Online Non-Negative Multi-Modality Feature Template Learning for RGB-Assisted Infrared Tracking

Article *in* IEEE Access · May 2019

DOI: 10.1109/ACCESS.2019.2916895

CITATIONS
15

READS
157

6 authors, including:

Mang Ye
Wuhan University
46 PUBLICATIONS **1,463** CITATIONS
SEE PROFILE

Bineng Zhong
National Huaqiao University
131 PUBLICATIONS **3,281** CITATIONS
SEE PROFILE

Huiyu Zhou
University of Leicester
310 PUBLICATIONS **5,396** CITATIONS
SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Submitted to IEEE Access

Digital Object Identifier

Online Non-negative Multi-modality Feature Template Learning for RGB-assisted Infrared Tracking

XIANGYUAN LAN[1], MANG YE[1], RUI SHAO[1], BINENG ZHONG[2], DEEPAK KUMAR JAIN[3], HUIYU ZHOU[4]

[1]Department of Computer Science, Hong Kong Baptist University, HK, China (e-mail: xiangyuanlan@life.hkbu.edu.hk, {mangye, ruishao}@comp.hkbu.edu.hk)
[2]School of Computer Science and Technology, Huaqiao University, Xiamen, China (e-mail: bnzhong@hqu.edu.cn)
[3]Key Laboratory of Intelligent Air-Ground Cooperative Control for Universities in Chongqing, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China (e-mail: deepak@cqupt.edu.cn)
[4]Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K.(e-mail: hz143@leicester.ac.uk)

This work was supported in part by Hong Kong Baptist University Tier 1 Start-up Grant. The work of H. Zhou was supported by UK EPSRC under Grant EP/N011074/1, Royal Society-Newton Advanced Fellowship under Grant NA160342, and European Union's Horizon 2020 research and innovation program under the Marie-Sklodowska-Curie grant agreement No. 720325. The work of B. Zhong was supported by the National Natural Science Foundation of China under Grants 61572205 and 61802135. This work of D. K. Jain was supported in part by the Key Laboratory of Intelligent Air-Ground Cooperative Control for Universities in Chongqing and the Key Laboratory of Industrial IoT and Networked Control, Ministry of Education, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China.
Corresponding author: Mang Ye

ABSTRACT Infrared sensors have been deployed in many video surveillance systems because of the insensibility of their imaging procedure to some extreme conditions (e.g. low illumination condition, dim environment). To reduce the human labor in video monitoring and perform intelligent infrared video understanding, an important issue we need to consider is how to locate the object of interest in consecutive video frames accurately. Therefore, developing a robust object tracking algorithm for infrared videos is necessary. However, the infrared information may not be reliable (e.g. thermal crossover), and appearance modeling with only the infrared modality may not be able to achieve good results. To address this issues, with the wide deployment of RGB-infrared camera systems, this paper proposes an infrared tracking framework in which information from RGB-modality will be exploited to assist the infrared object tracking. Specifically, within the tracking framework, in order to deal with the contaminated features caused by large appearance variations, an online non-negative feature template learning model is designed. The non-negative constraint enables the model to capture the local part-based characteristic of the target appearance. To ensure more important modality contribute more in appearance representation, an adaptive modality importance weight learning scheme is also incorporated in the proposed feature learning model. To guarantee the model optimality, an iterative optimization algorithm is derived. Experimental results on various RGB-infrared videos show the effectiveness of the proposed method.

INDEX TERMS Optical image processing, sensor fusion, computer vision

I. INTRODUCTION

INFRARED sensors, which form images by capturing the infrared radiation of subjects, is more effective to record informative videos under some extreme conditions (e.g. low illumination conditions, dim environment). Therefore, infrared sensors have been employed in many video surveillance systems for security monitoring, traffic management, etc..The past decade has witnessed the rapid development of AI technology in many fields, such as computer vision (e.g. video surveillance [1]–[22], image and video classification [23]–[36], image retrieval [37], [38], image quality assessment and processing [39]–[45]), unmanned vehicle [46], machine learning (e.g. [47]–[64]), and biometric security (e.g. [65]). With the massive video data generated from infrared sensors, to save the time and human labor, video understanding and analysis using artificial intelligence techniques is required. To this end, a key step is how to accurately locate the position or infer the motion status of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2916895, IEEE Access

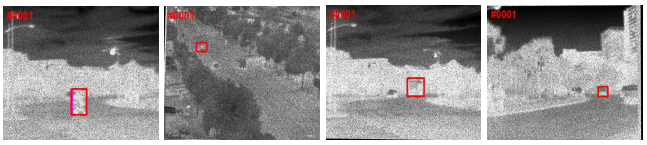

FIGURE 1: Illustration of some video frames of infrared-modality when the infrared information is not reliable for appearance modeling.

the object. Therefore, developing a robust infrared tracking is very important and useful for many applications such as underwater image perception [66]–[69], multi-sensor image understanding [70], unmanned aerial vehicle imagery [71], video surveillance [72], [73], and several infrared trackers have been developed [74]–[78] with improved tracking results.

However, information from infrared videos is not always reliable [79]. For example, for the case of thermal crossover in which the tracked target cannot be distinguished from the background because of the similar temperature, appearance modeling with single infrared modality may suffer the loss of discriminability, which means the tracker may be distracted into the background. Figure 1 illustrates some examples when the infrared information is not reliable. Therefore, more informative appearance cues should be incorporated to construct a more robust appearance model for infrared tracking. The rapid development of multispectral imaging techniques has brought the wide application of RGB-infrared dual-camera systems. Compared with infrared cameras, although the imaging procedure of the visible spectrum camera is more sensitive to some extreme conditions (e.g. in the darkness of nighttime), it can characterize more visual details such as color, texture for appearance modeling. As such, exploiting the reliable cues from RGB modality can compensate the weakness of infrared cues. Therefore, to enhance the performance of infrared tracking, it is useful to further integrate the reliable RGB information for appearance modeling.

To perform effective RGB-assisted infrared object tracking, there are two issues which should be considered. The appearance variations such as occlusions, cluttered background would usually be encountered during the tracking process, which would contaminate the training tracking samples and degrade the tracking performance. Therefore, how to effectively learn reliable cues of RGB or infrared modality from the potentially contaminated samples is the first issue to address. In addition, different modalities may contribute different roles to appearance modeling, and some extreme situations (e.g. thermal crossover) may also degrade the reliability of some modalities. Therefore, how to adaptively evaluate the reliability of different modalities is another important issues which should be considered.

Although several RGB-infrared tracking algorithms have been proposed, they may not well handle the aforementioned issues. Some typical feature fusion methods such as feature concatenation [80], sum rule [81] have been exploited to integrate the RGB and infrared modalities. However, these methods do not consider the reliability issues of different modalities. There are also some other methods such as [82] which regard tracking on RGB-infrared modalities as two independent tasks and fuse the results of different tasks to determine the final positions. However, the RGB and infrared modalities is not integrated for appearance modeling during the tracking process and thereby the reliability of different modalities is not adaptively evaluated, which limits the performance. Although several methods such as [83] attempt to weight the reliability of different modalities using some heuristic methods, the reliability weights are not coupled with the fusion model, which may not be an optimal estimation of reliability. Moreover, most of the aforementioned methods do not explicitly handle the contamination caused by large appearance change, and the performance would be degraded if large appearance variation happens.

To address the aforementioned issues, this paper propose a new learning model for RGB-assisted infrared object tracking. The proposed model aims to integrate the modality reliability weight estimation,uncontaminated modality feature template learning and fusion into a unified optimization framework. Within this framework, tracking samples are decontaminated during the feature learning process while modality reliability is adaptively evaluated. In addition, inspired by the non-negative matrix factorization [84], we incorporate the nonnegative constraint into the optimization framework, which guides the learned feature templates to capture the local part-based decomposition of the target appearance. An online optimization algorithm is derived to the learn the modal parameters.

In general, the contributions of this work are summarized as follows:

- We propose a new framework for infrared tracking. The proposed framework is able to integrate reliable RGB information with infrared modality for appearance modeling.
- We propose a robust nonnegative feature templates learning model to perform feature learning and fusion, and reliability weight estimation of multiple modality data.
- We derive an simplified version of the learning model which reduce the computation complexity and derive an effective optimization algorithm to learn the modal parameters.

The rest of this paper is organized as follows. Section II will provide some reviews of related works on infrared tracking and RGB-assisted infrared tracking. Section III will present the proposed method and the related optimization algorithm. The implementation details will be given in Section IV. The experimental results and the conclusion will be provided in Sections V-B and VI, respectively.

II. RELATED WORKS

This section will briefly introduce and discuss some related works on object tracking based on single infrared modality

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2916895, IEEE Access

and the fusion of RGB and infrared modalities. For more comprehensive literature review of object tracking, interested readers can refer to [85]–[89].

A. RGB-INFRARED OBJECT TRACKING

Based on level set model, Bunyak *et al.* proposed a unified framework for moving object segmentation and tracking [90]. A fusion-based tracking framework is proposed to combine the tracking results generated by multiple spatiogram trackers corresponding to RGB and infrared modalities for final target position decision [82]. A probabilistic background model is designed in [81] to infer and fuse the confidence map of target for tracking, where the confidence maps of RGB and infrared modalities are aggregated based on sum rule. Inspired by the success in sparse representation-based classification [91], several sparsity-based tracking algorithms have been proposed. They exploit the feature concatenation [80], joint sparsity regularization [83], [92], low rank regularization [93], collaborative discriminative learning [94], feature learning [95] to combine multi-modality information for appearance modeling.

B. INFRARED OBJECT TRACKING

Performing object tracking in infrared video has received great interests in recent years [74]. In [75], an adaptive weighted patch-based appearance model is proposed to deal with non-rigid deformation for infrared tracking. Based on background subtraction, a novel multiple-target tracking-before-detection method with δ generalized labeled multi-Bernoulli filter is developed to track the objects as pixel set [76]. To exploit the powerful representation ability of convolutional neural network for appearance modeling in infrared video, multi-layer convolutional features are exploited in multi-correlation-filter-based infrared tracking [77]. In [78], the infrared tracking is treated as the similarity verification task and a hierachical spatial-aware Siamese network is developed. However, the appearance model of this model is developed only based on infrared modality, which may not be effective if the infrared modality is not reliable.

III. PROPOSED METHODS

The novel aspects of the proposed methods will be described in this section. First, the online non-negative multi-modality feature template learning model will be introduced, and then the optimization algorithm for model parameter estimation will be derived.

A. ONLINE NON-NEGATIVE MULTI-MODALITY FEATURE TEMPLATE LEARNING

To derive our model for multi-modality feature template learning, we need to consider what criteria would be used to guide the learning process. Considering that the learning feature template should have a good capability for appearance modeling, the first objective is that the learned feature templates should have good representation ability to model the target appearance variation. Let $Y^k = [y_1^k, \ldots, y_N^k] \in \mathbb{R}^{d^k \times N}$, $k = 1, \ldots, K$ denote the samples of the tracked object in RGB and infrared modalities collected by the tracker for model learning, where N is the number of samples, and M is the number of modalities (M = 2 for our problem). Then the first objective can be formulated as follows:

$$Y^k = D^k X^k + E^k, k = 1, \ldots, K \tag{1}$$

where $D^k = [D_1^k, \ldots, D_c^k] \in \mathbb{R}^{d^k \times c}$ denote the feature templates in the k-th modality which will be learned in our model, X^k is the reconstruction coefficient matrix which would be used to reconstruct the object using the linear combination of feature templates, E^k is the error term which would be used to capture the contaminated features of the samples of the k-th modality.

How to decontaminate the tracking samples and learn feature templates of multiple modalities for accurate representation of the tracked object under appearance variation, is the key problems which should be considered to derive our learning model. The learned feature template should be robust to different appearance variations to achieve a better representation accuracy. Therefore, for each feature template, it should encode some specific variation of the object appearance. As such, during the model learning procedure, some constraints should be imposed to adaptively activate (or select) informative template to handle the variation. In addition, the error term should be enforced to characterize the outliers caused by appearance variations for decontamination of the samples. Based on the aforementioned consideration, the multi-modality feature templates and the contaminated features can be estimated via solving the following optimization problem:

$$\begin{aligned} \min_{\{X^k, D^k, E^k\}} & \sum_{k=1}^{K} \begin{pmatrix} \frac{1}{2}\|Y^k - D^k X^k - E^k\|_2^2 \\ +\lambda_1 \|X^k\|_1 + \lambda_2 \|E^k\|_1 \end{pmatrix} \\ & s.t.\ (D_{(\cdot,i)}^k)^T D_{(\cdot,i)}^k \leq 1 \end{aligned} \tag{2}$$

where the first term $\sum_{k=1}^{K} \| \bullet \|_2^2$ which encode the reconstruction error, represents the total representation accuracy of the tracked object using the learned multi-modality feature templates, and the second and the third term are the sparsity constraint on the reconstruction coefficient matrix and the error terms using the ℓ_1 norm function. The proposed model is the integration multiple sparse representation- trackers of different modalities [96]–[100]. We can see that optmizing the first term can ensure that the appearance modeling with the learned feature templates can achieve as good accuracy as possible, minimizing the second term can enforce only a small number of feature templates will be selected for handing appearance variation, and minimizing the third term enable the error terms to capture the outliers in the contaminated features with the same merit in [101].

Inspired by the non-negative dictionary learning [102] and matrix factorization model [84], to make the learned feature templates able to capture the local part-based decomposition of the target appearance, we further introduce the non-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2916895, IEEE Access

negative constraint on the matrices of sparse coefficients and the feature templates as follows:

$$\min_{\{X^k,D^k,E^k\}} \sum_{k=1}^{K} \begin{pmatrix} \frac{1}{2}\|Y^k - D^k X^k - E^k\|_2^2 \\ +\lambda_1\|X^k\|_1 + \lambda_2\|E^k\|_1 \end{pmatrix} \quad (3)$$
$$s.t.\ X^k \geq \mathbf{0}, D^k \geq \mathbf{0}, (D^k_{(\cdot,i)})^T D^k_{(\cdot,i)} \leq 1$$

where $\mathbf{0}$ denotes all zeros' matrices with the same size of X^k and D^k, respectively.

Since some modalities may not be reliable and features from different modalities may contribute different roles for appearance modeling, we further incorporate an adaptive importance weight learning scheme into the feature learning model, which is shown as follows:

$$\min_{\{X^k,D^k,E^k\}} \sum_{k=1}^{K} \begin{pmatrix} \frac{1}{2}(\alpha^k)^2\|Y^k - D^k X^k - E^k\|_2^2 \\ +\lambda_1\|X^k\|_1 + \lambda_2\|E^k\|_1 \end{pmatrix} \quad (4)$$
$$s.t.\ X^k \geq \mathbf{0}, D^k \geq \mathbf{0}, \alpha^k \geq 0, \sum_{k=0}^{K} \alpha^k = 1$$
$$(D^k_{(\cdot,i)})^T D^k_{(\cdot,i)} \leq 1 \quad (5)$$

Here the important weight α is dynamically optimized and undated during the tracking process via minimizing the weighted sum of the reconstruction error, which ensures that feature templates from more reliable modality play more important role in the sparse representation. We use α^2 instead of α to avoid the trivial solution that the weights corresponding to the lowest reconstruction error is 1 and the other weights is 0.

Model Simplification The main focus of the learning model is to learn the feature templates and reliability weights for appearance modeling. However, to decontaminate the features of the tracking samples, solving the problem in (4) requires the estimation of $\{E^k\}$, which introduces more unknown variables and may increase the computational complexity. Therefore, derived a simplified model is required. Inspired by the online robust non-negative dictionary learning [102], we exploit the Huber loss function to remove the variable $\{E^k\}$ and model the reconstruction loss in (4), which is formulated as follows:

$$\min_{\{X^k,D^k,\alpha^k\}} L = \sum_{k=1}^{K} \begin{pmatrix} \frac{(\alpha^k)^2}{2} \sum_{i=1}^{d^k}\sum_{j=1}^{N} g_\theta\left((Y^k - D^k X^k)_{(ij)}\right) \\ +\lambda_1\|X^k\|_1 \end{pmatrix} \quad (6)$$
$$s.t.\ X^k \geq \mathbf{0}, \alpha^k \geq 0, \sum_{k=1}^{K} \alpha^k = 1, (D^k_{(\cdot,i)})^T D^k_{(\cdot,i)} \leq 1$$
$$D^k \geq \mathbf{0}$$

where $g_\theta(\bullet)$ is the Huber loss function, i.e.

$$g_\theta(a) = \begin{cases} \frac{1}{2}a^2 & |a| < \theta \\ \theta|a| - \frac{1}{2}\theta^2 & \text{else} \end{cases} \quad (7)$$

B. OPTIMIZATION

Since the proposed model involves three blocks of parameters $\{D^k\}$, $\{x^k\}$, and $\{\alpha^k\}$, we derive an iterative optimization algorithm to alternative update $\{D^k\}$, $\{x^k\}$, and $\{\alpha^k\}$.

$\{X^k\}$**-subproblem**: With fixed α^k and D^k, Problem (6) is separable, and solving each separated problem is equivalent to solve the following problem:

$$\min_{\{X^k\}} \frac{1}{2}\sum_{i=1}^{d^k}\sum_{j=1}^{N} g_\theta\left((Y^k - D^k X^k)_{(ij)}\right) + \lambda_1\|X^k\|_1 \quad (8)$$
$$s.t.\ X^k \geq \mathbf{0}$$

Following [102], we utilize the following updating rule to update X^k until convergence, i.e.

$$(X^k_{ij})^t = \frac{(X^k_{ij})^{t-1}\left[((W^k)^{t-1} \odot Y^k)^T D^k\right]_{ij}}{\left[((W^k)^{t-1} \odot (D^k((X^k)^{t-1})^T))^T D^k\right]_{ij} + \gamma} \quad (9)$$

where $(\cdot)^t$ denote the value of t-th iteration, $\odot$ denote the element-wise product, and $W^k = [w^k_{ij}]$ is

$$w^k_{ij} = \begin{cases} 1 & |r_{ij}| < \theta \\ \frac{\theta}{|r^k_{ij}|} & \text{else} \end{cases} \quad (10)$$

where $r^k_{ij} = Y^k_{ij} - D^k_{i,\cdot} X^k_{\cdot,k}$

$\{D^k\}$**-subproblem**: With fixed α^k and X^k we employ the projected gradient decent to update D^k. In $t+1$-th iteration, we utilize the surrogate function to express Huber loss as a weighted ℓ_2 loss function, and then aim to solve the following problem:

$$\min_{\{D^k\}} L = \frac{1}{2}\sum_{i=1}^{d^k}\sum_{j=1}^{N}(w^k_{ij})^t(Y^k_{ij} - D^k_{i,\cdot}X^k_{\cdot,j})^2 + \lambda_1\|X^k\|_1 \quad (11)$$
$$s.t.\ D^k \geq \mathbf{0}, (D^k_{(\cdot,i)})^T D^k_{(\cdot,i)} \leq 1$$

By taking the derivative of L in (11), we can obtain:

$$\frac{\partial l}{\partial (D^k_{i,\cdot})^T} = (U^k_i)^t (D^k_{i,\cdot})^T - (V^k_i)^t \quad (12)$$

where $(U^k_i)^t = \sum_{j=1}^{c}(w^k_{ij})^t (X^k_{\cdot,j})(X^k_{\cdot,j})^T$, and $(V^k_i)^t = \sum_{j=1}^{c}(w^k_{ij})^t (X^k_{\cdot,j}) y_{ij}$. Then the projected gradient decent can be performed by

$$\widehat{(D^k_{i,\cdot})^t} \leftarrow \max\left((D^k_{i,\cdot})^t - \tau(D^k_{i,\cdot})^t((U^k_i)^t)^T + \tau((V^k_i)^t)^T, \mathbf{0}\right) \quad (13)$$
$$(D^k_{\cdot,j})^{t+1} \leftarrow \frac{\widehat{(D^k_{\cdot,j})^t}}{\|\widehat{(D^k_{\cdot,j})^t}\|_2}$$

Here Eq.(13) first performs the gradient decent and then projects to the intersection of non-negative orthant and the Euclidean norm cone. The step size τ is set to 0.2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2916895, IEEE Access

$\{\alpha^k\}$**-subproblem**: With $\{X^k, D^k\}$ fixed, let $R^k = \frac{1}{2}\sum_{i=1}^{d^k}\sum_{j=1}^{N} g_\theta\left((Y^k - D^k X^k)_{(ij)}\right)$, then the problem in (6) can be reduced to

$$\min_{\{\alpha^k\}} \sum_{k=1}^{K} (\alpha^k)^2 R^k \quad s.t. \sum_{k=1}^{K} \alpha^k = 1, \alpha^k \geq 0,\ k = 1, \ldots, K \tag{14}$$

By taking the derivatives of the Lagrange function of (14) i.e. $\mathcal{L}\left(\{\alpha^k\}, \beta\right) = \sum_{k=1}^{K}(\alpha^k)^2 R^k + \beta(\sum_{k=1}^{K}\alpha^k - 1)$, and setting it to be zeros. we can obtain $\alpha^k r^k + \beta = 0$. Based on the equality $\sum_{k=1}^{K}\alpha^k = 1$, we can derive $\alpha^{k'} = \frac{(R^{k'})^{-1}}{\sum_{k=1}^{K}(R^k)^{-1}}$.

The optimization algorithm alteratively updates the three blocks of variables until convergence.

IV. IMPLEMENTATIONS

This section mainly introduces some key implementation details of the proposed tracking algorithm.

A. TARGET APPEARANCE REPRESENTATION AND POSITION DETERMINATION

The proposed tracker is implemented within the particle filtering framework. Based on the collected tracking samples and the background samples, following the implementation in [102], in order to remove the effect of cluttered background and preserve the informative features, we exploit the ℓ_1-regularized logistic regression to construct the feature selection mask P^k whose elements are 0 or 1. Given the learned multi-modality feature templates $D^k, k = 1, \ldots, K$ learned from the model (15) that encode distinctive properties of the target appearance, to enhance the discriminability of the tracking models, we further augmented the feature templates matrix D^k with some randomly sampled background samples B^k. After the feature selection masks are applied, we exploit them to represent the target candidates of RGB-infrared modalities. Since the multi-modality feature template is learned under non-negative and sparsity constraint, we estimate the reconstruction coefficients of the target candidate based on sparse representation model under the same constraint as follows:

$$\min_{\{x^k, z^k\}} \sum_{k=1}^{K} \frac{(\alpha^k)^2}{2} \sum_{i=1}^{d^k}\sum_{j=1}^{N} g_\theta\left(\left(P^k(y^k - [D^k x^k + B^k z^k])\right)_{(ij)}\right) + \lambda_1 \|X^k\|_1 \tag{15}$$

$$s.t.\ X^k \geq \mathbf{0}, \alpha^k \geq 0, \sum_{k=1}^{K}\alpha^k = 1$$

The sparse coding algorithm derived in Section III-B is utilized to solve (15). After obtaining the sparse coefficients x^k and z^k, the observation likelihood for each particles can be derived as

$$p(o_t^i | s_t^i) \propto \exp\left(\sum_{k=1}^{K} \alpha_k \left(\eta \|D^k x^k\|_1 - \|B^k z^k\|_1\right)\right) \tag{16}$$

where s_t^i denote the i-th particle in the t-th frame. The observation likelihood ensure that the the target particle should be the one which can be well represented by the learned multi-modality feature templates and is poorly represented by the background, which means a good discrimination between the target and the background can be achieved.

B. MODEL INITIALIZATION AND UPDATING

The bounding box of the tracked target is initialized manually according to the annotation data. To initialize the tracking model, we randomly sample 20 image patches which have small shift from the target position in the first frame as the positive examples, and 100 image patches as the negative examples. The reliability weights are initialized to be the same, i.e. 0.5.

Since the object appearance will change during the tracking process, and it would also encounter some appearance variation, the tracking model should be properly updated. Inspired by the online learning model [103], our model updating should preserve some historical appearance information of the tracked target to alleviate the risk of drifting. Therefore, we introduce the forgetting factor r to combine the historical information with the target appearance in current frame. After obtaining the tracking result in t-th frame, we update the U_i^k and V_i^k which can be regarded as the sufficient statistics as follows:

$$(U_i^k)^t \leftarrow r(U_i^k)^{t-1} + (w_i^k)^t((x^k)^t)((x^k)^t)^T \tag{17}$$

$$(V_i^k)^t \leftarrow r(V_i^k)^{t-1} + (w_i^k)^t((x^k)^t)((y_{i,.}^k)^t) \tag{18}$$

where $(y^k)^t$ denote the target sample in t-th frame of $k-$th modality, $(x^k)^t$ is the sparse codes of the target sample of of different modalities, $(U_i^k)^t$ and $(V_i^k)^t$ denote U_i^k and V_i^k in t-th frame. Depend on the degree of changes of the target appearance, following the idea of [102] the model updating is performed every 3 or 5 frames.

V. EXPERIMENTS

This section first introduces the experimental settings, and then presents the experimental results.

A. EXPERIMENTAL SETTINGS

Fifteen RGB-infrared video pairs which are captured by infrared and visible cameras are adopted to evaluate the tracking performance. These videos cover large appearance variations such as occlusion, large scale changes and poor illumination conditions. To make sure hat the tracked object appears at the same position in each video frame of infrared and RGB modalities, video frame alignment and registration have been applied on these videos. Totally 10 methods are run for comparison. They are STC [104], CT [105], MIL [106],

L1 [80], JSR [92], CN [107], KCF [108], MEEM [109], RPT [110], STUCK [111]. The JSR and L1 methods are developed for RGB-infrared tracking, while the other methods are originally developed for tracking in RGB modality. Following the implementation as introduced in [83], we can implement the multi-modality version of these trackers. Some of the tracking results for these multi-modality trackers on these videos data can be obtained from [83].

The θ, λ_1, and η is set to 0.01, 0.01 and 5, respectively. The forgetting parameter r for model updating is set to 0.99. In each frame, to sample the target candidates, 600 particles is generated within the framework of particle filtering.

B. EXPERIMENTAL RESULTS

We use two metrics i.e. overlapping rate and success rate to quantitatively evaluate the tracking performance. We define the overlapping rate as $\frac{area(B_1 \bigcap B_2)}{area(B_1 \bigcup B_2)}$ where B_1 and B_2 are the bounding box generated by the tracker and the groundtruth. A tracking success is counted if the overlapping rate measured from the tracking result in a video frame is greater than 0.5. The percentage of video frames in which the tracking success happens is regarded as success rate. The results of the compared trackers and our proposed method in terms of overlapping rate and success rate can be found in Tables 1 and 2. In summary, the quantitative results recorded in Tables 1 and 2 show that the proposed tracker obtains the best accuracy among all the compared trackers in the fifteen RGB-infrared videos. The performance of the proposed tracker stays in the rank of top 3 in 14 videos in terms of success rate and in 11 videos in terms of overlapping rate. Specifically, as shown in Figure 2 the proposed tracker is more able to handle to some large variations, such as occlusion (e.g. *Exposure2*#28, *FastCar2*#26), thermal crossover (e.g. *Gathering*#256, cluttered background (e.g. *BlueCar*, *BusScale*) This is because the proposed model can explicitly decontaminate the training samples during the template learning process, which make it more robust to outlier caused by large appearance variation. In addition, the adaptively determined reliability weight enables more reliable modality contribute more in appearance modeling. The adaptive integration of RGB modality make it less sensitive to the issues of thermal crossover in infrared tracking.

Figure 3 show some qualitative comparison of the overlapping rate of the compared trackers in a frame-by-frame manner. It can be found that the proposed tracker achieves a relatively higher overlapping rate in general.

VI. CONCLUSION

In this paper, we propose an online multi-modality feature template learning model for infrared tracking with RGB information. By integrating multi-modality feature learning and fusion, feature decontamination, and modality reliability evaluation into a unified optimization framework, the proposed infrared tracker can achieve a better tracking results. To reduce the computational complexity, we further derive the simplified but simplified forms of the learning and the corresponding optimization algorithm. Comparison experimental results with other 10 trackers shows the effectiveness of the proposed tracker.

Since the proposed algorithm can not run in real time, one of our future work will be focus on how to improve the tracking efficiency. There are two directions which can be further explored. First, we can develop more efficient optimization algorithms to obtain the optimal solution. Second, we can exploit more scientific computation techniques (e.g paralleled computing) with advanced hardware (e.g. GPU) to increase the efficiency.

REFERENCES

[1] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. Siow Mong Goh, "Anomalynet: An anomaly detection network for video surveillance," IEEE Trans. Information Forensics and Security, pp. 1–1, 02 2019, DOI:10.1109/TIFS.2019.2900907.

[2] W. Zhang, W. Zhang, and J. Gu, "Edge-semantic learning strategy for layout estimation in indoor environment," IEEE Transactions on Cybernetics, 2019, DOI:10.1109/TCYB.2019.2895837.

[3] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," IEEE Trans. Pattern Anal. Mach. Intell., 2018, DOI:10.1109/TPAMI.2018.2875002.

[4] S. Zhang, X. Lan, H. Yao, H. Zhou, D. Tao, and X. Li, "A biologically inspired appearance model for robust visual tracking," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 10, pp. 2357–2370, 2017.

[5] Y. Qi, L. Qin, S. Zhang, Q. Huang, and H. Yao, "Robust visual tracking via scale-and-state-awareness," Neurocomputing, vol. 329, pp. 75–85, 2019.

[6] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in Proc. IJCAI, 2018, pp. 1092–1099.

[7] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," IEEE Trans. Image Process., 2019, DOI:10.1109/TIP.2019.2893066.

[8] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," IEEE Trans. Circuits Syst. Video Techn., 2017, DOI: 10.1109/TCSVT.2017.2718188.

[9] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in Proc. AAAI, 2018, pp. 7501–7508.

[10] S. Zhang, X. Lan, Y. Qi, and P. C. Yuen, "Robust visual tracking via basis matching," IEEE Trans. Circuits Syst. Video Techn., vol. 27, no. 3, pp. 421–430, 2017.

[11] W. Zhang, S. Hu, K. Liu, and Z. Zha, "Compact appearance learning for video-based person re-identification," IEEE Trans. Circuits Syst. Video Techn., 2018, DOI:10.1109/TCSVT.2018.2865749.

[12] J. Li, A. J. Ma, and P. C. Yuen, "Semi-supervised region metric learning for person re-identification," International Journal of Computer Vision, vol. 126, no. 8, pp. 855–874, 2018.

[13] W. Zhang, B. Ma, K. Liu, and R. Huang, "Video-based pedestrian re-identification by adaptive spatio-temporal appearance model," IEEE Trans. on Image Processing, vol. 26, no. 4, pp. 2042–2054, 2017.

[14] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in Proc. ECCV, 2018, pp. 2651–2664.

[15] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," IEEE Trans. Multimedia, vol. 18, no. 12, pp. 2553–2566, 2016.

[16] X. Chang, Y. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 8, pp. 1617–1632, 2017.

[17] Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, J. Lim, and M.-H. Yang, "Hedging deep features for visual tracking," IEEE Trans. Pattern Anal. Mach. Intell., 2018.

[18] B. Zhong, B. Bai, J. Li, Y. Zhang, and Y. Fu, "Hierarchical tracking by reinforcement learning-based searching and coarse-to-fine verifying," IEEE Trans. Image Processing, vol. 28, no. 5, pp. 2331–2341, 2019.

TABLE 1: Overlapping Rate. The best three results are shown in red, blue and green.

Sequence	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	JSR	L1	**Proposed Method**
Exposure4	**0.57**	0.35	0.37	0.23	**0.66**	0.55	0.2	0.56	0.56	0.51	**0.63**
Gathering	**0.77**	0.36	0.58	0.69	**0.77**	0.33	**0.81**	0.37	0.22	0.19	0.72
BlackCar	0.24	0.31	0.21	0.22	**0.33**	0.23	0.21	0.24	0.23	**0.64**	**0.64**
BlueCar	0.37	0.27	0.34	0.4	**0.65**	0.47	0.4	0.4	0.4	**0.63**	**0.78**
BusScale	0.47	0.45	0.46	0.49	**0.57**	0.52	0.51	0.51	0.54	**0.72**	**0.83**
Exposure2	0.32	0.37	0.31	0.32	**0.48**	0.3	0.32	0.32	0.35	**0.82**	**0.78**
FastCar2	**0.57**	0.53	0.43	0.48	0.51	0.49	0.5	0.54	**0.56**	0.34	**0.65**
FastCarNig	0.46	**0.75**	0.36	0.36	0.63	0.41	0.43	0.43	0.38	**0.75**	**0.66**
Tricycle	0.68	0.64	0.62	**0.71**	**0.72**	**0.73**	0.64	0.64	0.67	0.57	0.57
Minibus1	0.53	0.05	0.52	0.55	0.06	0.38	**0.56**	0.05	0.53	**0.69**	**0.74**
Motorbike	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	0.3	**0.31**	**0.31**	0.3	**0.5**	**0.7**
Cycling	**0.62**	0.47	0.51	**0.64**	0.55	0.03	0.61	**0.63**	0.49	0.36	0.53
MinibusNig	0.54	0.55	0.54	0.55	**0.68**	0.55	0.57	0.59	0.33	**0.74**	**0.75**
Crossing	**0.8**	0.62	0.67	**0.81**	**0.8**	**0.8**	**0.81**	0.79	0.64	0.65	0.74
BusScale1	0.4	0.41	0.43	0.39	**0.67**	0.44	0.43	0.42	0.47	**0.66**	**0.74**
Average	0.51	0.43	0.44	0.48	**0.56**	0.44	0.49	0.45	0.44	**0.58**	**0.7**

TABLE 2: Success Rate. The best three results are shown in red, blue and green.

Sequence	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	JSR	L1	**Proposed Method**
Exposure4	0.63	0.47	0.18	0.25	**0.9**	0.56	0.21	0.59	0.64	**0.73**	**0.83**
Gathering	0.97	0.35	0.84	0.97	**1**	0.3	**1**	0.37	0.09	0.14	**0.99**
BlackCar	0.12	0.16	0.1	0.12	**0.29**	0.12	0.12	0.12	0.15	**0.83**	**0.84**
BlueCar	0.33	0.33	0.28	0.38	**0.94**	0.46	0.38	0.38	0.44	**0.68**	**1**
BusScale	0.48	0.4	0.46	0.44	**0.61**	0.53	0.5	0.51	0.56	**0.82**	**1**
Exposure2	0.2	0.26	0.2	0.2	**0.45**	0.16	0.2	0.2	0.19	**1**	**1**
FastCar2	**0.55**	0.48	0.35	0.43	0.48	0.5	0.53	**0.55**	**0.57**	0.45	**0.8**
FastCarNig	0.31	**0.93**	0.28	0.28	0.73	0.26	0.28	0.28	0.39	**1**	**0.74**
Tricycle	0.98	0.72	**0.99**	**1**	**1**	0.98	0.85	0.75	0.93	0.56	0.68
Minibus1	**0.59**	0.04	0.54	0.58	0.05	0.32	0.54	0.04	0.49	**0.69**	**0.99**
Motorbike	0.14	**0.16**	0.14	0.13	0.13	0.12	0.14	0.14	0.12	**0.48**	**0.98**
Cycling	**0.71**	0.43	0.53	**0.71**	0.68	0.02	**0.71**	**0.71**	0.48	0.33	**0.71**
MinibusNig	0.51	0.49	0.55	0.51	**0.92**	0.51	0.54	0.55	0.36	**1**	**1**
Crossing	**1**	0.82	**1**	**1**	**1**	**1**	**1**	**1**	0.79	0.95	**1**
BusScale1	0.33	0.34	0.36	0.27	**0.87**	0.45	0.36	0.36	0.47	**0.76**	**0.87**
Average	0.52	0.43	0.45	0.48	**0.67**	0.42	0.49	0.44	0.44	**0.69**	**0.9**

[19] Q. Zhou, B. Zhong, Y. Zhang, J. Li, and Y. Fu, "Deep alignment network based multi-person tracking with occlusion and motion reasoning," IEEE Trans. Multimedia, 2018, DOI:10.1109/TMM.2018.2875360.

[20] W. Zhang, Y. Li, W. Lu, X. Xu, Z. Liu, and X. Ji, "Learning intra-video difference for person re-identification," IEEE Trans. Circuits Syst. Video Techn., 2018, DOI:10.1109/TCSVT.2018.2872957.

[21] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," IEEE Trans. Cybernetics, vol. 47, no. 5, pp. 1180–1197, 2017.

[22] Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, and M. Yang, "Structure-aware local sparse coding for visual tracking," IEEE Trans. Image Processing, vol. 27, no. 8, pp. 3857–3869, 2018.

[23] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," IEEE Trans. Image Processing, vol. 28, no. 1, pp. 291–301, 2019.

[24] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," Neurocomputing, vol. 291, pp. 187–194, 2018.

[25] J. Zheng, X. Cao, B. Zhang, X. Zhen, and X. Su, "Deep ensemble machine for video classification," IEEE Trans. Neural Netw. Learning Syst., vol. 30, no. 2, pp. 553–565, 2019.

[26] M. Pang, Y.-m. Cheung, B. Wang, and R. Liu, "Robust heterogeneous discriminative analysis for face recognition with single sample per person," Pattern Recognition, vol. 89, pp. 91–107, 2019.

[27] D. K. Jain, Z. Zhang, and K. Huang, "Random walk-based feature learning for micro-expression recognition," Pattern Recognition Letters, vol. 115, pp. 92–100, 2018.

[28] G. Ding, Y. Guo, K. Chen, C. Chu, J. Han, and Q. Dai, "DECODE: Deep confidence network for robust image classification," IEEE Transactions on Image Processing, 2019, DOI:10.1109/TIP.2019.2902115.

[29] Y. Zhang, K. N. Ngan, L. Ma, and H. Li, "Objective quality assessment of image retargeting by incorporating fidelity measures and inconsistency detection," IEEE Trans. Image Processing, vol. 26, no. 12, pp. 5980–5993, 2017.

[30] W. Zhang, W. Zhang, K. Liu, and J. Gu, "A feature descriptor based on local normalized difference for real-world texture classification," IEEE Trans. Multimedia, vol. 20, no. 4, pp. 880–888, 2018.

[31] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," IEEE Trans. Geoscience and Remote Sensing, vol. 57, no. 2, pp. 911–923, 2019.

[32] J. T. Zhou, I. W. Tsang, S.-s. Ho, and K.-R. Müller, "N-ary decomposition for multi-class classification," Machine Learning, DOI:10.1007/s10994-019-05786-2.

[33] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," Pattern Recognition Letters, 2017.

[34] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, "Spatial sequential recurrent neural network for hyperspectral image classification," IEEE J Sel. Topics in Appl. Earth Observ. and Remote Sensing, vol. 11, no. 11, pp. 4141–4155, 2018.

[35] F. Liu, J. Wu, L. Li, L. Jiao, H. Hao, and X. Zhang, "A hybrid method of SAR speckle reduction based on geometric-structural block and adaptive neighborhood," IEEE Trans. Geoscience and Remote Sensing, vol. 56, no. 2, pp. 730–748, 2018.

[36] X. Zhang, Z. Gao, L. Jiao, and H. Zhou, "Multifeature hyperspectral image classification with local and nonlocal spatial information via

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2916895, IEEE Access

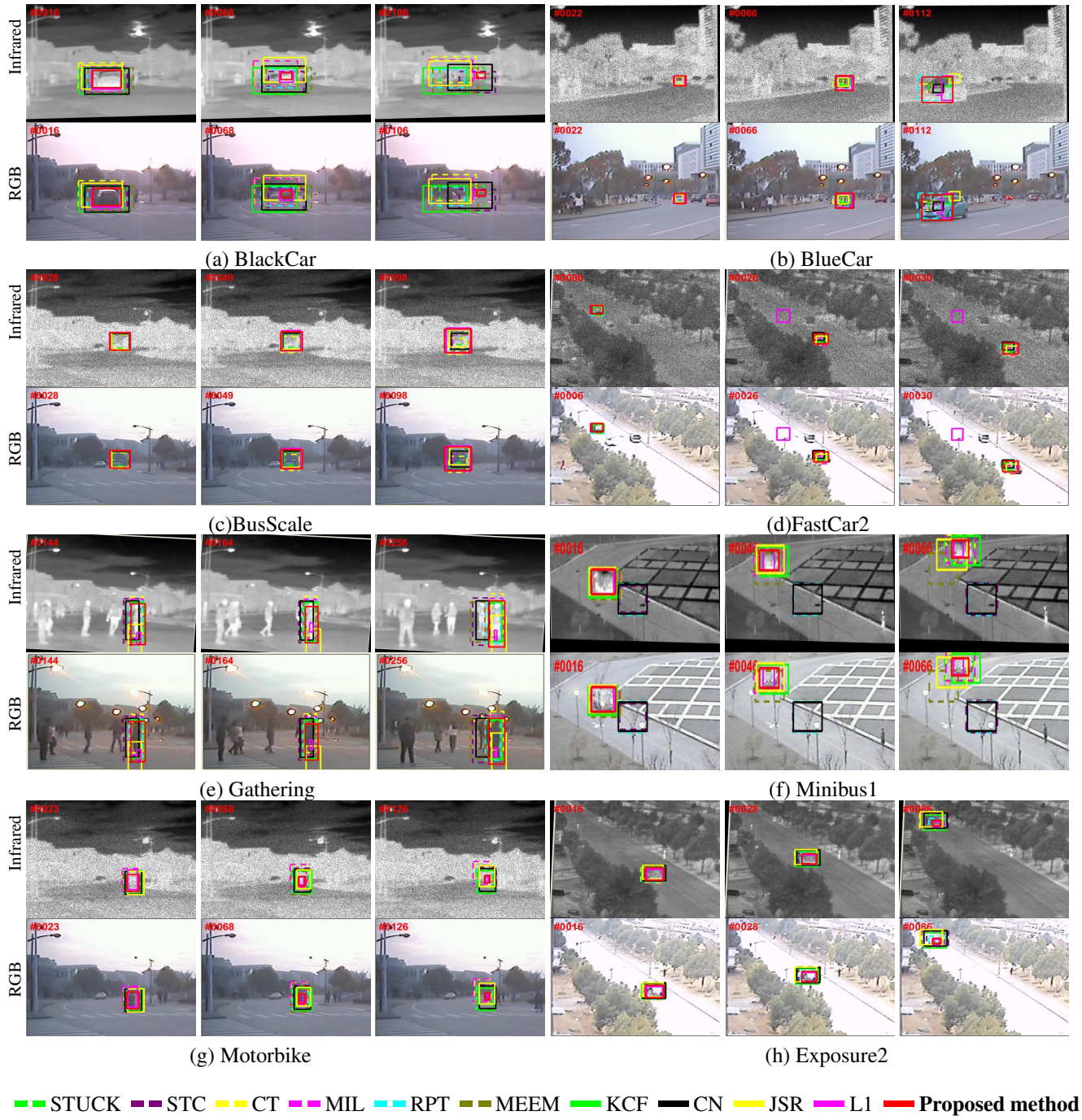

FIGURE 2: Qualitative comparison results on some video frames of RGB and infrared modality under some challenging situations, which includes large scale variation (e.g. *BusScale, Minibus1*), poor illumination conditions(e.g. *Exposure2*), thermal crossover (e.g. *Motorbike, Gathering*), occlusion (e.g. *BlackCar*). For each sub-figure, video frames of RGB modality are shown in the top row while video frames of infrared modality are shown in the bottom one.

markov random field in semantic space," IEEE Trans. Geoscience and Remote Sensing, vol. 56, no. 3, pp. 1409–1424, 2018.

[37] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, and L. Shao, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1993–2007, 2019.

[38] J. T. Zhou, H. Zhao, X. Peng, M. Fang, Z. Qin, and R. S. M. Goh, "Transfer hashing: From shallow to deep," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 12, p. 6191Ű6201, 2018.

[39] S. Ren, D. K. Jain, K. Guo, T. Xu, and T. Chi, "Towards efficient medical lesion image super-resolution based on deep residual networks," Signal Processing: Image Communication, vol. 75, pp. 1 – 10, 2019, DOI:10.1016/j.image.2019.03.008.

[40] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in Proc. CVPR, 2016, pp. 4792–4800.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2916895, IEEE Access

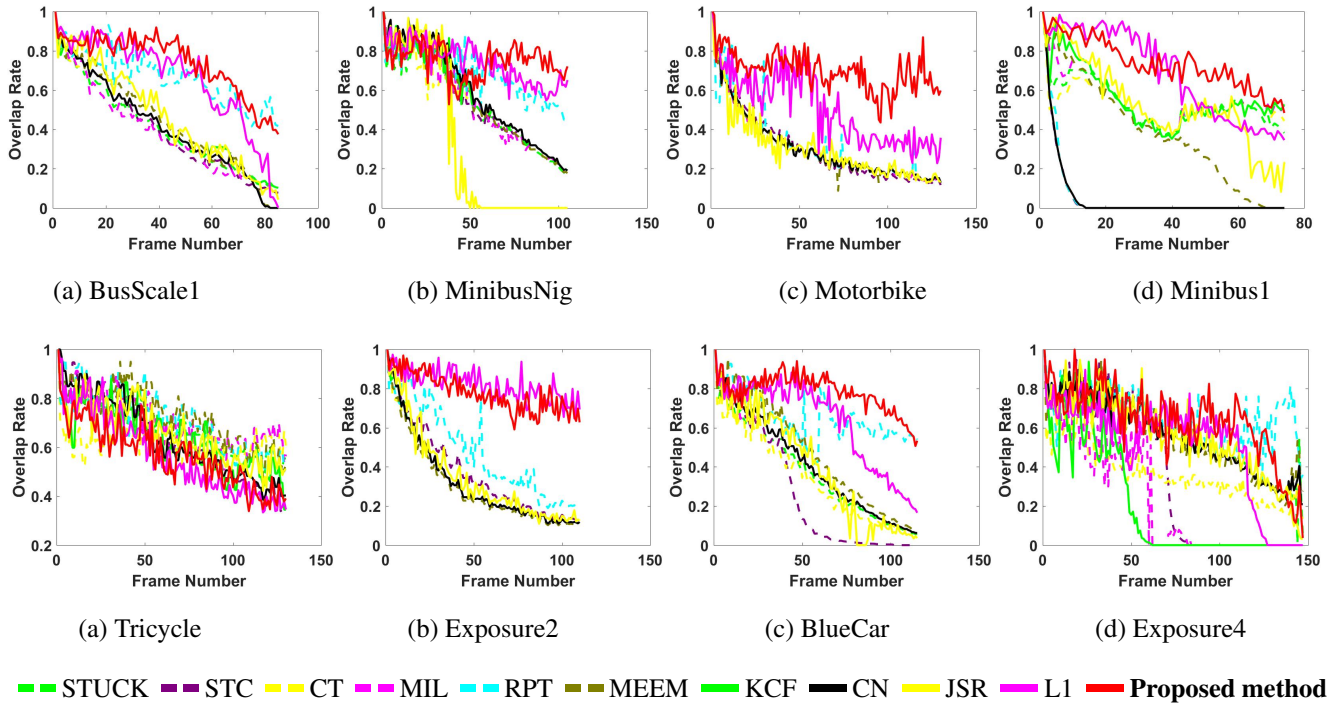

FIGURE 3: Frame-by-Frame Quantitative Comparison of 11 trackers on 8 challenging videos in terms of overlapping rate. The vertical axis indicates the overlapping rate and the horizontal axis is the frame index

[41] X. Zhang, J. Zhang, C. Li, C. Cheng, L. Jiao, and H. Zhou, "Hybrid unmixing based on adaptive region segmentation for hyperspectral imagery," IEEE Trans. Geoscience and Remote Sensing, vol. 56, no. 7, pp. 3861–3875, 2018.

[42] J. An, X. Zhang, H. Zhou, and L. Jiao, "Tensor-based low-rank graph with multimanifold regularization for dimensionality reduction of hyperspectral images," IEEE Trans. Geoscience and Remote Sensing, vol. 56, no. 8, pp. 4731–4746, 2018.

[43] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K. Ma, "ESIM: edge similarity for screen content image quality assessment," IEEE Trans. Image Processing, vol. 26, no. 10, pp. 4818–4831, 2017.

[44] Z. Ni, H. Zeng, L. Ma, J. Hou, J. Chen, and K. Ma, "A gabor feature-based quality assessment model for the screen content images," IEEE Trans. Image Processing, vol. 27, no. 9, pp. 4516–4528, 2018.

[45] W. Ren, J. Zhang, X. Xu, L. Ma, X. Cao, G. Meng, and W. Liu, "Deep video dehazing with semantic segmentation," IEEE Trans. Image Processing, vol. 28, no. 4, pp. 1895–1908, 2019.

[46] W. Zhang, K. Song, X. Rong, and Y. Li, "Coarse-to-fine uav target tracking with deep reinforcement learning," IEEE Trans. Automation Science and Engineering, 2018, DOI:10.1109/TASE.2018.2877499.

[47] B. Zhang, A. Perina, C. Li, Q. Ye, V. Murino, and A. Del Bue, "Manifold constraint transfer for visual structure-driven optimization," Pattern Recognition, vol. 77, pp. 87–98, 2018.

[48] M. Pang, B. Wang, Y.-M. Cheung, and C. Lin, "Discriminant manifold learning via sparse coding for robust feature extraction," IEEE Access, vol. 5, pp. 13 978–13 991, 2017.

[49] R. Shao, X. Lan, and P. C. Yuen, "Feature constrained by pixel: Hierarchical adversarial deep domain adaptation," in ACM MM, 2018, pp. 220–228.

[50] Y. Wang, C. Xu, C. Xu, and D. Tao, "Packing convolutional neural networks in the frequency domain," IEEE Trans. Pattern Anal. Mach. Intell., DOI:10.1109/TPAMI.2017.2780094.

[51] M. Pang, Y.-m. Cheung, R. Liu, J. Lou, and C. Lin, "Toward efficient image representation: Sparse concept discriminant matrix factorization," IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[52] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," IEEE Trans. Image Processing, vol. 27, no. 8, pp. 4066–4079, 2018.

[53] W. Zhang, Q. Chen, W. Zhang, and X. He, "Long-range terrain perception using convolutional neural networks," Neurocomputing, vol. 275, pp. 781–787, 2018.

[54] Q. Wang, Z. Qin, F. Nie, and X. Li, "Spectral embedded adaptive neighbors clustering," IEEE Trans. Neural Netw. Learn. Syst., 2018, DOI:10.1109/TNNLS.2018.2861209.

[55] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured autoencoders for subspace clustering," IEEE Trans Image Process, vol. 27, no. 10, pp. 5076–5086, Oct 2018.

[56] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, and X. Li, "Hierarchical feature selection for random projection," IEEE Trans. Neural Netw. Learn. Syst., 2018, DOI:10.1109/TNNLS.2018.2868836.

[57] X. Chang, Z. Ma, M. Lin, Y. Yang, and A. G. Hauptmann, "Feature interaction augmented sparse learning for fast kinect motion detection," IEEE Trans. Image Processing, vol. 26, no. 8, pp. 3911–3920, 2017.

[58] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," IEEE Trans. Image Processing, vol. 25, no. 7, pp. 3249–3260, 2016.

[59] Z. Huang, H. Zhu, J. T. Zhou, and X. Peng, "Multiple marginal fisher analysis," IEEE Trans Industr Electron, 2018, DOI:10.1109/TIE.2018.2870413.

[60] J. T. Zhou, M. Fang, H. Zhang, C. Gong, X. Peng, Z. Cao, and R. S. M. Goh, "Learning with annotation of various degrees," IEEE Trans. Neural Netw. Learning Syst., DOI:10.1109/TNNLS.2018.2885854.

[61] X. Peng, C. Lu, Y. Zhang, and H. Tang, "Connections between nuclear norm and frobenius norm based representation," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 1, pp. 218–224, 2018.

[62] C. Gong, D. Tao, W. Liu, L. Liu, and J. Yang, "Label propagation via teaching-to-learn and learning-to-teach," IEEE Trans. Neural Netw. Learning Syst., vol. 28, no. 6, pp. 1452–1465, 2017.

[63] C. Gong, T. Liu, J. Yang, and D. Tao, "Large-margin label-calibrated support vector machines for positive and unlabeled learning," IEEE Transactions on Neural Networks and Learning Systems, 2019, DOI:10.1109/TNNLS.2019.2892403.

[64] D. K. Jain, A. Kumar, S. R. Sangwan, G. N. Nguyen, and P. Tiwari, "A particle swarm optimized learning model of fault classification in web-apps," IEEE Access, vol. 7, pp. 18 480–18 489, 2019.

[65] R. Shao, X. Lan, and P. C. Yuen, "Joint discriminative learning of deep

dynamic textures for 3d mask face anti-spoofing," IEEE Trans. Inf. Forensics Security, vol. 14, no. 4, pp. 923–938, 2019.

[66] Y. Li, H. Lu, J. Li, X. Li, Y. Li, and S. Serikawa, "Underwater image de-scattering and classification by deep neural network," Computers & Electrical Engineering, vol. 54, pp. 68–77, 2016.

[67] H. Lu, Y. Li, S. Nakashima, H. Kim, and S. Serikawa, "Underwater image super-resolution by descattering and fusion," IEEE Access, vol. 5, pp. 670–679, 2017.

[68] H. Lu, Y. Zhang, Y. Li, Q. Zhou, R. Tadoh, T. Uemura, H. Kim, and S. Serikawa, "Depth map reconstruction for underwater kinect camera using inpainting and local image mode filtering," IEEE Access, vol. 5, pp. 7115–7122, 2017.

[69] H. Lu, T. Uemura, D. Wang, J. Zhu, Z. Huang, and H. Kim, "Deep-sea organisms tracking using dehazing and deep learning," Mobile Networks and Applications, DOI:10.1007/s11036-018-1117-9.

[70] W. Zhao, H. Lu, and D. Wang, "Multisensor image fusion and enhancement in spectral total variation domain," IEEE Trans. Multimedia, vol. 20, no. 4, pp. 866–879, 2018.

[71] Z. Liu, Z. Wang, H. Lu, and D. Wang, "Online vehicle tracking in aerial imagery," in Proc. IScIDE, 2017, pp. 335–345.

[72] X. Zhu, X. Song, X. Chen, Y. Bai, and H. Lu, "Size aware correlation filter tracking with adaptive aspect ratio estimation," TIIS, vol. 11, no. 2, pp. 805–825, 2017.

[73] X. Wang, B. Fan, S. Chang, Z. Wang, X. Liu, D. Tao, and T. S. Huang, "Greedy batch-based minimum-cost flows for tracking multiple objects," IEEE Trans. Image Processing, vol. 26, no. 10, pp. 4765–4776, 2017.

[74] M. Felsberg, A. Berg, G. Hager, J. Ahlberg, M. Kristan, J. Matas, A. Leonardis, L. Cehovin, G. Fernandez, T. Vojír et al., "The thermal infrared visual object tracking vot-tir2015 challenge results," in Proc. ECCV Workshop, 2015, pp. 76–88.

[75] J. Wang and J. Zhang, "Robust object tracking in infrared video via adaptive weighted patches," Mathematical and Computational Applications, vol. 22, no. 1, p. 3, 2017.

[76] C. Li and W. Wang, "Detection and tracking of moving targets for thermal infrared video sequences," Sensors, vol. 18, no. 11, 2018.

[77] Q. Liu, X. Lu, Z. He, C. Zhang, and W.-S. Chen, "Deep convolutional neural networks for thermal infrared object tracking," Knowledge-Based Systems, vol. 134, pp. 189–198, 2017.

[78] X. Li, Q. Liu, N. Fan, Z. He, and H. Wang, "Hierarchical spatial-aware siamese network for thermal infrared object tracking," Knowledge-Based Systems, vol. 166, pp. 71 – 81, 2019.

[79] M. S. W. S. J. Paul Retief, C. J. Willers, "Prediction of thermal crossover based on imaging measurements over the diurnal cycle," Proc.SPIE, vol. 5097, pp. 5097 – 5097 – 12, 2003.

[80] 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,, 2011, pp. 1–8.

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

[82] C. Ó. Conaire, N. E. O'Connor, and A. F. Smeaton, "Thermo-visual feature fusion for object tracking using multiple spatiogram trackers," Mach. Vis. Appl., vol. 19, no. 5-6, pp. 483–494, 2008.

[83] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," IEEE Trans. Image Processing, vol. 25, no. 12, pp. 5743–5756, 2016.

[84] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Advances in Neural Information Processing Systems 13, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., 2001, pp. 556–562.

[85] S. Zhang, H. Yao, X. Sun, and X. Lu, "Sparse coding based visual tracking: Review and experimental comparison," Pattern Recognit., vol. 46, no. 7, pp. 1772–1788, 2013.

[86] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. v. d. Hengel, "A survey of appearance models in visual object tracking," ACM Trans. Intell. Syst. Technol., vol. 4, no. 4, pp. 58:1–58:48, 2013.

[87] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," IEEE Trans. Image Process., vol. 24, no. 12, pp. 5630–5644, 2015.

[88] Y. Wu, J. Lim, and M. Yang, "Object tracking benchmark," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1834–1848, 2015.

[89] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," Neurocomputing, vol. 74, no. 18, pp. 3823–3831, 2011.

[90] 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. WACV, 2007.

[91] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," Proceedings of the IEEE, vol. 98, no. 6, pp. 1031–1044, 2010.

[92] 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, 2012.

[93] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for rgb-infrared object tracking," Pattern Recogn. Lett., 2018, DOI:10.1016/j.patrec.2018.10.002.

[94] X. Lan, M. Ye, S. Zhang, and P. C. Yuen, "Robust collaborative discriminative learning for rgb-infrared tracking," in Proc. AAAI, 2018, pp. 7008–7015.

[95] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust rgb-infrared tracking system," IEEE Trans. Industrial Electronics, DOI: 10.1109/TIE.2019.2898618.

[96] X. Lan, A. J. Ma, and P. C. Yuen, "Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation," in Proc. CVPR, 2014, pp. 1194–1201.

[97] X. Lan, A. Ma, P. Yuen, and R. Chellappa, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," IEEE Trans. Image Process., vol. 24, no. 12, pp. 5826–5841, Dec 2015.

[98] X. Lan, P. C. Yuen, and R. Chellappa, "Robust MIL-based feature template learning for object tracking," in Proc. AAAI, 2017, pp. 4118–4125.

[99] X. Lan, S. Zhang, and P. C. Yuen, "Robust joint discriminative feature learning for visual tracking," in Proc. IJCAI, 2016, pp. 3403–3410.

[100] X. Lan, S. Zhang, P. C. Yuen, and R. Chellappa, "Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker," IEEE Trans. Image Processing, vol. 27, no. 4, pp. 2022–2037, 2018.

[101] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 210–227, 2009.

[102] N. Wang, J. Wang, and D.-Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," in Proc. ICCV, 2013, pp. 657–664.

[103] J. Mairal, F. R. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," Journal of Machine Learning Research, vol. 11, pp. 19–60, 2010.

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

[105] K. Zhang, L. Zhang, and M.-H. Yang, "Fast compressive tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 10, pp. 2002–2015, 2014.

[106] B. Babenko, M. 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, 2011.

[107] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in Proc. CVPR. IEEE, 2014, pp. 1090–1097.

[108] 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, 2015.

[109] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in Proc. ECCV, 2014, pp. 188–203.

[110] Y. Li, J. Zhu, and S. C. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in Proc. CVPR, 2015, pp. 353–361.

[111] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M. Cheng, S. L. Hicks, and P. H. S. Torr, "Struck: Structured output tracking with kernels," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 10, pp. 2096–2109, 2016.

XIANGYUAN LAN received the B.Eng. degree in computer science and technology from the South China University of Technology, China, in 2012, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong in 2016. He is currently a Research Assistant Professor with Hong Kong Baptist University. His current research interests include intelligent video surveillance and biometric security.

MANG YE received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2013 and in 2016. He is currently a Ph.D student at Department of Computer Science, Hong Kong Baptist University. His research interests focus on multimedia content analysis and retrieval, computer vision and pattern recognition.

RUI SHAO received the B.Eng. degree in Electronic Information Engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. He is currently pursuing the Ph.D degree in Computer Science from Hong Kong Baptist University, Hong Kong. His current research interests include computer vision and pattern recognition.

BINENG ZHONG received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively. From 2007 to 2008, he was a Research Fellow with the Institute of Automation and Institute of Computing Technology, Chinese Academy of Science. From September 2017 to September 2018, he was a visiting scholar in Northeastern University, Boston, MA, USA. Currently, he is an professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. His current research interest is computer vision.

DEEPAK KUMAR JAIN is working as an Assistant Professor at Institute of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China . He received the Bachelor of Engineering degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, India, in 2010, the Master of Technology degree from the Jaypee University of Engineering and Technology, India, in 2012, and the Ph.D. degree from the Institute of Automation, University of Chinese Academy of Sciences, Beijing, China. He was an awardee of CAS-TWAS Presidential fellowship from 2014-2018. He was invited as "Foreign Experts" by Shandong Taian Administration of foreign Expert Affairs. He has presented several papers in peer-reviewed conferences and has published numerous studies in science cited journals. His research interests include deep learning, machine learning, pattern recognition, and computer vision

HUIYU ZHOU received a Bachelor of Engineering degree in Radio Technology from Huazhong University of Science and Technology of China, and a Master of Science degree in Biomedical Engineering from University of Dundee of United Kingdom, respectively. He was awarded a Doctor of Philosophy degree in Computer Vision from Heriot-Watt University, Edinburgh, United Kingdom. Dr. Zhou currently is a Reader at Department of Informatics, University of Leicester, United Kingdom.

Dr. Zhou serves as the Editor-in-Chief of Recent Advances in Electrical and Electronic Engineering and Associate Editor of IEEE Transaction on Human-Machine Systems, and is on the Editorial Boards of several refereed journals. He has authored over 180 peer-reviewed papers in the field. His research work has been or is being supported by U.K. EPSRC, MRC, EU, Royal Society, Leverhulme Trust, Puffin Trust, Invest NI, and industry.