Learning Modality-Consistency Feature Templates: A Robust RGB-Infrared Tracking System

Xiangyuan Lan, Mang Ye, Rui Shao, Bineng Zhong, Pong C. Yuen, *Senior Member*, and Huiyu Zhou

***Abstract*—With a large number of video surveillance systems installed for the requirement from industrial security, the task of object tracking, which aims to locate objects of interest in videos, is very important. Although numerous tracking algorithms for RGB videos have been developed in the decade, the tracking performance and robustness of these systems may be degraded dramatically when the information from RGB video is unreliable (e.g. poor illumination conditions or very low resolution). To address this issue, this paper presents a new tracking system which aims to combine the information from RGB and infrared modalities for object tracking. The proposed tracking systems is based on our proposed machine learning model. Particularly, the learning model can alleviate the modality discrepancy issue under the proposed modality consistency constraint from both representation patterns and discriminability, and generate discriminative feature templates for collaborative representations and discrimination in heterogeneous modalities. Experiments on a variety of challenging RGB-infrared videos demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—Multimodal sensor fusion, tracking system, video surveillance system.**

I. Introduction

DEVELOPING an intelligent video analysis system is very important for many industrial applications, such as burglar alarms, entrance systems, transportation management, etc. A key task of an intelligent video analysis system is to perform motion perception of the objects of interest. To this end, developing a robust object tracking system for object localization and moving status inference is very important. Visual tracking still remains a challenging task due to many unpredictable variations and poor environmental conditions such as severe occlusion, large illumination changes, dim lighting, low image resolution, etc. Efforts have been made in this field, and the decade has witnessed numerous tracking algorithms proposed to handle a variety of research issues [1]–[29]. However, most of these tracking algorithms are developed for object tracking in RGB videos, and they extract some visual features from RGB video frames for appearance modeling of the tracked object. When the information from RGB videos in not reliable (e.g. under poor lighting conditions), these trackers may fail to track the objects stably. This would limit them to be employed in practical industrial systems (e.g. a video surveillance system can be used for security monitoring in the night time).

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

Fig. 1. Illustration of some video frames from RGB and infrared modalities. **Top**: RGB **Bottom**: infrared

With the increasingly economical and affordable cost of multispectral sensors, equipping industrial systems with dual-camera systems, which contain both thermal infrared and visible, has become more and more popular. In addition, the rapid development of multispectral imaging techniques make it more effective to capture images or videos in RGB and infrared modalities for many industrial applications. Different from visible spectrum cameras, an infrared camera forms images by capturing the infrared radiation of a subject instead of using visible light. Therefore, they are more effective in imaging under poor lighting conditions. However, when the temperature of the background and the tracked objects are similar (i.e. infrared radiations are similar), infrared images may be of low quality and the RGB information may be more effective. Therefore, to develop a robust tracking system for practical industrial applications and perform effective object tracking in challenging practical scenarios, combining information from the RGB and infrared modalities for appearance modeling of the tracked object is necessary.

Therefore, to effectively perform object tracking in RGB-infrared modalities, a key problem is how to appropriately

This work was supported in part by Hong Kong Research Grants Council RGC/HKBU12254316 and 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 Grant 61572205.

X. Lan, M. Ye, R. Shao and P. C. Yuen are with the Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, PR China (e-mail: xiangyuanlan@life.hkbu.edu.hk; {mangye, ruishao, pcyuen}@comp.hkbu.edu.hk).

B. Zhong is with School of Computer Science and Technology, Huaqiao University, Xiamen 362100, China (e-mail:bnzhong@hqu.edu.cn)

H. Zhou is with the Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: hz143@leicester.ac.uk)

extract and combine reliable information from the RGB and infrared modalities. To solve this problem, there are two research issues which should be addressed. First, the tracked object usually encounters different appearance variations within each modality such as large illumination changes and occlusion as illustrated in Figures 1(a) and (d). These variations usually contaminates the tracking samples, and constructing (updating) the tracking model using such contaminated samples may degrade the tracking performance. Therefore, effectively extracting informative features of both RGB and infrared modalities from these potentially contaminated samples to deal with the sample contamination issue, and exploiting their complementary to handle appearance changes, is very important for robust RGB-infrared tracking. Second, due to the heterogeneity in multi-modality data, the visual properties of RGB and infrared images are intrinsically different (e.g. texture and color as shown in Figure 1). Such difference may result in a gap between the statistical properties of the features of these two modalities. Bridging the gap between the heterogeneous modalities to address cross-modality discrepancy issue, which means mining the consistency or correlation between these modalities, is also essential for effective modality fusion. While exploiting the modality consistency or correlation is important, it has been shown that incorporating modality specific characteristics can further improve the fusion performance because of modality complementarity [30]–[33].

Several algorithms have been developed to perform RGB-infrared tracking and show effectiveness in some scenarios. However, most of them do not explicitly consider handling either or both aforementioned issues, which may limit their modality fusion performance. One type of approaches exploits some fusion techniques such as score-level fusion (sum rule) [34], feature concatenation [35], tracking decision fusion [36] to combine information from RGB-infrared modalities. These methods (e.g. the feature concatenation method [35]) usually treat the RGB-infrared modalities as homogeneous feature channels, and ignore their heterogeneous characteristics. In addition, some of them regard these modalities independently (e.g. combining the independent tracking results in two separate modalities in [36]), which may not well exploit the correlation or consistency between different modalities for fusion. In general, such type of methods can not well address the cross-modality discrepancy issue. Another type of methods consider appearance modeling of RGB and infrared modalities as different learning tasks and perform tracking in the framework of multi-task learning such as multi-task joint sparse representation [37], [38]. Through multi-task learning, such kind of methods can exploit the correlation of different learning tasks for RGB-infrared tracking to some extent. However, these methods usually exploit some strong constraints on fusion such as the joint sparsity constraint which enforces all the representation patterns to be the same and the modality-specific properties are not exploited in appearance modeling.

To address the aforementioned issues, we develop a new feature learning model for RGB-infrared object tracking. The feature learning model aims to extract informative feature templates of multiple modalities from potentially contaminated tracking samples for appearance modeling. To address the cross-modality discrepancy issue, the learned feature templates of the heterogeneous modalities are constrained to achieve modality consistency in the following two aspects: 1) representation consistency, which means the representation should have necessary consistency to preserve the sharable modality invariant properties for enhancing the representation ability [30]; and 2) discriminability consistency, which implies that consistent discrimination decision should be reached by the discriminators of multiple modalities [39] and this will help to strengthen the discrimination power of the tracking model. In addition, through a new soft regularization scheme on modality consistency modeling, different from other existing methods such as [37], [38] which employ some strict constraints to model the consistency only, the proposed model can further exploit the specific properties in discriminability and representation ability for appearance modeling. Moreover, an optimization algorithm based on implicit differentiation on fixed-point equations is derived, which ensures that the representation ability and discriminability of the modality fusion model can be jointly optimized.

In summary, the contributions of this work are listed as follows:

- A RGB-infrared tracking system is developed for industrial applications.
- A learning model is proposed to extract informative feature templates and exploit the modality consistency in discriminability and representation ability for modality fusion based appearance modeling.
- An effective optimization algorithm is derived to learn the feature template learning model.

The rest of this paper is organized as follows. In Section II, we first review related works on RGB-Infrared object tracking and sparse representation-based visual tracking. In Section III, we present our proposed tracking models as well as the corresponding learning algorithm. We describe the implementation details in Section IV. Experimental analysis and conclusion are given in Sections V and VI, respectively.

II. Related Work

In this section, we first review related works in RGB-infrared tracking. Then some sparse representation-based trackers which are related to the methodologies exploited in our proposed tracking system are also introduced. For a more comprehensive summary of tracking methods, interested readers can refer to [40]–[43].

A. RGB-Infrared Tracking

Various algorithms have been designed for RGB-infrared object tracking. Bunyak *et al.* developed a level set based RGB-infrared moving object segmentation and tracking framework [44]. An algorithm for fusing the tracking results of multiple spatiogram trackers on RGB and infrared modalities was proposed in [36]. In [34], confidence maps from RGB and infrared modalities were aggregated for pedestrian tracking by using sum rule based on a probabilistic background model. To

enhance the tracking robustness, several sparse representation-based trackers are developed in which some fusion models such as feature concatenation [35], group sparsity [37], low rank regularization [45] were exploited for modality combination. However, these methods may fail to effectively and jointly utilize the modality consistency and specificity.

B. Sparse Representation-Based Tracking

Sparse representation has been widely applied in many computer vision tasks [46], such as image classification [47], object recognition [48]. Inspired by the success in sparse representation, Mei and Ling [49] proposed the ℓ_1 tracker which exploits sparse representation via ℓ_1 minimization for appearance modeling in tracking. The method in [49] shows some effectiveness in dealing with appearance variations, such as occlusion. Along this line, various tracking algorithms based on sparse representation have been proposed [40]. To more effectively handle the appearance variations especially the local deformation and partial occlusion, local sparse appearance model is developed such as the local patch dynamical graph learning [50], the structural sparse model [51]. Since optimizing the sparse representation/coding models is usually of high computational complexity, there are some developed algorithms which also consider to enhance the computational efficiency by reducing the feature dimension [52], constructing circular shift matrices [53]. To capture the intrinsic characteristics of the tracked object, several sparse representation-based feature learning methods are developed based on dictionary learning [54]–[56], subspace learning [20], [57], etc.. Since exploiting one single feature extracted from the RGB modality (e.g. gray intensity) may not be able to deal with complicated appearance changes, multiple sparse representation-based tracking algorithms which fuse multiple features from one single or multiple modalities have been developed such as the multiple sparse representation model [6], [58], [59], the collaborative representation model [38], collaborative discriminative learning [60]. Zhang *et al.* [61] developed the output constraint transfer model for kernelized multiple channel correlation filters. As mentioned, these methods focus on exploiting the consistency or correlation among multiple features (modalities), and the modality-specific characteristics are not well utilized, which may limit their fusion performance.

III. Proposed method

This section describes the novel aspects of the proposed tracking system: 1) modality-consistency sparse representation framework, which is a general framework for learning feature templates for sparse representation in RGB-infrared modalities under the modality consistency constraint; and 2) discriminability-consistency constrained feature template learning, which further imposes the constraint of the discriminability consistency for feature template learning. Then optimization algorithms for deriving the tracking model is presented.

a) Modality-Consistency Sparse Representation Framework: During the tracking process, the samples of the tracked target are collected by the tracker itself for learning or updating the tracking model. Let $Y_1^m = [y_1^m, \ldots, y_n^m] \in \mathbb{R}^{d^m \times N_1}$, $m = 1, \ldots, M$ denote the target samples of the M-th modality obtained by the tracker for model training, and N_1 is the number of the target samples in the training set, and M is the number of the modality in the training samples (M = 2 in our rgb-infrared tracking system). Since the tracked target would encounter some appearance variations during the tracking process, feature learning should be performed to capture the intrinsic properties of the tracked target for appearance modeling. Because of the limited training samples, large-scale off-line trained deep models may not be appropriate. Inspired by the success of sparse representation in computer vision [46], one objective of the proposed learning model is to exploit the correlation of different modalities and learn multi-modal feature templates for sparse representation of the tracked object. Let

$$Y_1^m = D_1^m X_1^m + E_1^m, m = 1, \ldots, M \tag{1}$$

where $D_1^m = [D_{1(\cdot,1)}^m, \ldots, D_{1(\cdot,c)}^m] \in \mathbb{R}^{d^m \times c}$ are the learned feature templates in the m-th modality, $X^m = [X_1^m, \ldots, X_c^m] \in \mathbb{R}^{c \times N_1}$ are the corresponding reconstruction coefficient vectors for the tracking samples, and $E^m \in \mathbb{R}^{d^m \times N_1}$ are the error matrix that may be caused by the variations of the tracked object. Because the reconstruction coefficients of different modalities can be regarded as the representation patterns of different modalities and they are closely related to the feature templates, to exploit the correlation or consistency of different modalities and perform effective fusion of different modalities during the learning process, the modality-consistency constraint should be incorporated into the feature template and sparse representation framework, which implies that the consistency constraint should be imposed on the reconstruction coefficient vectors X^m to guide the multi-modal feature template learning. Then the modality-consistency sparse representation framework can be formulated as

$$\begin{aligned} \min_{\{X_1^m, D_1^m, E_1^m\}} & \frac{1}{2} \sum_{m=1}^{M} \|E_1^m\|_F^2 + \lambda_1 \sum_{m=1}^{M} \|X_1^m\|_1 + \lambda_2 \mathcal{R}(\{X_1^m\}) \\ \text{s.t. } & Y_1^m = D^m X_1^m + E_1^m, m = 1, \ldots, M \\ & \|D_{1(\cdot,c)}^m\|_2 \leq 1, m = 1, \ldots, M, c = 1, \ldots, C \end{aligned} \tag{2}$$

where the first term $\sum_{m=1}^{M} \|\cdot\|_F^2$ aims to minimize the total reconstruction error using the learned feature templates of the multiple modalities, the second term $\sum_{m=1}^{M} \|\cdot\|_1$ is the sparsity regularization which can be utilized to select the informative templates in different modalities for the reconstruction of the tracked object, the third terms $\mathcal{R}(\{W^m\})$ is the regularizer which is incorporated to characterize the consistency of different modalities, $D_{1(\cdot,c)}^m$ denote the c-th column (template) of D_1^m and the inequality provides the unit ℓ_2 norm constraint .

To bridge the gap between the heterogeneous modalities for effective modality fusion and characterize the consistency over the representations of different modalities, the regularizer in (2) should be able to mine the similarity among different modalities. A straightforward approach to enforce the similarity between different modalities is simply to make all the representation pattern the same. However,this approach ignores the specificities which exist in different heterogeneous modalities, which means the complementary advantages of different modality representations cannot be well exploited.

Instead of using such kind of strict constraints, the regularizer in (2) can be defined as

$$\mathcal{R}(\{X_1^m\}) = \sum_{m=1}^{M}\sum_{n=1}^{N_1}\|x_n^m - x_n^0\|_2^2 \tag{3}$$

where x_n^0 is the consensus representation of different modalities for the n-th sample, which captures the consistency in the representations of different modalities. By minimizing the regularization term, the representation W^m will be constrained to be close to W^0, which makes the representations of different modalities $W^m, m = 1, \ldots, M$ similar to each other. The constraint on the representations of different modalities also guide the learning of the representation-consistency multi-modal feature templates.

b) Discriminability-Consistency Constrained Feature Template Learning: Although the modality-consistency feature template learning framework is able to learn feature templates which capture the consistency between different modalities, it cannot guarantee the strong discriminability of the learned templates for foreground and background separation, which means the appearance model with learned multi-modal feature templates may suffer the loss of robustness under the cluttered background. Therefore, discriminability regularization should be considered in the feature template learning model. Let $X^m = [X_1^m, X_2^m] = [x_1^m, \ldots, x_N^m] \in \mathbb{R}^{d^m \times N}$ be the training samples which consists of the $m-$th modality of the tracked target samples $X_1^m \in \mathbb{R}^{d^m \times N_1}$ and the background samples $X_2^m \in \mathbb{R}^{d^m \times N_2}$ where $N = N_1 + N_2$, $L = [L_1, \ldots, L_N]$ be the zero-one label matrix where $L_n = [1, 0]^T (L_n = [1, 0]^T)$ means the $n-$th samples is the foreground (background), D^m. To ensure the consistency in the discriminability level, inspired by the label-consistent dictionary learning [62], we incorporate the discriminability-consistency constraint into the sparsity-based multi-modal feature template learning framework:

$$\min_{\{W^m, D^m\}} \sum_{m=1}^{M} \ell(D^m, W^m; X^m, L, Q) + \frac{\lambda_3}{2}\sum_{m=1}^{M}\|W^m\|_F^2 \tag{4}$$

where $\ell(D^m, W^m; X^m, L, Q) =$

$$\frac{\alpha_1}{2}\sum_{m=1}^{M}\|W^m X^m - L\|_F^2 + \frac{\alpha_2}{2}\sum_{m=1}^{M}\|X^m - Q\|_F^2, \tag{5}$$

$\|W^m X^m - L\|_F^2$ measures the prediction loss using the linear classifier of modality M on the training samples, $\|X^m - Q\|_F^2$ measure the discrimination of the sparse representation, $Q = [Q_1, \ldots, Q_N] \in R^{C \times N}$ is the zero-one matrix, $Q_{n,k} = 1$ means the sample $x_n^m, m = 1, \ldots, M$ belongs to the same class with the k-th feature template, and $Q_{n,k} = 0$ means the sample $x_n^m, m = 1, \ldots, M$ belongs to the different class with the k-th feature template. We can see the objective function in (4) imposes the discriminability-consistency regularization from two perspectives. First, the first term introduces the constraint that the discrimination results (i.e. output of classifiers) of different modalities should be consistent. Second, the second term enforces that the target sample can be only well represented by the feature templates of the target class, which means the indexes of the non-zero elements of the sparse representation of the target samples and the background samples are different and the ones of different modalities of the same samples should be consistent. This constraint ensures that the feature patterns of the target samples and the background samples can be easily saperated and the feature patterns of different modalities of the same sample should be similar, which implicitly enhance the discriminability of the tracking model and impose the discriminability consistency regularizer.

c) Putting them all together: Based on the aforementioned analysis, the tracking model can be formulated into the following modality-consistent feature template framework:

$$\min_{\{W^m, D^m\}} \sum_{m=1}^{M} \frac{\alpha_1}{2}\|W^m X^m - L\|_F^2 + \frac{\alpha_2}{2}\|X^m - Q\|_F^2 + \frac{\lambda_3}{2}\|W^m\|_F^2 \tag{6}$$

$$\begin{aligned}
\text{s.t. } & \{X^m, X^0, E^m\} = \arg\min_{\{X^m, X^0, E^m\}} \mathcal{F}(X^m, X^0, E^m) \\
& Y^m = D^m X^m + E^m \\
& \|D^m_{(\cdot,c)}\|_2 \le 1 \\
& m = 1, \ldots, M, c = 1, \ldots, C
\end{aligned}$$

where

$$\mathcal{F}(X^m, X^0, E^m) = \begin{array}{l} \frac{1}{2}\sum_{m=1}^{M}\|E^m\|_F^2 + \lambda_1 \sum_{m=1}^{M}\|X^m\|_1 \\ \quad +\lambda_2 \sum_{m=1}^{M}\sum_{n=1}^{N}\|x_n^m - x_n^0\|_2^2 \end{array} \tag{7}$$

The learning framework performs the sparse representation and feature template learning under the constraint of modality consistency. The optimization algorithm for solving (6) will be presented in Section III(d).

d) Optimization: To reduce the number of the optimal variables for more efficient optimization, we transform the problem into

$$\min_{\{W^m, D^m\}} \sum_{m=1}^{M} \frac{\alpha_1}{2}\|W^m X^m - L\|_F^2 + \frac{\alpha_2}{2}\|X^m - Q\|_F^2 + \frac{\lambda_3}{2}\|W^m\|_F^2 \tag{8}$$

$$\begin{aligned}
\text{s.t. } & \{X^m\} = \arg\min_{\{X^m\}} \begin{array}{l} \frac{1}{2}\sum_{m=1}^{M}\|Y^m - D^m X^m\|_F^2 + \lambda_1 \sum_{m=1}^{M}\|X^m\|_1 \\ \quad +\frac{\lambda_2}{2}\sum_{m=1}^{M}\|X^m - \frac{1}{M}\sum_{k=1}^{M} X^k\|_2^2 \end{array} \\
& \|D^m_{(\cdot,c)}\|_2 \le 1, m = 1, \ldots, M, c = 1, \ldots, C
\end{aligned}$$

where the solution $X^0 = \frac{1}{M}\sum_{k=1}^{M} X^k$ can be obtained by taking the derivative of (6) with respect to X^0 and setting it to be zero. Since all the tracking samples cannot be obtained at the same time, we adopt stochastic gradient decent to optimize (8). Let $y^m \in \mathbb{R}^{d^m}, m = 1, \ldots, M$ denote a sample of multi-modalities for gradient computation, $x^m \in \mathbb{R}^C, m = 1, \ldots, M$ be the sparse coefficients, l is the label vector of the sample, and q is the target sparse coefficients. The main difficulty in the gradient computation is how to obtain the gradients of x^m with respect to D because D is not explicitly defined in the discriminability regularization (4) but it is related to x^m in the sparse representation model in (8). To address this issue, implicit differentiation with the chain rule is exploited to derive the gradient indirectly. For the simplicity

of derivation and presentation, we define new variables Y', D', X', W', and L', where $Y' = [(y^1)^T, \ldots, (y^M)^T]^T$, $X' = [(x^1)^T, \ldots, (x^M)^T]^T$, $L' = [l^T, \ldots, l^T]^T \in \mathbb{R}^{2\cdot M}$, $Q' = [q^T, \ldots, q^T]^T$ and

$$D' = \begin{bmatrix} D^1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & D^M \end{bmatrix}, W' = \begin{bmatrix} W^1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & W^M \end{bmatrix}.$$

Then by taking the gradient of the objective function of the sparse representation step shown in (8) and setting it to be zero, the following equation can be obtained.

$$(D')^T \left(D'X' - Y'\right) + AX' = -\text{sign}(X') \tag{9}$$

where

$$A = \begin{bmatrix} \frac{M-1}{M}\mathcal{I}_C & -\frac{1}{M}\mathcal{I}_C & \ldots & -\frac{1}{M}\mathcal{I}_C \\ -\frac{1}{M}\mathcal{I}_C & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ -\frac{1}{M}\mathcal{I}_C & \ldots & \ldots & \frac{M-1}{M}\mathcal{I}_C \end{bmatrix} \in \mathbb{R}^{(C\cdot M)\times(C\cdot M)}$$

and $\mathcal{I}_C \in \mathbb{R}^{C\times C}$ is an identity matrix. By applying the strategy of implicit differentiation on (16), the differential of X' with respect to $D^m_{(i,j)}$ can be derived as $\frac{\partial X'}{\partial D^m_{(i,j)}} =$

$$\left[(D')^T(D') + A\right]^{-1}\left[(E^m_{(i,j)})^T(Y' - D'X') - (D')^T E^m_{(i,j)}X'\right] \tag{10}$$

where $E^m_{(i,j)} \in \mathbb{R}^{\sum_{k=1}^{M} d^k + C\cdot M}$ is a matrix with zero elements except the element in the $(\sum_{k=1}^{m-1} d^k + i)$-th row and the $[(m-1)C+j]$-th column is 1. Then the derivative of $\ell(\cdot)$ with respect to $D^m_{(i,j)}$ can be computed by using the chain rule

$$\frac{\partial l}{\partial D^m_{(i,j)}} = \frac{\partial l}{\partial X'}\frac{\partial X'}{\partial D^m_{(i,j)}} \tag{11}$$

where $\frac{\partial l}{\partial X'} = \begin{pmatrix} \alpha_1[(X')^T(W')^T(W') - (L')^T W'] \\ +\alpha_2[(X')^T - (Q')^T] \end{pmatrix}$. The derivative of l with respect to W^m is

$$\frac{\partial l}{\partial W^m} = \alpha_1((x^m)(x^m)^T(W^m)^T - (x^m)l^T) + \lambda_3 W^m \tag{12}$$

Based on these derivatives based on one training sample, stochastic gradient can be exploited to update the optimal variables iteratively. We adopt the learning rate updating scheme in [63] which sets it to be $\min(\tau, \tau T/(10t))$ where t means the gradient decent is performed in the t-th iteration and T is the total number of iterations. As noted in [63], the convergence of the stochastic gradient decent algorithm to the stationary point can be achieved under some assumptions, and the learning rate should be well tuned to achieve a better performance. Under our setting, the learning rate will keep fixed in the first $\frac{T}{10}$ iterations, and then descrease in the $\frac{1}{t}$ annealing strategy, which avoids the case where the learning rate descrease too quickly. In each iteration, we permute the training samples and compute the gradient based on each sample to update the parameters iteratively which are shown as follows:

$\{x^m\}$**-sparse learning subproblem**: With fixed $\{D^m\}$, given the randomly drawn training sample of multiple modalities $y^m, m = 1, \ldots, M$, their sparse presentations of multiple modalities can be obtained by solving the following sparse learning problem:

$$\min_{\{x^m\}} \begin{array}{c} \frac{1}{2}\sum_{m=1}^{M}\|y^m - D^m x^m\|_F^2 + \lambda_1 \sum_{m=1}^{M}\|x^m\|_1 \\ +\frac{\lambda_2}{2}\sum_{m=1}^{M}\|x^m - \frac{1}{M}\sum_{k=1}^{M} x^k\|_2^2 \end{array} \tag{13}$$

The objective function consists of smooth components (quadratic terms) and non-smooth one (ℓ_1 norm regularization). The Accelerated Proximal Gradient Method [64] can be adopted to solve the problem.

$\{D^m, W^m\}$**-subproblem**: Before performing gradient deceent, the learning rate τ can be updated by $\tau^t = \min(\tau^{t-1}, \tau^{t-1}T/(10t))$. Then with fixed $\{x^m\}$, $\{D^m\}$ and $\{x^m\}$ can be updated as follows:

$$\begin{aligned} D^m &= D^m - \tau^t(\frac{\partial l}{\partial D^m})^T \\ D^m_c &= \frac{D^m_c}{\|D^m_c\|_2}, \text{for } c = 1, \ldots, C \\ W^m &= W^m - \tau^t(\frac{\partial l}{\partial W^m})^T \end{aligned} \tag{14}$$

where $\frac{\partial l}{\partial D^m} = (\frac{\partial l}{\partial D^m_{(i,j)}})_{d^m\times C}$.

IV. Implementation Details

This section mainly introduces the key implementation details of the proposed multi-modal tracker.

A. Initialization

In the initial frame, we randomly sample N_0 target and background samples of RGB-infrared modalities and then extract the feature descriptors of each sample, which are denoted as Y^m. For each modality, we adopt K-SVD [65] to initialize the feature templates for target samples and background samples, which are denoted as D^m_F and D^m_B respectively. With the intialed feature temlpates $D^m_0 = [D^m_F, D^m_B]$, the sparse coefficiets X^m can be estimated. Then the classifiers for each modality can be estimated as $W^m = \arg\min_W \|WX^m - L\|_F^2 + \lambda_3\|W\|_F^2$. The tracker randomly samples 200 examples which is close to the ground truth in the first frame as positive samples, and 200 examples which do not overlap with the ground truth bounding box as negative samples. For using the K-SVD algorithm in initializing feature templates, following the setting in [54], the number of templates in the target dictionary and the background dictionary are both set to 200. In each iteration, the L1-regularized least square problem is solved to obtain the sparse codes in which the parameter associated with the L1-regularization is set to 0.5. There are totally 5 iterations used in the K-SVD algorithm.

B. Particle Filtering for Target Position Decision

The tracking algorithm is implemented in the particle filtering framework. Within this framework, we estiminate the tracking results at Frame t by maximizing a posteriori:

$$\tilde{h}_t = \arg\max_{h_t^i} p(h_t^i|Z_t) \tag{15}$$

where $Z_t = \{z_j | j = 1, \ldots, t\}$ denote the set of observation variables up to Frame t, z_j is the observation variable at Frame

j, and h_t^i is the state variable of the i-th particle at Frame t. The objective is that we want to approximate the true posterior by a set of particles with different states h_t^i, and the posterior probability $p(h_t^i|Z_t)$ is recursively computed as

$$p(h_t|Z_T) \propto p(z_t|h_t) \int p(h_t|h_{t-1})p(h_{t-1}|Z_{t-1})dh_{t-1} \quad (16)$$

where $p(h_t|h_{t-1})$ and $p(z_t|h_t)$ denote the motion model and the observation model. The motion state is defined as $h_t = [b_1, b_2, b_3, b_4, b_5, b_6]$ which encodes the horizontal and vertical translation, scale, rotation angle, skew and aspect ratio respectively, and the motion model is define as $p(h_t|h_{t-1}) = N(h_{t-1}, \Gamma)$ where Γ is a 6-by-6 diagonal matrix. With the learned feature templates, given a target candidate sample, the sparse coefficients can be obtained by solving (2). Then based on the learned feature templates and the classifiers, the observation likelihood function can be defined as

$$p(o_t|s_t) \propto \exp\left(-\sum_{m=1}^{M}(\|z_{tr}^m - D^m x^m\|_2^2 + \rho\|[1,0]^T - W^m x^m\|_2^2)\right) \quad (17)$$

where z_{tr}^m is the weighted average feature of the samples of the m-th modality in the sample pool, x^m is the sparse coefficient of the sample of the m-th modality, h is the label vector which represents the target label. Since a good target candidate should be well represented by the learned feature templates and have high confidence to be the positive class $[1,0]^T$, we use the joint decision measurement based on the reconstruction error using the feature templates and the classification accuracy of multiple modalities to decide the target position. Such joint decision measurement exploits the representation ability and the discriminability of the learned feature templates.

To obtain the weighted average feature of the samples of the m-th modality, the proposed tracker maintains a sample pool which contains a fixed number P of the features of tracked samples in RGB-infrared modalities, which are denoted as $\{z_p^k\}, k = 1, \ldots, M, p = 1, \ldots, P$. When the features of the most recent tracking results in multiple modalities are added into the sample pool, the samples captured from the older frames will be removed. This strategy will maintain the adaptivity of the tracking model. In the sample pool, each sample is associated with a weight $\omega_p, p = 1, \ldots, P$ which is set to be the value of the observation likelihood in (17). Before computing the weighted average feature z_{tr}^m, the sample weight is normalized as $\omega_p' = \frac{\omega_p}{\sum_{p=1}^{P}\omega_p}$.

V. Experiments

This section first describes the experimental setting, and then provides the analysis of the experimental results.

A. Experimental Setting

Fifteen RGB-infrared video pairs which are captured by visible and thermal sensors are adopted for experimental evaluation. These videos cover some challenging scenarios, such as occlusion, poor lighting condition, and large scale variations. To perform more effective modality fusion, these videos are well aligned and registered, which means the track target in the RGB and infrared modalities almost share the same location in each frame.Ten methods are used for comparison, which includes STRUCK [66], STC [67], CT [68], MIL [69], RPT [70], MEEM [71], KCF [72], CN [73], ℓ_1 [35], and JSR [37] method. Among these method, expect the ℓ_1 and the JSR methods which are proposed for RGB-infrared object tracking, all the other methods are originally designed for RGB single modality tracking. Following the setting in [38], we exploit the multi-modality version of these methods by concatenating the features of the RGB and infrared modalities as the input of these trackers. Some results of these methods can be obtained from [38]. The parameters λ_1, λ_2, λ_3, α_1, and α_2 are set to 0.5, 10^{-6}, 10^{-6}, 0.4, 0.6, and the learning rate is initialized as 0.2. In each frame, given the image patches of the tracking results or the training samples, we extract the HOG features from both RGB and infrared modalities [74] as representation.

B. Experimental Results

Two metrics are adopted to quantitatively measure the tracking accuracy. They are overlapping rate and success rate. The overlapping rate is defined as $\frac{area(S_1 \bigcap S_2)}{area(S_1 \bigcup S_2)}$ where S_1 and S_2 are the bounding box of the tracker and the groundtruth. The tracking in each frame is considered to be performed successfully if the overlapping rate is greater than 0.5. Then the success rate is defined as the percentage of video frames in which a tracking success is achieved. Tables I and II show the overlapping rate and the success rate of all the compared trackers in the fifteen videos. We can see that the proposed tracker ranks in top two on fourteen videos in terms of success rate and overlapping rate. Compared with other trackers, our proposed tracker is more robust under some challenging scenarios such as occlusion (e.g. *OccCar-1*), thermal crossover (e.g. *Cycling*), scale changes (e.g. *BusScale*) which are illustrated in Fig.2. Specially, compared with other multi-modal trackers such as the JSR method and the l_1 tracker which also exploit the consistency in the multi-modal representation, our proposed method further exploit the consistency in discriminability level, which facilitate the discrimination between the tracked object and the background. Similar to these dictionary learning-based tracking algorithms [75]–[77] which shows learning features under sparsity constraint can be more effective to deal with appearance variations such as occlusion, the proposed algorithm can learn some feature templates under the sparsity constraint which are not contaminated by some external variations such as partial occlusion. When the tracker encounters the variation of thermal crossover, the fusion of RGB and infrared information make the tracker less sensitive to such kind of variation because only the infrared modality is affected by thermal crossover but the RGB information are still reliable. On the contrary, when there is large illumination changes or it is in low illumination conditions which make the RGB modality not reliable, fusing the infrared modality will enhance the robustness of the proposed tracker to different variations, and thus will perform much better than the one only use RGB information.

Fig. 3 show the frame-by-frame quantitative comparison in terms of overlapping rate. We can see that the proposed

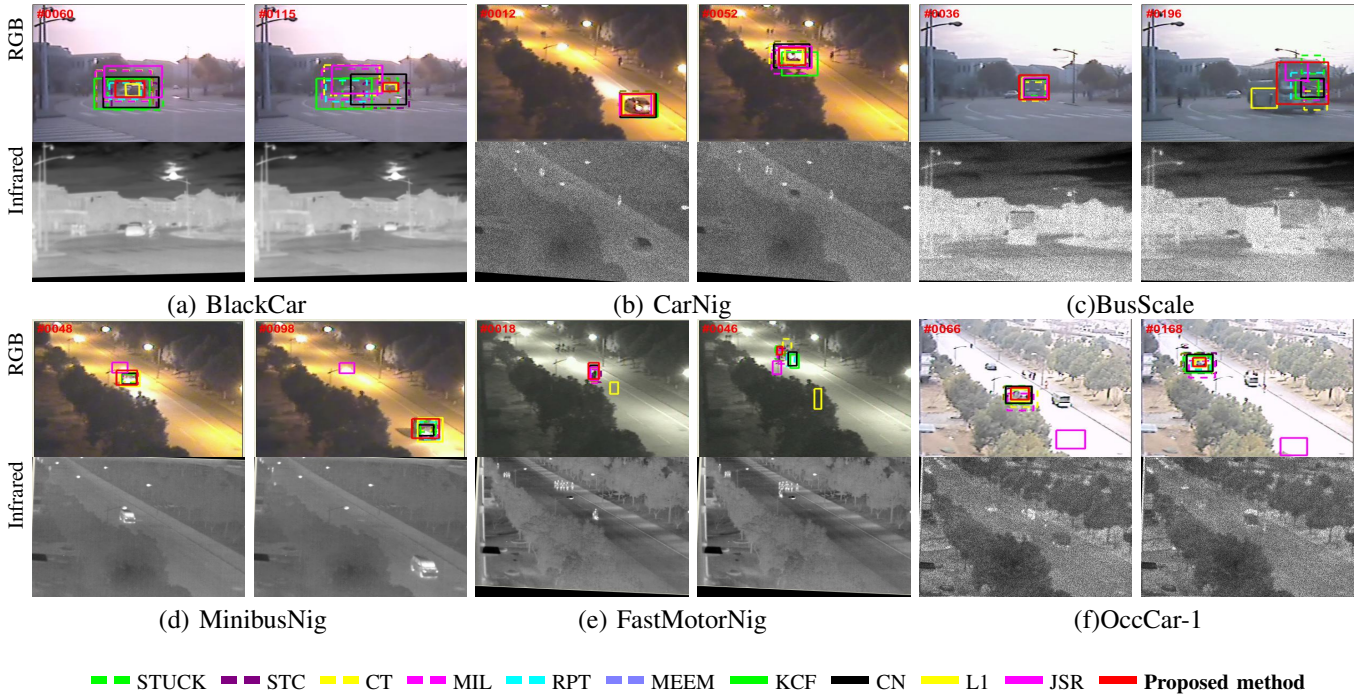

Fig. 2. Qualitative results on some frames of video in RGB and infrared modality with some challenging factors, such as occlusion (e.g. *OccCar-1*),Thermal crossover (e.g. *Cycling*), scale changes (e.g. *BusScale*), low illumination (e.g. *CarNig*, *FastMotorNig*). For each sub-figure, images of RGB modality are shown in the top row while images of infrared modality are shown in the bottom row.

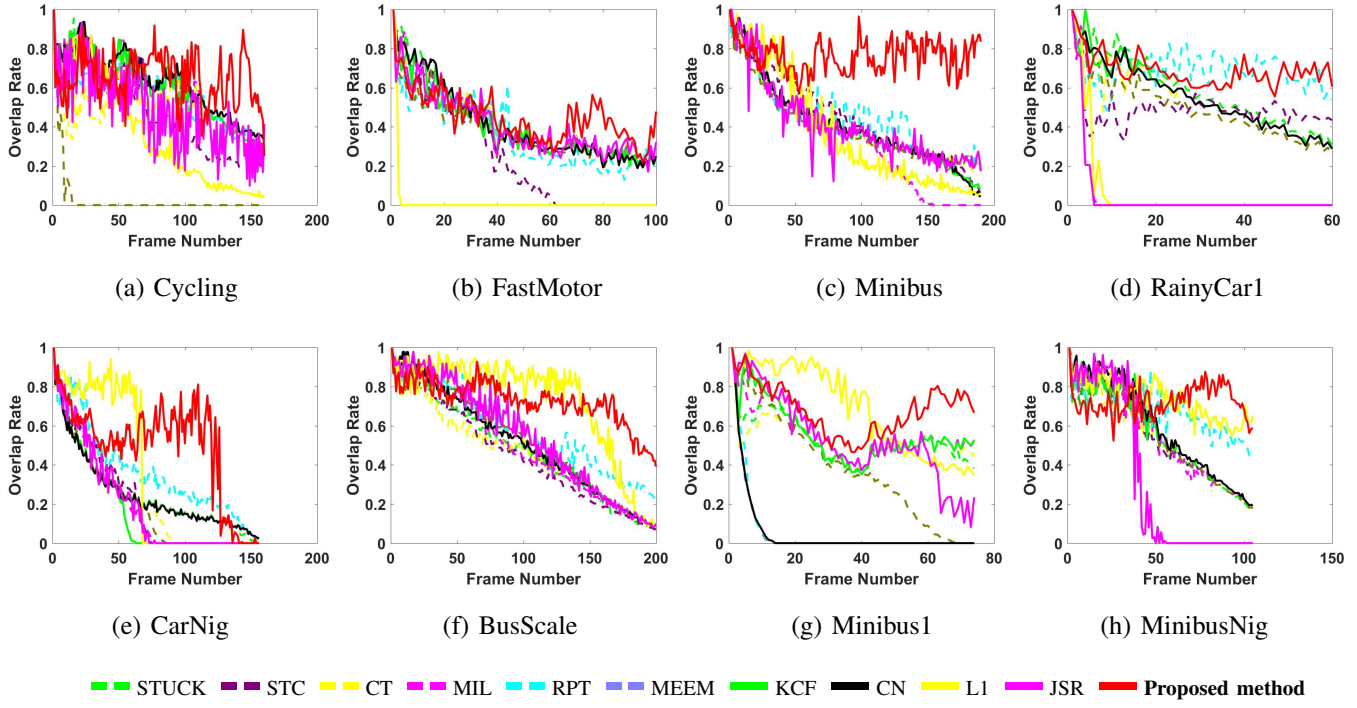

Fig. 3. Quantitative comparison of 11 trackers on 8 challenging videos in terms of overlapping rate. The horizontal axis is the frame index and the vertical axis indicates the overlapping rate.

TABLE I
OVERLAPPING RATE. THE BEST THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN.

	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	L1	JSR	**Proposed Method**
BlackCar	0.24	0.31	0.21	0.22	0.33	0.23	0.21	0.24	0.64	0.23	0.52
BlueCar	0.37	0.27	0.34	0.4	0.65	0.47	0.4	0.4	0.63	0.4	0.63
BusScale	0.47	0.45	0.46	0.49	0.57	0.52	0.51	0.51	0.72	0.54	0.73
Exposure2	0.32	0.37	0.31	0.32	0.48	0.3	0.32	0.32	0.82	0.35	0.58
MinibusNig	0.54	0.55	0.54	0.55	0.68	0.55	0.57	0.59	0.74	0.33	0.73
FastMotorNig	0.51	0.54	0.41	0.55	0.43	0.57	0.35	0.37	0.08	0.48	0.6
Motorbike	0.31	0.31	0.31	0.31	0.31	0.3	0.31	0.31	0.5	0.3	0.53
CarNig	0.25	0.21	0.2	0.18	0.36	0.19	0.16	0.25	0.35	0.2	0.5
Cycling	0.62	0.47	0.51	0.64	0.55	0.03	0.61	0.63	0.36	0.49	0.65
FastMotor	0.43	0.24	0.43	0.42	0.36	0.37	0.4	0.41	0.02	0.41	0.45
Minibus	0.43	0.46	0.42	0.34	0.43	0.39	0.41	0.41	0.37	0.42	0.76
OccCar-1	0.45	0.46	0.43	0.33	0.68	0.41	0.45	0.45	0.82	0.07	0.69
RainyCar1	0.58	0.5	0.55	0.07	0.69	0.49	0.55	0.55	0.07	0.05	0.69
GoTogether	0.79	0.63	0.09	0.75	0.77	0.71	0.66	0.71	0.65	0.49	0.74
Minibus1	0.53	0.05	0.52	0.55	0.06	0.38	0.56	0.05	0.69	0.53	0.68
Average	0.46	0.39	0.38	0.41	0.49	0.39	0.43	0.41	0.5	0.35	0.63

TABLE II
SUCCESS RATE. THE BEST THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN.

	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	L1	JSR	**Proposed Method**
BlackCar	0.12	0.16	0.1	0.12	0.29	0.12	0.12	0.12	0.83	0.15	0.57
BlueCar	0.33	0.33	0.28	0.38	0.94	0.46	0.38	0.38	0.68	0.44	0.82
BusScale	0.48	0.4	0.46	0.44	0.61	0.53	0.5	0.51	0.82	0.56	0.93
Exposure2	0.2	0.26	0.2	0.2	0.45	0.16	0.2	0.2	1	0.19	0.66
MinibusNig	0.51	0.49	0.55	0.51	0.92	0.51	0.54	0.55	1	0.36	1
FastMotorNig	0.37	0.72	0.52	0.55	0.54	0.55	0.48	0.51	0.09	0.45	0.75
Motorbike	0.14	0.16	0.14	0.13	0.13	0.12	0.14	0.14	0.48	0.12	0.5
CarNig	0.13	0.19	0.13	0.13	0.21	0.13	0.13	0.13	0.43	0.17	0.67
Cycling	0.71	0.43	0.53	0.71	0.68	0.02	0.71	0.71	0.33	0.48	0.86
FastMotor	0.32	0.19	0.33	0.33	0.3	0.2	0.27	0.3	0.02	0.27	0.35
Minibus	0.27	0.42	0.27	0.24	0.25	0.21	0.27	0.27	0.32	0.24	1
OccCar-1	0.32	0.44	0.27	0.24	0.89	0.21	0.32	0.32	1	0.08	0.94
RainyCar1	0.58	0.35	0.55	0.08	0.98	0.45	0.57	0.57	0.07	0.05	1
GoTogether	1	0.87	0.01	1	1	0.98	0.93	1	0.81	0.61	0.99
Minibus1	0.59	0.04	0.54	0.58	0.05	0.32	0.54	0.04	0.69	0.49	0.92
Average	0.4	0.36	0.33	0.38	0.55	0.33	0.41	0.38	0.57	0.31	0.8

trackers run stably in most of the videos which shows the stability of the proposed tracker.

Running time: Because the tracking model is optimized iteratively, the proposed tracker cannot run in real time and its FPS is 0.7.

Potential failure cases: Once the testing videos in low frame rate and dramatic scale changes happen, the proposed tracker may not work well since particles with limited scale variance may not be able to cover such kind of scale changes. As shown in the videos *BlackCar* and *Exposure2* which contain large scale changes, the proposed tracker does not perform well. In addition, since the proposed tracker is not explicitly designed to handle occlusion, it may not work well when full occlusion happens.

VI. CONCLUSION

In this paper, we proposed a robust RGB-Infrared tracking system for industrial applications. A modality-consistency Feature template learning algorithm is proposed to learn multi-modal feature template for appearance modeling. The learning algorithm achieve the modality consistency in the learned feature template in two levels: representation and discriminability. The experimental results on fifteen videos show its effectiveness. Since the proposed tracking algorithm cannot perform in real time, one of our future work will develop more efficient learning algorithms to improve the computational complexity. In addition, how to dynamically weight the importance of different modalities for modality fusion is another issue which will be further studied.

REFERENCES

[1] 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.

[2] 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.

[3] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, 2014.

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

[5] J. Han, E. J. Pauwels, P. M. de Zeeuw, and P. H. N. de With, "Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment," *IEEE Trans. Consumer Electronics*, vol. 58, no. 2, pp. 255–263, 2012.

[6] C. Chen, J. Xin, Y. Wang, L. Chen, and M. K. Ng, "A semisupervised classification approach for multidomain networks with domain selection," *IEEE Trans. Neural Netw. Learning Syst.*, 2018, DOI:10.1109/TNNLS.2018.2837166.

[7] 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.

[8] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Trans. Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018.

[9] 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.

[10] C. Gong, K. Fu, A. Loza, Q. Wu, J. Liu, and J. Yang, "Pagerank tracker: From ranking to tracking," *IEEE Trans. Cybernetics*, vol. 44, no. 6, pp. 882–893, 2014.

[11] W. Zhang and W.-K. Cham, "Gradient-directed multiexposure composition," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2318–2323, 2012.

[12] 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.

[13] 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.*, DOI: 10.1109/TNNLS.2018.2827036.

[14] 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.

[15] J. T. Zhou, I. W. Tsang, S.-s. Ho, and K.-R. Müller, "N-ary decomposition for multi-class classification," *Machine Learning*, 2018.

[16] 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.

[17] 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.

[18] 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.

[19] G. Ding, W. Chen, S. Zhao, J. Han, and Q. Liu, "Real-time scalable visual tracking via quadrangle kernelized correlation filters," *IEEE Trans. Intelligent Transportation Systems*, vol. 19, no. 1, pp. 140–150, 2018.

[20] 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*, 2018, DOI:10.1109/TIFS.2018.2868230.

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

[22] S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," in *Proc. ECCV*, pp. 558–573, 2018.

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

[24] B. Yang, A. J. Ma, and P. C. Yuen, "Body parts synthesis for cross-quality pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, DOI:10.1109/TCSVT.2017.2789224.

[25] B. Yang, A. J. Ma, and P. C. Yuen, "Learning domain-shared group-sparse representation for unsupervised domain adaptation," *Pattern Recognit.*, vol. 81, pp. 615–632, 2018.

[26] R. Shao, X. Lan, and P. C. Yuen, "Feature constrained by pixel: Hierarchical adversarial deep domain adaptation," in *ACM MM'18*, pp. 220–228, 2018.

[27] 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.

[28] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, and Q. Ni, "Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning," *IEEE Trans. on Industrial Electronics*, 2018, DOI:10.1109/TIE.2018.2873547.

[29] 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 Trans. on Image Process.*, vol. 28, no. 4, pp. 1993–2007, Apr. 2019.

[30] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition," in *Proc. ICCV*, pp. 1125–1133, Dec. 2015.

[31] J. Hu, J. Lu, and Y.-P. Tan, "Sharable and individual multi-view metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, DOI: 10.1109/TPAMI.2017.2749576.

[32] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1233–1246, 2015.

[33] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proc. CVPR*, pp. 5344–5352, Jun. 2015.

[34] 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.

[35] 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,*, pp. 1–8, 2011.

[36] 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.

[37] 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.

[38] 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.

[39] A. Shrivastava, M. Rastegari, S. Shekhar, R. Chellappa, and L. S. Davis, "Class consistent multi-modal fusion with binary features," in *Proc. CVPR*, pp. 2282–2291, 2015.

[40] 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.

[41] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: an experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, 2014.

[42] 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.

[43] S. Salti, A. Cavallaro, and L. di Stefano, "Adaptive appearance modeling for video tracking: Survey and evaluation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4334–4348, 2012.

[44] 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.

[45] 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.

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

[47] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, 2012.

[48] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Multi-observation visual recognition via joint dynamic sparse representation," in *Proc. ICCV*, pp. 595–602, 2011.

[49] X. Mei and H. Ling, "Robust visual tracking using l1 minimization," in *ICCV*, 2009.

[50] C. Li, L. Lin, W. Zuo, J. Tang, and M.-H. Yang, "Visual tracking via dynamic graph learning." *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, DOI:10.1109/TPAMI.2018.2864965.

[51] T. Zhang, C. Xu, and M.-H. Yang, "Robust structural sparse tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, DOI:10.1109/TPAMI.2018.2797082.

[52] S. Zhang, H. Zhou, F. Jiang, and X. Li, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 11, pp. 1749–1760, 2015.

[53] T. Zhang, A. Bibi, and B. Ghanem, "In defense of sparse tracking: Circulant sparse tracker," in *Proc. CVPR*, pp. 3880–3888, 2016.

[54] F. Yang, Z. Jiang, and L. S. Davis, "Online discriminative dictionary learning for visual tracking," in *Proc. WACV*, pp. 854–861. IEEE, 2014.

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

[56] 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.

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

[58] 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.

[59] 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*, pp. 1194–1201, 2014.

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

[61] B. Zhang, Z. Li, X. Cao, Q. Ye, C. Chen, L. Shen, A. Perina, and R. Ji, "Output constraint transfer for kernelized correlation filter in tracking," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 693–703, 2017.

[62] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, 2013.

[63] J. Mairal, F. R. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, 2012.

[64] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
[65] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
[66] 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.
[67] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. ECCV*, pp. 127–141, 2014.
[68] 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.
[69] 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.
[70] Y. Li, J. Zhu, and S. C. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *Proc. CVPR*, pp. 353–361, 2017.
[71] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in *Proc. ECCV*, pp. 188–203, 2014.
[72] 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.
[73] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. CVPR*, pp. 1090–1097. IEEE, 2014.
[74] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, pp. 886–893, 2005.
[75] N. Wang, J. Wang, and D.-Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," in *Proc. ICCV*, pp. 657–664, 2013.
[76] J. Xing, J. Gao, B. Li, W. Hu, and S. Yan, "Robust object tracking with online multi-lifespan dictionary learning," in *Proc. ICCV*, pp. 665–672, 2013.
[77] T. Zhou, F. Liu, H. Bhaskar, and J. Yang, "Robust visual tracking via online discriminative and low-rank dictionary learning," *IEEE Trans. Cybernetics*, vol. 48, no. 9, pp. 2643–2655, 2018.

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 a professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. His current research interest is computer vision.

Pong C. Yuen received his B.Sc. degree in Electronic Engineering with First Class Honours in 1989 from City University of Hong Kong, and his Ph.D. degree in Electrical and Electronic Engineering in 1993 from The University of Hong Kong. He joined the Hong Kong Baptist University in 1993, and served as the Head of Department of Computer Science from 2011-2017. Currently he is a Professor at the Department of Computer Science and Associate Dean of Science Faculty, Hong Kong Baptist Unviersity.

Dr. Yuen served as Associate Editor of IEEE Transactions on Information Forensics and Security from 2014 - 2018, and received the Outstanding Editorial Board Service Award in 2018. Currently, he is the Vice President (Technical Activities) of the IEEE Biometrics Council, Editorial Board Member of Pattern Recognition, and Senior Editor of SPIE Journal of Electronic Imaging. He also serves as a member of Hong Kong Research Grant Council Engineering Panel. He has been serving the Director of IAPR/IEEE Winter School on Biometrics since 2017. He is a Fellow of IAPR.

Dr. Yuen's current research interests include video surveillance, human face recognition, biometric security and privacy.

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. He 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.