two different locations with only people moving, which is obviously not sufficient to evaluate grayscale-thermal tracking algorithms. Other two grayscale-thermal datasets are collected by Torabi et al. [6], Bilodeau et al. [7]. Most of them suffer from their limited size, low diversity and high bias. This work addresses these issues and creates a reasonable size grayscale-thermal video dataset that provides comprehensive benchmark.

III. Adaptive Tracking via Collaborative Sparse Representation

In this section, we introduce an adaptive grayscale-thermal object tracking method in Bayesian filtering framework.

A. Overview of Our Approach

Our tracking method utilizes Bayesian filtering technique for uncertainty reasoning. To define the likelihood function, we represent each object of interest as a set of basis, i.e., object templates, and adaptively update the template set on the fly. To track an object over time, we measure each candidate region by how well it could be sparsely reconstructed from the template basis. For grayscale-thermal tracking, we maintain a template set for each modality. This is actually a multi-task learning problem if we consider the reconstruction of target region in each modality as an individual task.

When the target is occasional perturbation or malfunction in one modality, other modalities can complement information to avoid model drift in tracking process. Therefore, we introduce a weighted variable for each modality, and make it optimized independently.

An object template used consists of two types of basis for each modality: one from the tracked foreground regions, and another from the surrounding background regions. We refer them to *positive basis* and *negative basis*, respectively. Thus, the likelihood function is defined by a discriminative likelihood score based on the reconstruction residual by positive basis and the reconstruction residual by negative basis. This discriminative likelihood is capable of capturing the surrounding contrast information as the target is moving in the scene.

B. Review: Bayesian Filtering

We first review the Bayesian filtering based tracking method. Let $\mathbf{Z}_t^{[M]} = [\mathbf{z}_1^{[M]}, \mathbf{z}_2^{[M]}, \ldots, \mathbf{z}_t^{[M]}]$ denote the observation set generated from M different modalities, where the operator $[M]$ indicates the set of integers between 1 and M: $[M] = \{1, 2, \ldots, M\}$, *e.g.*, $\mathbf{z}_t^{[M]} = \{\mathbf{z}_t^1, \mathbf{z}_t^2, \ldots, \mathbf{z}_t^M\}$. grayscale-thermal data used in this paper is the special case with $M = 2$. Given $\mathbf{Z}_t^{[M]}$ and the state variable $\mathbf{x}_t$, we can compute the optimal state $\hat{\mathbf{x}}_t$ by Maximum A Posterior (MAP) estimation,

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}_{t,i}} P(\mathbf{x}_{t,i}|\mathbf{Z}_t^{[M]}), \tag{1}$$

where $\mathbf{x}_{t,i}$ indicates the state of the i-th sample at time t. We factorize Eq. (1) by Bayesian rules as follows,

$$P(\mathbf{x}_t|\mathbf{Z}_t^{[M]}) \propto P(\mathbf{z}_t^{[M]}|\mathbf{x}_t) \int P(\mathbf{x}_t|\mathbf{x}_{t-1}) P(\mathbf{x}_{t-1}|\mathbf{Z}_{t-1}^{[M]}) d\mathbf{x}_{t-1}, \tag{2}$$

where $P(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $P(\mathbf{z}_t^{[M]}|\mathbf{x}_t)$ are the motion model and the likelihood model, respectively. We utilize six independent affine parameters, including deformable and translated information, to represent the variation of motion, and model the dynamic process by the Gaussian distribution,

$$P(\mathbf{x}_t|\mathbf{x}_{t-1}) = N(\mathbf{x}_t; \mathbf{x}_{t-1}, \sigma_p), \tag{3}$$

where σ_p denotes a diagonal covariance matrix whose elements are the variations of the affine parameters, and its setting depends on motion variations of the target object. Based on the current tracking result, we predict a set of candidate regions in the next frame according to the motion model. The likelihood term $P(\mathbf{z}_t^{[M]}|\mathbf{x}_t)$ will be defined in a discriminative way to measure the confidence of each candidate object region.

C. Collaborative Sparse Representation

We describe an object of interest with two template bases for each modality, one is extracted from the foreground regions, and another from the surrounding background regions. Let $\mathbf{Y}^m = [\mathbf{y}_1^m, \mathbf{y}_2^m, \ldots, \mathbf{y}_K^m] \in R^{d \times K}$ denote the candidate set, where $m = 1, \ldots, M$, d and K denote the feature dimension and the number of candidates, respectively. The positive and negative template bases are denoted by $\mathbf{T}_{pos}^{[M]} = [\mathbf{t}_1^{[M]}, \mathbf{t}_2^{[M]}, \ldots, \mathbf{t}_p^{[M]}]$ and $\mathbf{T}_{neg}^{[M]} = [\mathbf{t}_{p+1}^{[M]}, \mathbf{t}_{p+2}^{[M]}, \ldots, \mathbf{t}_{p+q}^{[M]}]$, where p and q indicate the number of positive and negative templates. Thus, we construct M dictionaries as $\mathbf{T}^{[M]} = [\mathbf{T}_{pos}^{[M]}, \mathbf{T}_{neg}^{[M]}]$, and reconstruct the candidate samples as follows:

$$\begin{aligned} \min_{\mathbf{C}, \alpha^{[M]}} & \sum_{m=1}^{M} \frac{(\alpha^m)^s}{2} \|\mathbf{Y}^m - \mathbf{T}^m \mathbf{C}^m\|_F^2 + \lambda \|\mathbf{C}\|_{2,1} \\ & + \sum_{m=1}^{M} (\phi^m (\alpha^m)^s + (1 - \alpha^m)^s), \\ s.t.\ & \mathbf{C} \succeq 0, \alpha^{[M]} \succeq 0, \end{aligned} \tag{4}$$

Algorithm 1 Optimization Procedure to Eq. (4)

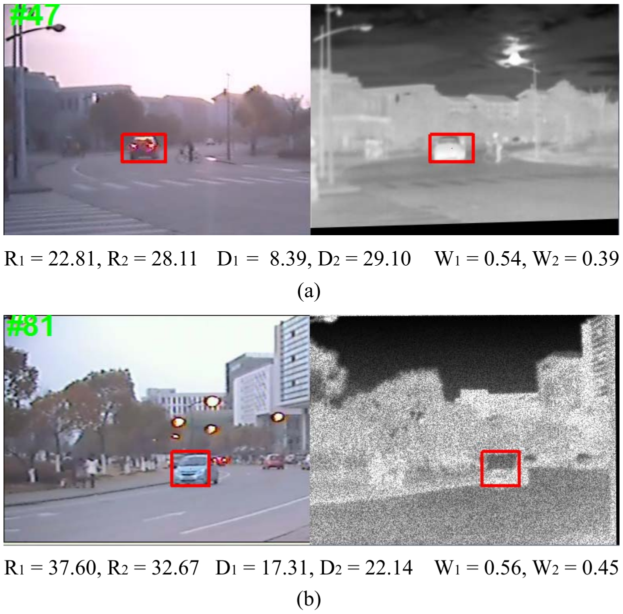
```
Input: The template set $\mathbf{T}^{[M]}$, the candidate set $\mathbf{Y}^{[M]}$;
    Set $\alpha^m = \frac{1}{M}$ $(m = 1, 2, ..., M)$;
    $\varepsilon = 10^{-15}$, $maxIter = 3$.
Output: $\mathbf{C}$, $\alpha^{[M]}$.
for $k = 1 : maxIter$ do
    Update $\mathbf{C}$ by solving Eq. (5);
    if $k == 1$ then
        $\beta = \max_m(\|\mathbf{Y}^m - \mathbf{T}^m\mathbf{C}^m\|_F^2) * 2$;
    end if
    Update $\alpha^{[M]}$ by Eq. (11);
    if $|J_k - J_{k-1}| < \varepsilon$ then
        Terminate the loop.
    end if
end for
```

where $\mathbf{C} = [\mathbf{C}^1; \mathbf{C}^2; \ldots; \mathbf{C}^M]$, and $\|\mathbf{C}\|_{2,1}$ encourages that one candidate shares the same pattern across different modalities. $\mathbf{T}^m$ is the dictionary of the m-th modality, which consists of both positive and negative basis templates of the m-th modality. α^m is the reliable weight for the m-th modality, and $s \in (1, \infty)$ is a fuzzy parameter [29], [30]. λ is a balance parameter. The nonegative constraint $\mathbf{C} \succeq 0$ is used to avoid the meaningless negative elements, as each element of $\mathbf{C}$ represents the similarity between the corresponding candidate and template. The last term of Eq. (4) is the regularization of $\alpha^{[M]}$, which avoids a degenerate solution of $\alpha^{[M]}$ while allowing them to be specified independently. $\phi^{[M]}$ are the adaptive parameters which can be set manually if modal discriminative abilities are known beforehand, and also be updated online to adapt its variations. In implementation, we empirically update it online as $\phi^m = \beta \exp(-\frac{|\max(\hat{\mathbf{C}}_{opt-pos}^m) - \max(\hat{\mathbf{C}}_{opt-neg}^m)|}{\sigma_\phi})$, $m = 1, 2, \ldots, M$. Herein, β is a balance parameter initialized based on the reconstruction errors after the first iteration (see Algorithm 1 for details), and σ_ϕ is adaptively set to be $\max_{m=1}^{M}(|\max(\hat{\mathbf{C}}_{opt-pos}^m) - \max(\hat{\mathbf{C}}_{opt-neg}^m)|)$. $\hat{\mathbf{C}}_{opt-pos}^m$ and $\hat{\mathbf{C}}_{opt-neg}^m$ represent the previous positive and negative coefficients of the optimal candidate on template basis, respectively.

$\alpha^{[M]}$ in Eq. (4) can be adjusted online based on the modal reliabilities due to the following reasons. 1) In one modality, the reconstruction error can measure all candidate regions by how well it could be sparsely reconstructed from the template basis. Therefore, the qualities of different modalities can be reflected by their respective reconstruction errors. From the first term in Eq. (4) one can see that our method places larger weights on those modalities which have smaller reconstruction errors, resulting in a quality-aware weight optimization. Fig. 2 (a) shows one sample of this situation. 2) In some challenging scenarios, reconstruction errors are not enough to represent modal reliabilities for tracking task, as shown in Fig. 2 (b). To obtain more reliable weights, we introduce the adaptive parameters defined by the computed sparse codes to represent the modal discriminative abilities. This regularization leads to a discrimination-aware weight optimization. Fig. 2 justifies the effectiveness of the adaptive parameters.

D. Problem Optimization

Though the optimization problem in Eq. (4) is not jointly convex in $\mathbf{C}$ and $\alpha^{[M]}$, we can utilize the alternate strategy to optimize one variable with another fixed.



R1 = 22.81, R2 = 28.11 D1 = 8.39, D2 = 29.10 W1 = 0.54, W2 = 0.39

(a)

R1 = 37.60, R2 = 32.67 D1 = 17.31, D2 = 22.14 W1 = 0.56, W2 = 0.45

(b)

Fig. 2. Two typical samples with the optimized weights are shown in (a) and (b), respectively. R, D and W denote the reconstruction error, the adaptive parameter and the optimized weight, respectively. Subscript 1 is for grayscale video and 2 for thermal video.

Given fixed $\alpha^{[M]}$, Eq. (4) can be rewritten as

$$\min_{\mathbf{C}} \sum_{m=1}^{M} \frac{(\alpha^m)^s}{2} \|\mathbf{Y}^m - \mathbf{T}^m \mathbf{C}^m\|_F^2 + \lambda \|\mathbf{C}\|_{2,1} + \psi(\mathbf{C}), \quad (5)$$

where $\psi(\mathbf{C})$ is defined as

$$\psi(\mathbf{C}) = \begin{cases} 0, & if\ \mathbf{C}_{ij} \geq 0 \\ \infty, & else \end{cases}, \quad (6)$$

Eq. (5) can be efficiently solved by applying the Accelerated Proximal Gradient (APG) approach [31]. We denote

$$\begin{aligned} F(\mathbf{C}) &= \sum_{m=1}^{M} \frac{(\alpha^m)^s}{2} \|\mathbf{Y}^m - \mathbf{T}^m \mathbf{C}^m\|_F^2, \\ G(\mathbf{C}) &= \|\mathbf{C}\|_{2,1} + \psi(\mathbf{C}). \end{aligned} \quad (7)$$

Let initial $\mathbf{C}$ be zero, APG utilizes the following update equations at the k-th iteration:

$$\begin{aligned} \mathbf{A}^{k+1} &= \mathbf{C}^k + \rho^k (\mathbf{C}^k - \mathbf{C}^{k-1}), \\ \mathbf{C}^{k+1} &= prox_{\lambda, t^k, G}(\mathbf{A}^{k+1} - t^k \triangledown F(\mathbf{A}^{k+1})), \end{aligned} \quad (8)$$

where t^k denotes constant step size, updated using a line search algorithm [31], and the extrapolating parameter ρ^k is set to be $\frac{k}{k+3}$. The associated proximal optimization problem is defined as follows:

$$\begin{aligned} prox_{\lambda, G}(\mathbf{V}) &= \arg\min_{\mathbf{U}} \ G(\mathbf{U}) + \frac{1}{2\lambda} ||\mathbf{U} - \mathbf{V}||_F^2 \\ &= \max(\mathbf{0}, \ \arg\min_{\mathbf{U}} \ ||\mathbf{U}||_{2,1} + \frac{1}{2\lambda} ||\mathbf{U} - \mathbf{V}||_F^2). \end{aligned} \quad (9)$$

This paper employs the Sparse Modeling Software [32] to solve the proximal step in Eq. (9).

Given $\mathbf{C}$, Eq. (4) can be written as

$$\begin{aligned} \min_{\alpha^{[M]}} \sum_{m=1}^{M} & ((\alpha^m)^s (\frac{\|\mathbf{Y}^m - \mathbf{T}^m \mathbf{C}^m\|_F^2}{2} + \phi^m) \\ & + (1 - \alpha^m)^s), \ s.t. \ \alpha^{[M]} \geq 0, \end{aligned} \quad (10)$$

which has a closed-form solution:

$$\alpha^m = \frac{1}{1 + (\phi^m + \frac{\|\mathbf{Y}^m - \mathbf{T}^m \mathbf{C}^m\|_F^2}{2})^{\frac{1}{s-1}}}, \quad (11)$$

where $m = 1, 2, \ldots, M$. The whole optimization procedure is summarized in Algorithm 1, in which J_k denotes the k-th objective value of Eq. (4).

E. Discriminative Likelihood

For each candidate region, we utilize the reconstruction residues from both positive basis and negative basis to define its likelihood, *e.g.*, $P(\mathbf{y}_i|\mathbf{x}_t)$ for i-th candidate $\mathbf{y}_i$. In contrast, the past sparse representation based tracking algorithms usually employ positive template only [13]–[15]. In particular, their methods are likely to drift when target appearance is similar to the background [16].

For i-th candidate, we obtain the constructed errors on the positive and negative templates of m-th modal as

$$\begin{aligned} e_{i-pos}^m &= \|\mathbf{Y}_i^m - \mathbf{T}_{pos}^m \mathbf{C}_{i-pos}^m\|_F^2, \\ e_{i-neg}^m &= \|\mathbf{Y}_i^m - \mathbf{T}_{neg}^m \mathbf{C}_{i-neg}^m\|_F^2. \end{aligned} \quad (12)$$

For effective fusion of different modalities, we normalize the constructed errors of each modality into [0, 1]:

$$\hat{\mathbf{e}}^m = (\mathbf{e}^m - \min(\mathbf{e^m}))/(\max(\mathbf{e^m}) - \min(\mathbf{e^m})), \quad (13)$$

with $m = 1, 2, \ldots, M$, where $\min(\mathbf{e^m})$ and $\max(\mathbf{e^m})$ denote the minimum and maximum elements of $\mathbf{e}^m$, respectively. Note that this normalization method is usually obtain good performance even though different error vectors have different normalization rules. In fact, the normalized error vectors are further combined with the optimized reliable weights, and thus we can achieve a reliability-aware fusion for different modalities. The discriminative score of i-th candidate can be defined as follows,

$$P(\mathbf{y}_i^{[M]}|\mathbf{x}_t) \propto \frac{1}{1 + \exp\{-\sum_{m=1}^{M} \alpha^m (\hat{e}_{i-neg}^m - \hat{e}_{i-pos}^m)\}}. \quad (14)$$

Thus, we could select a candidate region that maximizes the Eq. (1), which integrates the motion term and the discriminative likelihood term.

F. Implementation Details

For each candidate region, we extract gray features as follows. First, we resize each candidate or template with fixed size and partition it into local patches. Then, we quantize the grayscale values in each local patch into vectors, and concatenate these vectors together as the feature representation. To this end, the feature could preserve the global structure and get rid of the local effect, such as appearance variations and partial occlusions.

We initialize positive template by collecting multiple patches from the manually annotated regions. To alleviate model drifting, we keep the first template in the positive template set unchanged. In addition, we use the past tracking result to replace the positive template which has the largest similarity with the new target appearance, if the largest similarity is larger than a specified threshold σ_t. Otherwise, we discard this bad sample without update, as there is a large appearance change or a part of the target is occluded. To adapt the variations in different scenes, we define the above threshold by mean distance of positive samples in initial frame in our experiments, where is on the other hand, negative templates are updated dynamically. For each frame, we sample positive patches from an annular region that keeps a few pixels away from the center of tracking result.

G. Difference With Related Work

It should be note that the proposed tracking algorithm based on collaborative sparse representation is significantly different from recently proposed approaches that use sparse representation for grayscale-thermal tracking [4], [10].

In [10], the image patches from different sources of each target candidate are concatenated into a one-dimensional vector that is then sparsely represented in the target template space. The solution of the sparsity in the representation can be achieved by solving an l_1-regularized least squares problem, and the tracking result is then determined by finding the candidate with the smallest approximation error. While [4] employs the joint sparse representation on both modalities, and the resultant tracking results are fused using min operation on the sparse representation coefficients.

The proposed algorithm is significantly different from [4] and [10] in several aspects. First, our algorithm assigns each modality with a weight that describes the modal reliability, and thus pursues a collaborative sparse representation for adaptive object tracking. Second, the quality-aware and discrimination-aware regularizations on weights make our algorithm robust in dealing with occasional perturbation or malfunction of individual sources. Third, our algorithm jointly optimizes the sparse codes and the weight variables of different modalities in an online way. The sparse codes are efficiently optimized by the APG method [31], and the weights are solved with closed-form solutions. Finally, our algorithm utilizes reconstruction residues from both positive and negative basis to define the discriminative likelihood, which can prevent the model drift effectively when the target appearance is similar to the background.

IV. Grayscale-Thermal Tracking Benchmark

In this section, we introduce the newly collected grayscale-thermal object tracking (GTOT) benchmark, including dataset construction with statistics analysis, baseline methods with both grayscale and grayscale-thermal inputs, and evaluation metrics.

A. GTOT Benchmark

We collected 50 grayscale-thermal video clips under different scenarios and conditions, e.g., office areas, public roads, and water pool, etc. Each grayscale video is paired with one thermal video. We manually annotated them with ground truth bounding boxes. All annotations are done by a full-time annotator, to guarantee consistency. Among these videos, there are 4 videos from OSU Color-Thermal [5] and LITIV [6]. Some typical samples are presented in Fig. 3. We introduce more details of the dataset in the following.

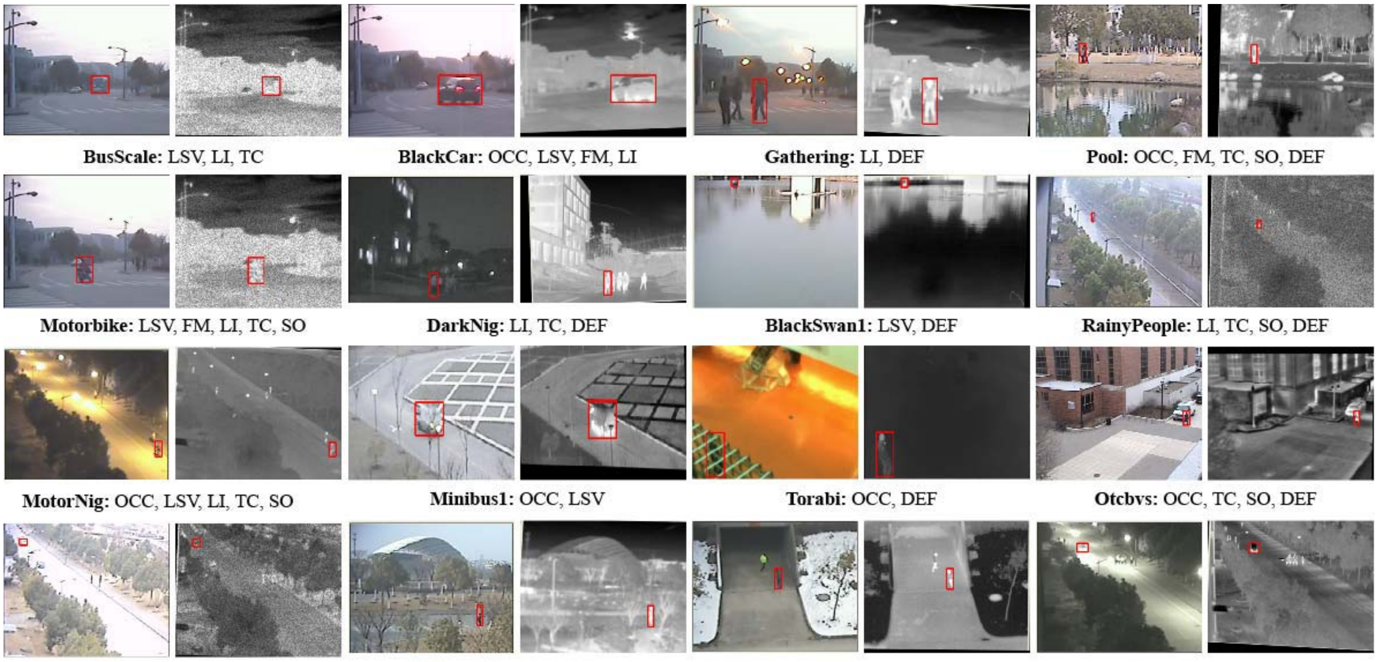
Hardware Setup. Our recording system consists of an online thermal image (MAG32) and a CCD camera (SONY TD-2073). We mount these two cameras in tripods, and make their views overlapped as much as possible for convenient alignment.

Alignment. Unlike industry registration in RGBD sensors, we manually construct the recording system, and develop an annotation tool to align grayscale-thermal videos in following way. We uniformly select a number of point correspondences in keyframe of the video pair, and compute the homography matrix by the least-square method. Then, the video pair can be aligned by applying the computed homography matrix to transform remaining frame pairs. This registration method can accurately align video pairs due to two main reasons. First, we carefully choose the planar and non-planar scenes to make the homography assumption effective. Second, since two camera views are almost coincident as we made, the transformation between two views is simple.

Annotation. Due to manually aligning in video pairs, we annotate the ground truth of dataset by drawing minimum bounding boxes covering the targets on both grayscale and thermal frames. In particular, all frames are manually annotated by one person to keep high consistency of dataset. When occlusion occurs, ground truth box only cover the visible portion of the target. When the target of one modal frame is ambiguous while another is distinguishable, we can not exactly annotate the ground truth of ambiguous one. In such circumstance, we define its ground truth as the ground truth of distinguishable one.

Statistics. We captured video pairs in sixteen scenes, including laboratory rooms, campus roads, play grounds and water pools, etc. We analyze the diversity and bias of this video dataset from the following aspects.

- *Object category*. We annotate both rigid objects and nonrigid objects. Rigid objects, including vehicles and human heads, usually move fast and have large scale variation over time. Non-rigid objects, including humans and swans, have very high degree of freedom, and usually increase the difficulty in tracking.
- *Object size*. Small objects frequently appear in visual surveillance. Tracking of small objects is important yet challenging because they do not include enough appearance information.
- *Illumination condition*. The video sequences are recorded under different weather conditions, including sunshine, overcast sky, rain and snow, which can bring big challenges in grayscale videos. In addition, the illumination of some grayscale videos significantly varies over time.
- *Thermal crossover*. When the targets have similar temperature with other objects or background, "thermal crossover" will occur in thermal videos. In such



BusScale: LSV, LI, TC
BlackCar: OCC, LSV, FM, LI
Gathering: LI, DEF
Pool: OCC, FM, TC, SO, DEF
Motorbike: LSV, FM, LI, TC, SO
DarkNig: LI, TC, DEF
BlackSwan1: LSV, DEF
RainyPeople: LI, TC, SO, DEF
MotorNig: OCC, LSV, LI, TC, SO
Minibus1: OCC, LSV
Torabi: OCC, DEF
Otcbvs: OCC, TC, SO, DEF
Exposure4: OCC, LSV, LI, TC, SO
WalkingOcc: OCC, TC, DEF
Quarreling: OCC,LSV,FM,LI,TC,SO,DEF
LightOcc: OCC, LSV, FM, LI

Fig. 3. Sample video pairs with annotated attributes from our grayscale-thermal dataset. The first frame of each sequence is shown with initial bounding box of the target object.

circumstance, thermal information will be ambiguous in tracking objects.

- *Scale variation.* Scale adaptation is important to alleviate model drift. Some of our videos have large scale variation, e.g., a car closing toward the camera. We extract histogram of relative area to the first frame bounding box in Fig. 4 (e) to reflect the statistics of scale variation.
- *Presence of occlusion.* Our videos cover several aspects of occlusion, e.g., how much the target is occluded, how long the target is occluded, and the target is completely occluded in grayscale (thermal) video while can be seen in thermal (grayscale) video.
- *Moving speed.* We also take the motion speed of the target into account. Fig. 4 (d) shows the histogram of center distance between consecutive frames, which can reflect the statistics of moving speed.

Table I summarizes attributes of the newly built video datasets. These fine-grained annotations allow us to analyze the attributed-sensitive performance of the object tracking methods. We present the attribute distribution in Fig. 4 (a).

TABLE I
LIST OF THE ATTRIBUTES ANNOTATED TO OUR GRAYSCALE-THERMAL DATASET

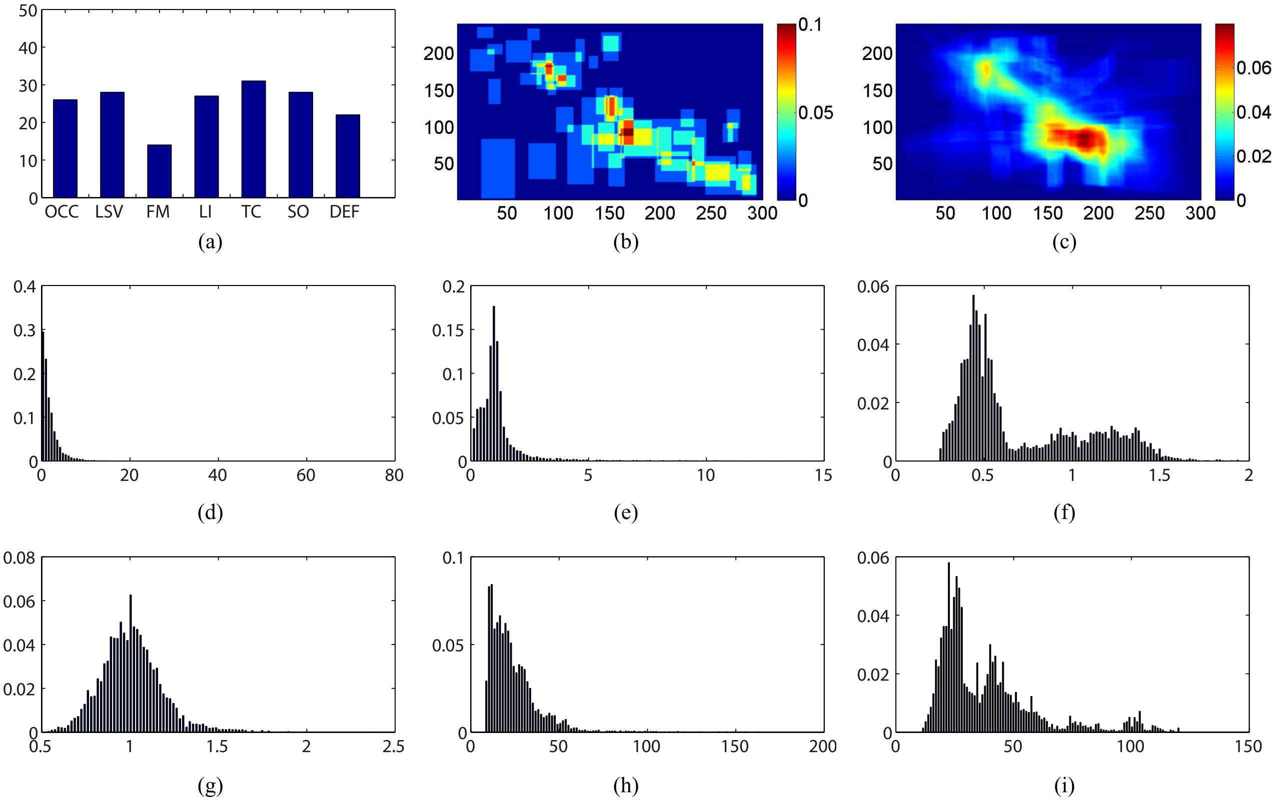
Attribute	Description
OCC	Occlusion - the target is partially or fully occluded.
LSV	Large Scale Variation - the ratio of the first bounding box and the current bounding box is out of the range [0.5, 1].
FM	Fast Motion - the motion of the ground truth is larger than 10 pixels.
LI	Low Illumination - the illumination in the target region is low.
TC	Thermal Crossover - the target has similar temperature with other objects or background.
SO	Small Object - the number of pixels in the ground truth bounding box is less than 400.
DEF	Deformation - non-rigid object deformation.

B. Baseline Methods

For evaluating the proposed approach and providing a comprehensive evaluation benchmark, we include some popular tracking methods as baselines into our GTOT benchmark.

On one hand, thirteen grayscale trackers are presented to demonstrate their performance in challenging scenarios, including DSST [33], RPT [35], MUSTer [34], MEEM [36], PCOM [40], CN [23], STC [37], SCM [16], KCF [24], CT [39], Struck [18], TLD [38] and MIL [19]. Table II summarizes these baselines. On the other hand, we also implement thirteen grayscale-thermal trackers for identifying the importance of thermal information, where eleven trackers of them are induced by grayscale trackers (DSST, Struck, SCM, KCF, CN, CT, MIL, STC, and TLD), and another two trackers are L1-PF [10] and JSR [4]. In particular, we concatenate features used in trackers from grayscale and thermal modalities as grayscale-thermal input of corresponding tracking algorithms.

Note that the model of JSR is the same as Ours-II, i.e., the reliable weights are removed in the proposed model, see Section V-C for more details. However, the main difference between JSR and Ours-II is that Ours-II utilizes reconstruction residues from both positive and negative basis to define the discriminative likelihood, while JSR fused the resultant tracking results using min operation on the sparse representation coefficients, which might be ineffective when target appearance is similar to the background.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 4. Dataset statistics. (a) Attribute distribution of entire dataset. (b) Distribution of first frame bounding boxes. (c) Distribution of all bounding boxes. (d) Distance between consecutive frames. (e) Relative area to first frame bounding box. (f) Width-height ratio of all bounding boxes. (g) Width-height ratio of first frame bounding boxes. (h) Width distribution of all bounding boxes. (i) Height distribution of all bounding boxes.

C. Evaluation Metrics

We utilize two widely used metrics, precision score and success plot, to evaluate the tracking performance. In particular, the final precision score and success score are defined as the best one of two modalities.

Precision score. Center Position Error is the Euclidean distance between the center locations of the tracked object and the ground truth bounding box, and usually employed to evaluate tracking precision. However, it will be invalid when the trackers are failed [19]. To evaluate overall performance, we employ the precision score used in recent literatures [19], [24], [41], the percentage of frames whose output location is within the given threshold distance of ground truth. Since many targets are small, we set the threshold to be 5 pixels instead of 20 pixels in other works [19], [24] to obtain the representative Precision Score (PS).

Success plot. Bounding box overlap is another effective evaluation metric. Given the output bounding box r_o and the ground truth bounding box r_g, the overlap score is defined as $S = \frac{|r_o \bigcap r_g|}{|r_o \bigcup r_g|}$, where $\bigcap$ and $\bigcup$ denote the intersection and union operators of two regions, and $|\cdot|$ indicates the number of pixels in the region. Setting a threshold t_o of overlapping area, we can calculate the success rate value, the ratio of the number of successful frames whose overlap S is larger than t_o. The success plot can be shown by the success rate at different t_o ($\in [0, 1]$). Unlike specifying a fixed threshold in CPE, we employ the Area Under Curve of success plots to define the Success Score (SS), and rank the tracking algorithms [41].

TABLE II

LIST OF THE BASELINE TRACKERS WITH THE USED FEATURES, THE MAIN TECHNIQUES AND THE PUBLISHED INFORMATION

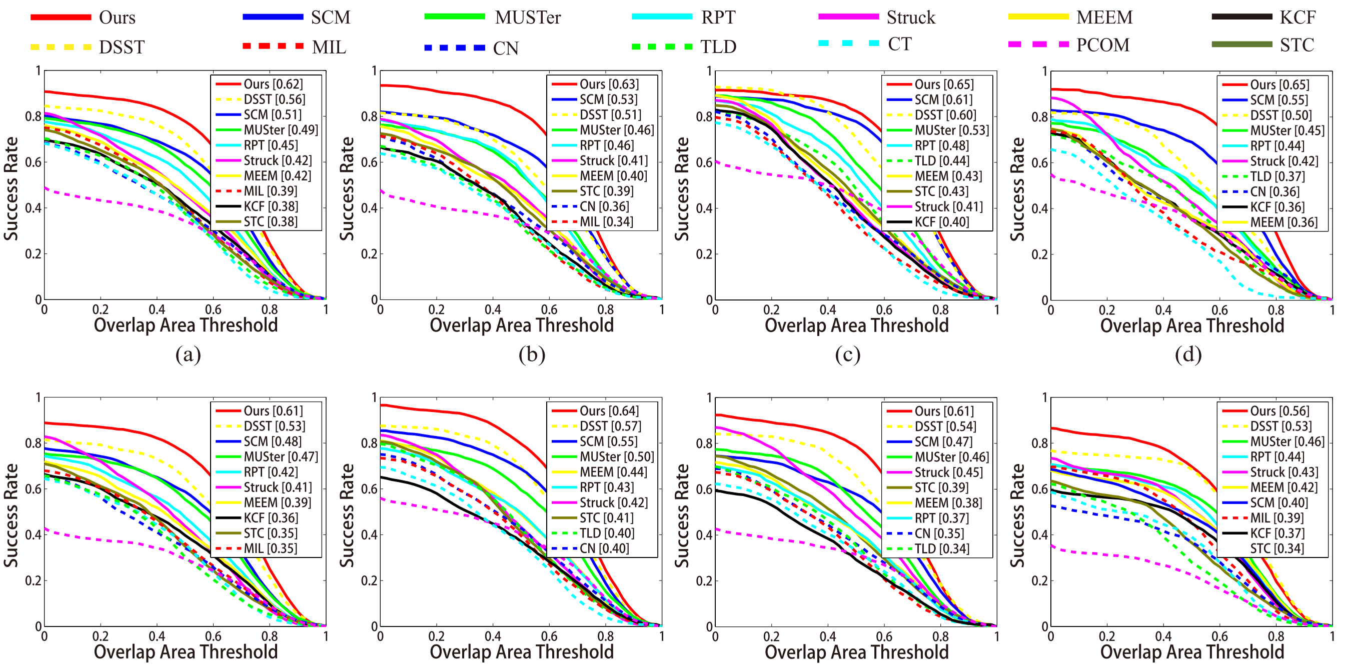
Baseline	Feature	Technique	Booketitle	Year
RPT	Grayscale	Sequential Monte Carlo	CVPR	2015
MUSTer	HOG	Correlation filter	CVPR	2015
DSST	HOG&Grayscale	Correlation filter	BMVC	2014
PCOM	Grayscale	MRF	CVPR	2014
MEEM	Lab	SVM	ECCV	2014
CN	Colors	Correlation filter	CVPR	2014
STC	Grayscale	Correlation filter	ECCV	2014
JSR	Colors	Sparse representation	INFOSCI	2014
SCM	Grayscale	Sparse representation	CVPR	2012
KCF	HOG	Correlation filter	ECCV	2012
CT	Haar	Naive Bayes classifier	ECCV	2012
L1-PF	Grayscale	Sparse representation	ICIF	2011
Struck	Haar	SVM	ICCV	2011
TLD	Binary pattern	P-N learning	TPAMI	2011
MIL	Haar	Boosting	CVPR	2009

V. EXPERIMENTS

In this section, we apply the proposed tracking method on our GTOT benchmark and compare with other popular tracking methods. The source codes, evaluation metrics, result figures will be provided with the benchmark for public usage in the community.

A. Parameter Settings

We detail the parameter settings of Algorithm 1 as follows. The balance parameter λ is to control the sparsity of reconstruction coefficients on both modalities. The results with different λ are shown in Table III, and thus we set it



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Success plots of the proposed trackers with state-of-the-art trackers on the entire dataset and several subsets. Herein, baseline trackers are with only grayscale input, and top ten trackers are shown in the legend for clarity. (a) Entire dataset. (b) OCC subset. (c) LSV subset. (d) FM subset. (e) LI subset. (f) TC subset. (g) SO subset. (h) DEF subset.

TABLE III

PRECISION SCORE/SUCCESS SCORE (PS/SS) OF THE PROPOSED METHOD WITH DIFFERENT PARAMETERS

Param	Setting	PS/SS	Param	Setting	PS/SS
λ	0.0001	71/60	Region Size	40×40	57/49
	0.001	75/62		32×32	75/62
	0.01	67/56		24×24	68/56
s	2	75/62	Patch Size	16×16	70/58
	3	70/59		8×8	75/62
	4	65/54		4×4	64/55

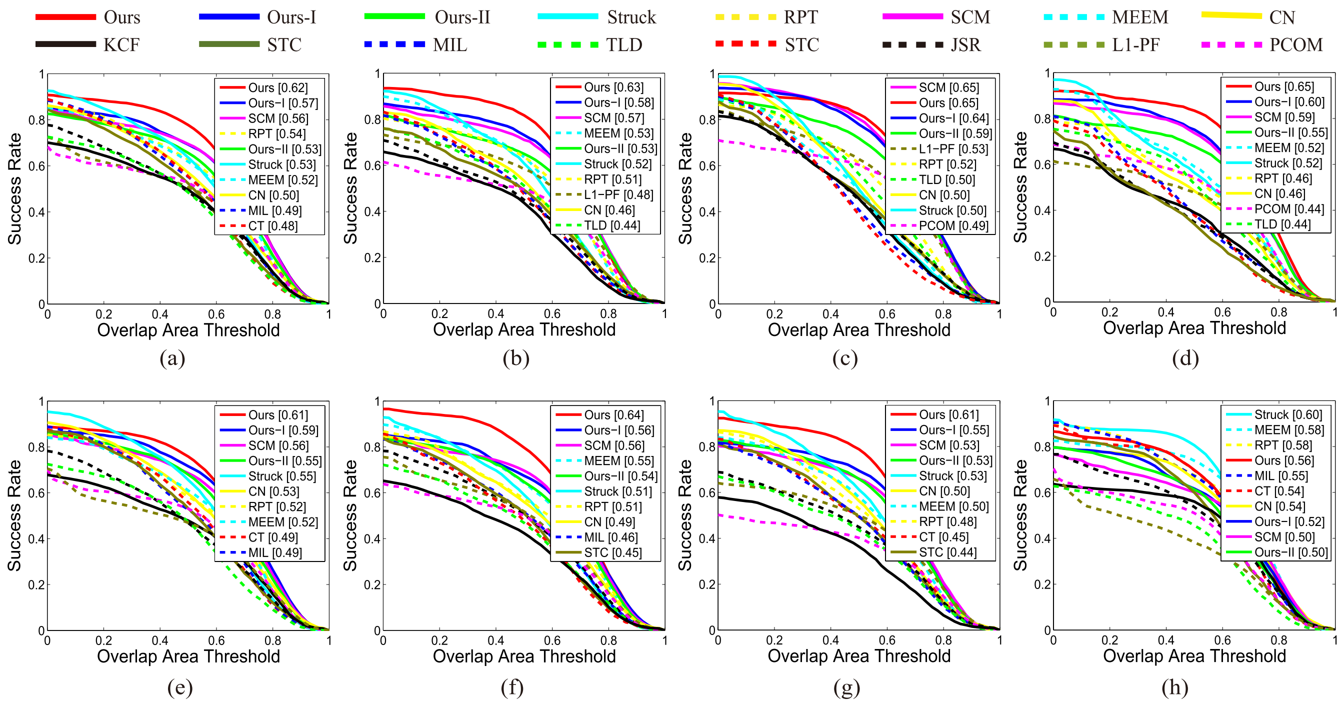
to be $\lambda = 0.001$. The fuzzy parameter $s \in (1, \infty)$ [30] is set to be 2 by observing the performance with different s in Table III. We resize each sample to be a region of 32×32 pixels, and evenly partition each patch into 8×8 patches for better performance, as shown in Table III. Note that the sizes of the normalized region and the partitioned patch are important for robust tracking. This is because that the feature representations generally play a crucial role in visual tracking task. In addition, s controls the variations among reliable weights, and its influences to the performance demonstrate the significance of the introduced modal reliable weights.

The negative and positive templates are updated in each frame to adapt the appearance variations of surrounding background and target object in time, respectively. In Bayesian filtering, the number of candidates is generally determined by the trade-off between the computational cost and the variance of the resulting estimates. Some works set a different number for each video sequence according to the motion variations of target object, but we fix it to be 200 on entire dataset for more fair comparison. Similarly, the positive and negative samples are set to be 20 and 200, respectively. σ_t determines that whether one tracking result is used to replace one template in the positive template set or not. Smaller σ_t will keep the adaptation of the template set to variations of object appearance, while easily introducing noisy templates, and vice versa. Therefore, we set it to be 1.5 and 0.8 for grayscale and thermal modalities for balancing the adaptation and clearness of the positive template set, respectively.

B. Comparison Results

To justify the importance of thermal information and the effectiveness of the proposed approach, we evaluate the compared methods with grayscale or grayscale-thermal inputs, where the baselines have been introduced in Section IV-B.

Overall performances. We first report the success plots and Success Score (SS) of grayscale trackers on entire dataset in Fig. 5 (a) and Table IV, respectively. From the evaluation results, we can observe that the proposed approach substantially outperforms the baseline trackers. This comparison clearly demonstrates the effectiveness of our approach for exploiting thermal information. The success plots and SS of grayscale-thermal trackers on entire dataset are then presented in Fig. 6 (a) and Table V respectively for justifying the effectiveness of the proposed method in adaptively exploiting grayscale-thermal information. We also employ Precision Score (PS) to evaluate the overall performance, as shown in Table VI. The results further suggest that our method can achieve robust tracking in challenging scenarios by employing grayscale-thermal information, and outperforms other state-of-the-art trackers with both grayscale and grayscale-thermal inputs.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6. Success plots of the proposed trackers with state-of-the-art trackers on the entire dataset and several subsets. Herein, baseline trackers are with grayscale-thermal input, and top ten trackers are shown in the legend for clarity. (a) Entire dataset. (b) OCC subset. (c) LSV subset. (d) FM subset. (e) LI subset. (f) TC subset. (g) SO subset. (h) DEF subset.

TABLE IV

SUCCESS SCORE (SS, %) OF SUCCESS PLOTS AND CORRESPONDING RANKINGS (IN PARENTHESIS) WITH DIFFERENT ATTRIBUTES. HEREIN, BASELINE TRACKERS ARE WITH ONLY GRAYSCALE INPUT. THE BOLD FONTS OF RESULTS INDICATE THE BEST PERFORMANCE, AND THE FINAL RANK IS THE AVERAGE RANK OF ALL ATTRIBUTES

Algorithm	**Rank**	**All**	**OCC**	**LSV**	**FM**	**LI**	**TC**	**SO**	**DEF**
Ours	1.00	**62(1)**	**63(1)**	**65(1)**	**65(1)**	**61(1)**	**64(1)**	**61(1)**	**56(1)**
DSST [34]	2.00	56(2)	51(3)	60(3)	50(3)	53(2)	57(2)	54(2)	53(2)
SCM [16]	3.13	51(3)	53(2)	61(2)	55(2)	48(3)	55(3)	47(3)	40(7)
MUSter [35]	3.88	49(4)	46(4)	53(4)	45(4)	47(4)	50(4)	46(4)	46(3)
RPT [36]	5.25	45(5)	46(4)	48(5)	44(5)	42(5)	43(6)	37(8)	44(4)
Struck [18]	6.00	42(6)	41(5)	41(8)	42(6)	41(6)	42(7)	45(5)	43(5)
MEEM [37]	6.50	42(6)	40(6)	43(7)	36(8)	39(7)	44(5)	38(7)	42(6)
STC [38]	8.00	38(8)	39(7)	43(7)	35(9)	35(9)	41(8)	39(6)	34(10)
TLD [39]	9.38	36(9)	32(10)	44(6)	37(7)	32(11)	40(9)	34(10)	29(13)
CN [23]	9.38	36(9)	36(8)	39(10)	36(8)	33(10)	40(9)	35(9)	31(12)
KCF [24]	9.38	38(8)	34(9)	40(9)	36(8)	36(8)	34(11)	29(13)	37(9)
MIL [19]	9.50	39(7)	34(9)	37(12)	33(10)	35(9)	37(10)	32(11)	39(8)
CT [40]	11.13	34(10)	32(10)	36(13)	28(12)	32(11)	34(11)	31(12)	32(11)
PCOM [41]	11.88	29(11)	28(11)	38(11)	31(11)	25(12)	34(11)	26(14)	19(14)

Attribute-based Performance. For evaluating trackers on subsets with different attributes to facilitate analysis of performance on different challenging factors, We present the success plots of all attributes in Fig. 5 (b-i) and Table IV for grayscale trackers, and Fig. 6 (b-i) and Table V for grayscale-thermal trackers, respectively.

For grayscale trackers, our method substantially outperforms all baselines in all attributes, demonstrating the effectiveness of our method in various challenging factors. The results also demonstrate the importance of thermal information in visual tracking, especially in LI and SO. In such scenarios, thermal sources can provide more reliable information.

For grayscale-thermal trackers, the proposed method achieve superior performance over other baselines in all attributes except for DEF, further validating effectiveness of our tracking method. In particular, for occasional perturbation or malfunction of one modality (e.g., LI, TC and SO), our method can effectively incorporate another modal information to track objects robustly.

C. Component Analysis

To justify the significance of the main components of the proposed model, we implement two special versions for comparative analysis, including: 1) *Ours-I*, that set $\phi = 0$ to remove the adaptive parameter on $\alpha^{[M]}$ in Eq. 4. 2) *Ours-II*, that fix $\alpha^m = \frac{1}{M}(m = 1, 2, .., M)$ to remove the weight variables in Eq. 4.

TABLE V

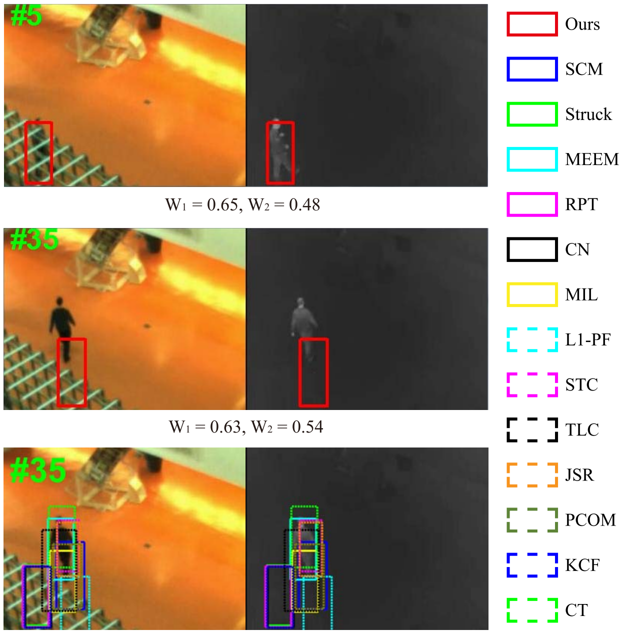
SUCCESS SCORE (SS, %) OF SUCCESS PLOTS AND CORRESPONDING RANKINGS (IN PARENTHESIS) WITH DIFFERENT ATTRIBUTES. HEREIN, BASELINE TRACKERS ARE WITH GRAYSCALE-THERMAL INPUT. THE BOLD FONTS OF RESULTS INDICATE THE BEST PERFORMANCE, AND THE FINAL RANK IS THE AVERAGE RANK OF ALL ATTRIBUTES

Algorithm	**Rank**	**All**	**OCC**	**LSV**	**FM**	**LI**	**TC**	**SO**	**DEF**
Ours	1.25	**62(1)**	**63(1)**	**65(1)**	**65(1)**	**61(1)**	**64(1)**	**61(1)**	56(3)
Ours-I	2.50	57(2)	58(2)	64(2)	60(2)	59(2)	56(2)	55(2)	52(6)
Ours-II	4.25	53(5)	53(4)	59(3)	55(4)	55(4)	54(4)	53(3)	50(7)
SCM [16]	3.13	56(3)	57(3)	65(1)	59(3)	56(3)	56(2)	53(3)	50(7)
Struck [18]	4.25	53(5)	52(5)	50(6)	52(5)	55(4)	51(5)	53(3)	**60(1)**
MEEM [37]	4.75	52(6)	53(4)	46(8)	52(5)	52(6)	55(3)	50(4)	58(2)
RPT [36]	4.88	54(4)	51(6)	52(5)	46(6)	52(6)	51(5)	48(5)	58(2)
CN [23]	5.88	50(7)	46(8)	50(6)	46(6)	53(5)	49(6)	50(4)	54(5)
MIL [19]	7.75	49(8)	43(10)	43(10)	39(9)	49(7)	46(7)	44(7)	55(4)
CT [40]	8.50	48(9)	43(10)	42(11)	37(10)	49(7)	43(10)	45(6)	54(5)
L1-PF [10]	8.88	43(11)	48(7)	53(4)	40(8)	40(12)	44(9)	42(8)	34(12)
STC [38]	9.13	46(10)	42(11)	50(6)	34(12)	48(12)	45(8)	44(7)	50(7)
TLD [39]	9.38	41(13)	44(9)	50(6)	44(7)	40(8)	40(11)	38(10)	36(11)
JSR [4]	10.00	43(11)	39(13)	44(9)	34(12)	43(9)	44(9)	39(9)	45(8)
PCOM [41]	10.00	42(12)	40(12)	49(7)	44(7)	42(10)	40(11)	33(11)	40(10)
KCF [24]	11.50	42(12)	36(14)	42(11)	35(11)	41(11)	37(12)	32(12)	44(9)

TABLE VI

PRECISION SCORE (PS) AND FRAME PER SECOND (FPS) OF THE COMPARED TRACKERS WITH BOTH GRAYSCALE (G) AND GRAYSCALE-THERMAL (G-T) INPUTS ON THE ENTIRE DATASET. THE BOLD FONTS OF RESULTS INDICATE THE BEST PERFORMANCE

Algorithm	**G**	**G-T**	Code Type	FPS
Ours	-	**75**	MATLAB & C++	1.6
Ours-I	-	67	MATLAB & C++	1.6
Ours-II	-	62	MATLAB & C++	3.4
MUSter [35]	**63**	-	MATLAB & C++	4.0
DSST [34]	68	-	MATLAB & C++	35.3
SCM [16]	58	55	MATLAB & C++	0.3
RPT [36]	54	-	MATLAB & C++	2.6
MEEM [37]	48	-	MATLAB & C++	4.9
STC [38]	48	61	MATLAB	225.5
KCF [24]	47	56	MATLAB & C++	124.1
L1-PF [10]	-	55	MATLAB & C++	5.1
CN [23]	47	66	MATLAB & C++	65.4
Struck [18]	46	68	C++	10.8
JSR [4]	-	46	MATLAB	0.8
TLD [39]	38	45	MATLAB & C++	2.7
MIL [19]	36	48	C++	3.7
PCOM [41]	32	-	MATLAB & C++	21.6
CT [40]	28	43	MATLAB	31.8



W$_1$ = 0.65, W$_2$ = 0.48

W$_1$ = 0.63, W$_2$ = 0.54

Fig. 7. One failure case of our method. W$_1$ and W$_2$ denote the optimized weights of grayscale and thermal sources, respectively. The third row presents the tracking results of other grayscale-thermal methods.

TABLE VII

COMPARISON RESULTS (PS/SS) OF THE PROPOSED APPROACH WITH SEVERAL TYPICAL SETTINGS OF THE RELIABLE WEIGHTS (G/T). HEREIN, ONE SETTING DENOTES THAT THE RELIABLE WEIGHTS OF GRAYSCALE AND THERMAL MODALITIES ARE FIXED ON ENTIRE DATASET

Ours	0.5/0.5	0.6/0.4	0.4/0.6	0.1/0.01	0.01/0.1
75/62	62/53	68/58	67/58	60/49	51/48

The evaluation results are reported in Fig. 6 and Table V, VI, and we can draw the following conclusions. 1) The complete algorithm outperforms Ours-I. This justifies the contribution of the proposed adaptive parameter ϕ on $\alpha^{[M]}$. 2) Our method substantially outperforms Ours-II. This demonstrates the significance of the introduced weighted variables.

It is worth mentioning that the reasonable estimations of reliable weights are important for robust adaptive tracking. In general, if a modality is unreliable or ambiguous, our approach will automatically assign it with a low weight to alleviate the effects of this modality in tracking. Without such weighting schema, the unreliable information or noises might contaminate the object model, leading to model drifting eventually. While it is possible to learn improper weights as shown in Fig. 7, the proposed method can work well for most of challenging scenes. To empirically justify the effectiveness of our approach, Table VII presents the

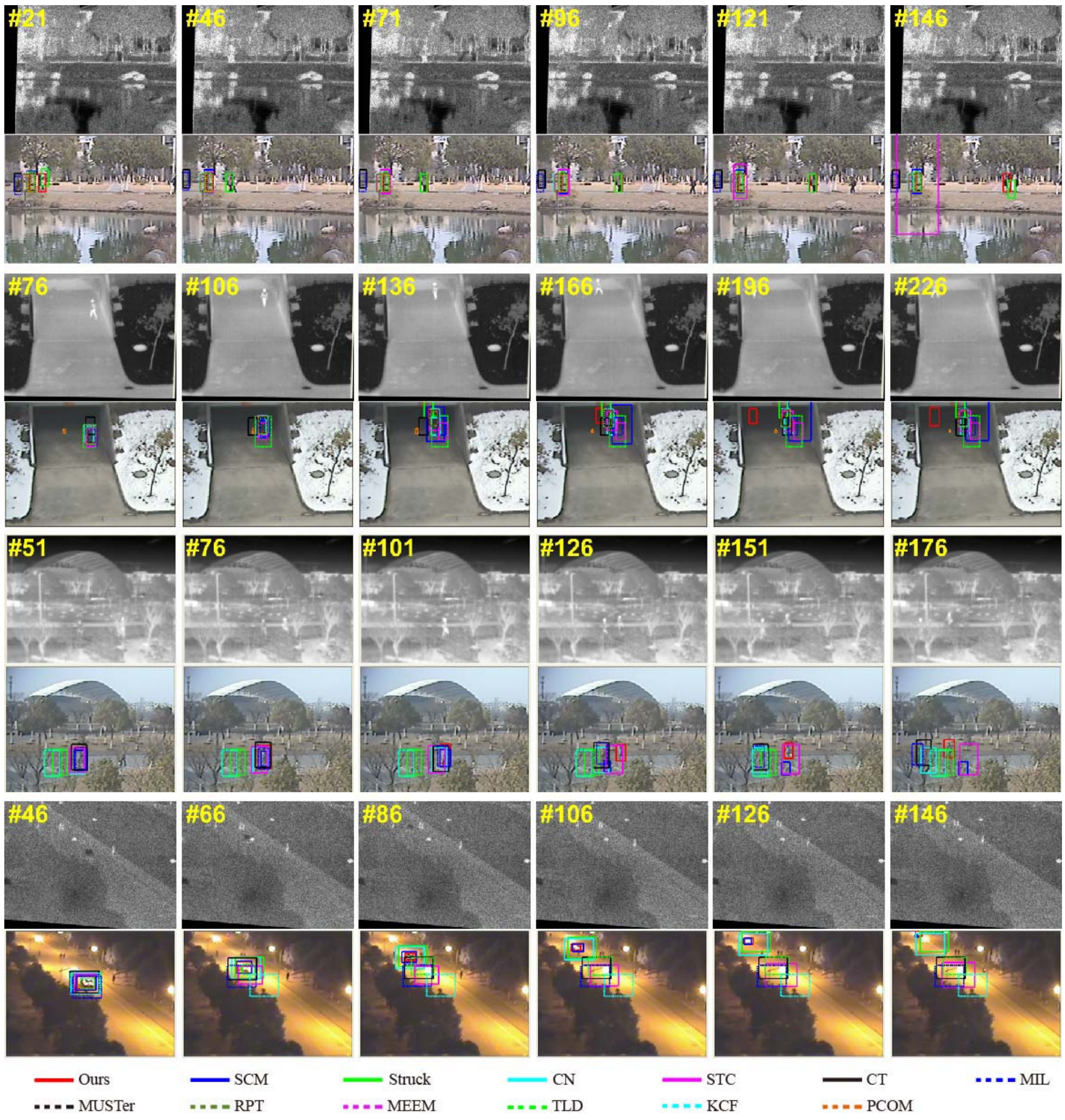


Fig. 8. Output bounding boxes of trackers with grayscale and grayscale-thermal inputs. The first two sequence pairs denote the grayscale tracking results, and the last two sequence pairs indicate the grayscale-thermal tracking results.

comparison results of the proposed approach with 5 typical settings of reliable weights, which further validates the effectiveness of the weight estimation of the proposed approach.

We also present one failure case generated by our method in Fig. 7. The reliable weights sometimes are wrongly estimated due to the effect of clutter background. From Eq. 4, we can see that the reliable weight of one modality is determined by its reconstruction error and adaptive parameter, i.e., the smaller reconstruction error and adaptive parameter are, the larger of the reliable weight is. When a more reliable modality has bigger reconstruction error (e.g., Fig. 2) and adaptive parameter, the computed reliable weight will be smaller. In such circumstance, our method will generate wrong tracking results. In Fig. 7, the thermal information is reliable to track the target object, but most grayscale-thermal methods are failed. This problem could be tackled by incorporating the measure of background clutter in optimizing reliable weights, and will be addressed in our future work.

In addition to reliable weight computation, our approach has another major limitation. The real-time tracking performance is very important for visual processing systems, but our algorithm has high computational cost due to inefficient optimization to the proposed model. In future work, we will alleviate it from two aspects. First, we will develop a fast solver to l_1-regularized representation [42]. Second, we will propose a strategy to reduce the number of sampled candidates while improving their qualities, which can improve the tracking efficiency without losing much accuracy.

D. Discussions

We observe from the evaluations that integrating grayscale data and thermal data will boost tracking performance. The improvements are even bigger while encountering certain challenges, i.e., low illuminations, small objects and thermal crossover, demonstrating the complementary benefits from grayscale-thermal data.

We can also observe that some modal information is redundant, and directly utilizing both of their information will lead to bad tracking results, as shown in Table VI. We can address this issue from two aspects. First, adaptively integrating different modal information (Our method and Ours-I) can automatically determine the contribution weights of different modalities to alleviate effect of redundant information. Second, feature reduction or selection techniques (e.g., CN [23] and SCM [16]) can compress or remove some useless information and achieve more robust tracking performance.

In addition, we found from evaluations that scale adaptation (e.g., the proposed methods, DSST [33] and SCM [16]) and context information (e.g., Struck [18] and STC [37]) are crucial for effective grayscale-thermal tracking.

VI. CONCLUSIONS

In this work, we proposed a robust adaptive tracking algorithm that integrated collaborative sparse representation in Bayesian framework, and built a comprehensive video benchmark (GTOT) for grayscale-thermal tracking. Extensive experiments on the new benchmarks demonstrated superior performance over other popular tracking methods. In future work, we will improve the efficiency of our method for realtime command, and expand the video dataset to include more challenging scenes, and evaluate more popular trackers as parts of the benchmark platform.

REFERENCES

[1] C. O. Conaire, N. E. Connor, E. Cooke, and A. F. Smeaton, "Comparison of fusion methods for thermo-visual surveillance tracking," in *Proc. Int. Conf. Inf. Fusion*, Jul. 2006, pp. 1–7.

[2] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, Jan. 2014.

[3] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weld: Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2016.2556586, 2016, in press.

[4] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Sci. China Inf. Sci.*, vol. 55, no. 3, pp. 590–599, Mar. 2012.

[5] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vis. Image Understand.*, vol. 106, no. 2, pp. 162–182, 2007.

[6] A. Torabi, G. Masse, and G.-A. Bilodeau, "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Comput. Vis. Image Understand.*, vol. 116, no. 2, pp. 210–221, 2012.

[7] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi, "Thermal–visible registration of human silhouettes: A similarity measure performance evaluation," *Inf. Phys. Technol.*, vol. 64, pp. 79–86, Mar. 2014.

[8] C. O. Conaire, N. E. Connor, and A. Smeaton, "Thermo-visual feature fusion for object tracking using multiple spatiogram trackers," *Mach. Vis. Appl.*, vol. 19, no. 5, pp. 483–494, Oct. 2008.

[9] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, "Geodesic active contour based fusion of visible and infrared video for persistent object tracking," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Feb. 2007, pp. 35–42.

[10] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Proc. Int. Conf. Inf. Fusion*, Jul. 2011, pp. 1–8.

[11] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1521–1528.

[12] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[13] X. Mei and H. Ling, "Robust visual tracking using L1 minimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2009, pp. 1436–1443.

[14] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2042–2049.

[15] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust L1 tracker using accelerated proximal gradient approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1830–1837.

[16] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1838–1845.

[17] C. Li, S. Hu, S. Gao, and J. Tang, "Real-time grayscale-thermal tracking via laplacian sparse representation," in *Proc. Int. Conf. Multimedia Modelling*, Jan. 2016, pp. 54–65.

[18] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 263–270.

[19] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[20] C. Li, L. Lin, W. Zuo, S. Yan, and J. Tang, "Sold: Sub-optimal low-rank decomposition for efficient video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5519–5527.

[21] C. Li, L. Lin, W. Zuo, W. Wang, and J. Tang, "An approach to streaming video segmentation with sub-optimal video segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1947–1960, May 2016.

[22] N. Jiang, W. Liu, and Y. Wu, "Order determination and sparsity-regularized metric learning adaptive visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1956–1963.

[23] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van De Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1090–1097.

[24] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[25] S. Song and J. Xiao, "Tracking revisited using RGBD camera: Unified benchmark and baselines," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 233–240.

[26] N. Cvejic *et al.*, "The effect of pixel-level fusion on object tracking in multi-sensor surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[27] A. Leykin and R. Hammoud, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Mach. Vis. Appl.*, vol. 21, no. 4, pp. 587–595, Jun. 2010.

[28] Y. Wu, H. Ling, J. Yu, F. Li, X. Mei, and E. Cheng, "Blurred target tracking by blur-driven tracker," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1100–1107.

[29] S. Bahrampour, A. Ray, N. M. Nasrabadi, and K. W. Jenkins, "Quality-based multimodal classification using tree-structured sparsity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4114–4121.

[30] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2, pp. 191–203, 1984.

[31] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2013.

[32] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for sparse hierarchical dictionary learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 487–494.

[33] M. Danelljan, G. Hager, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2014, pp. 1–11.

[34] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "MUlti-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 749–758.

[35] Y. Li, J. Zhu, and S. C. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 353–361.

[36] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.

[37] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 127–141.

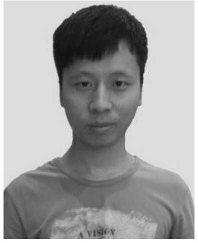
[38] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[39] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 864–877.

[40] D. Wang and H. Lu, "Visual tracking via probability continuous outlier model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3478–3485.

[41] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 2411–2418 .

[42] Z. Zhang and V. Saligrama. (Jun. 2014). "Rapid: Rapidly accelerated proximal gradient algorithms for convex minimization." [Online]. Available: https://arxiv.org/abs/1406.4445

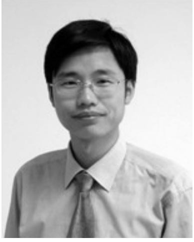


Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University. He has authored over 20 academic papers, including seven peer-reviewed articles in top-tier conferences and leading journals. His current research interests include computer vision, machine learning, and intelligent video analysis.

Hui Cheng received the B.Eng. degree in electrical engineering from Yan Shan University, Qinhuangdao, China, in 1998, the M.Phil. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2001, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, Hong Kong, in 2005. She was a Post-Doctoral Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, from 2006 to 2007. She is currently an Associate Professor with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China. Her current research interests include intelligent robots and networked control.

Shiyi Hu received the B.S. degree in information and computing science from Sun Yat-sen University, Guangzhou, China, in 2014, where he is currently pursuing the M.S. degree in computer science and technology. His current research interest is computer vision with a focus on visual tracking and multi-object tracking.

Xiaobai Liu received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China. He was a Post-Doctoral Research Fellow with the Center for Vision, Cognition, Learning and Art, University of California at Los Angeles, Los Angeles, CA, USA. He is currently an Assistant Professor with the Department of Computer Science, San Diego State University. He has authored over 30 peer-reviewed articles in top-tier conferences and leading journals. His current research interests include scene parsing with a variety of topics such as, joint inference for recognition and reconstruction, and commonsense reasoning. He received a number of awards for his academic contribution, including the 2013 Outstanding Thesis Award from the China Computer Federation.



Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, and machine learning.

Liang Lin (M'09–SM'15) received the B.S. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 1999 and 2008, respectively. From 2008 to 2009, he was a Post-Doctoral Research Fellow with the Department of Statistics, University of California at Los Angeles. He is currently a Full Professor with the School of Data and Computer science, Sun Yat-sen University, China. He has authored or co-authored over 100 papers in top-tier academic journals and conferences. His research focuses on new models, algorithms, and systems for intelligent processing and understanding of visual data, such as images and videos. He was a recipient of the Best Paper Runner-Up Award in ACM NPAR 2010, the Google Faculty Award in 2012, and the Best Student Paper Award in the IEEE ICME 2014. He has been serving as an Associate Editor of the IEEE TRANSACTIONS HUMAN–MACHINE SYSTEMS. He was supported by several promotive programs or funds for his works, such as the NSFC Excellent Young Scholars Program in 2016.