Robust Multi-modality Anchor Graph-based Label Prediction for RGB-Infrared Tracking

Xiangyuan Lan, Wei Zhang, *Member, IEEE*, Shengping Zhang, *Member, IEEE*, Deepak Kumar Jain, and Huiyu Zhou

***Abstract*—Given massive video data generated from different applications such as security monitoring and traffic management, to save cost and human labor, developing an industrial intelligent video analytic system, which can automatically extract and analyze the meaningful content of videos, is essential. For achieving the objective of motion perception in video analytic system, a key problem is how to perform effective tracking of object of interest so that the location and the status of the tracked object can be inferred accurately. To solve this problem, with the popularity of RGB-infrared dual camera systems, this paper proposes a new RGB-infrared tracking framework which aims to exploit information from both RGB and infrared modalities to enhance the tracking robustness. In particular, within the tracking framework, a robust multi-modality anchor graph-based label prediction model is developed, which is able to 1) construct a scalable graph representation of the relationship of the samples based on local anchor approximation; 2) defuse a limited amount of known labels to large amount of unlabeled sample efficiently based on transductive learning strategy; and 3) adaptive incorporate importance weights for measuring modality discriminability. Efficient optimization algorithms are derived to solve the prediction model. Experimental results on various multi-modality videos demonstrate the effectiveness of the proposed method.**

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

I. Introduction

THE last decade has witnessed a substantially great demand of the industrial intelligent video systems, which is driven by many applications such as video surveillance, traffic management, robotics, face recognition and so on [1]–[9]. With the help of intelligent video systems, extracting and analyzing the meaningful content of videos can be performed automatically, which can save cost and human labor. A key task of such kinds of video systems is to achieve the intelligent motion perception of objects of interests. To this end, developing a robust object tracking model which is able to locate the objects of interest and infer their motion status is very important. However, it is still challenging to perform robust object tracking due to many unpredictable variations and poor environmental conditions, such as occlusion, poor illumination conditions, scale changes and so on. Extensive studies on object tracking have been made in tracking research in the past decade [10], [11], and lots of tracking algorithms have been developed to deal with various kinds of research issues, such as the issues of model drift [12], feature selection and fusion [13], etc.. However, most of these tracking algorithms are designed based on RGB modality only. They construct the appearance model by using the visual features from the RGB video frames. In some extreme but common cases such as dim environment at night, the RGB information is not reliable and tracking failure may happen. Therefore, deploying an object tracker with RGB modality only in intelligent video systems may limit their real-world application (e.g. night time video surveillance systems).

With the great advancement of multi-spectral imaging technology, forming multi-spectral images or videos has become more and more effective. Besides, the increasingly lower cost and higher quality of multispectral imaging devices has brought the wide application of multi-spectral camera systems. RGB-Infrared (thermal) camera systems, which are able to capture images or videos of both RGB and infrared modali-



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

Fig. 1. Some video frames of RGB and infrared modalities which covers challenging factors for RGB-infrared tracking. **Top**: RGB **Bottom**: infrared

This work was supported in part by Hong Kong Baptist University Tier 1 Start-up Grant. The work of S. Zhang was supported in part by the National Natural Science Foundation of China under Grant 61872112. The 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. The work of H. Zhou was supported in part by the U.K. EPSRC under Grant EP/N011074/1, Royal Society-Newton Advanced Fellowship under Grant NA160342, and European Union's Horizon 2020 Research and Innovation Program through the Marie-Sklodowska-Curie under Grant 720325. *(Corresponding Author: Wei Zhang)*

X. Lan is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China (e-mail: xiangyuanlan@life.hkbu.edu.hk)

W. Zhang is with the School of Control Science and Engineering, and Institute of Brain and Brain-Inspired Science, Shandong University, China. (e-mail: davidzhang@sdu.edu.cn)

S. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264200, China (e-mail: s.zhang@hit.edu.cn)

D. K. Jain is with the Key Laboratory of Intelligent Air-Ground Cooperative Control for Universities in Chongqing, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: deepak@cqupt.edu.cn)

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

ties, have been widely deployed in many industrial systems. Compared with visible spectrum cameras which forms images by visible lights, the infrared cameras produce images by receiving the infrared radiation of a subject with a above-zero temperature, and thus it is much less sensitive to large illumination changes or poor lighting conditions. However, the information from infrared cameras are not always reliable. For example, when the tracked object is surrounded by background with similar temperature which may cause the issue of thermal crossover, the tracked target may not be distinguishable from the background in infrared image. However, the visual features extracted from RGB images may be more able to differentiate the target from background since it carries more visual characteristic such as color, texture, etc., which is beneficial for appearance modeling. Therefore, developing an effective model for integrating both RGB and infrared modalities for object tracking is essential for intelligent video systems.

However, it is challenging to perform an effective integration of both RGB and infrared modalities for tracking because several issues would limit the modality combination performance. First, as shown in Figure 1, large variations such as occlusion, illumination variation may be encountered during the tracking process, which would introduce outliers and contaminate the tracking examples. Modeling learning (updating) with these examples may affect the tracker performance negatively. Therefore, suppressing and removing the contaminated features from the tracking samples is required. Second, due to the dynamical changes of the background and the appearance variation, the reliability of different modalities would be different and keep changed during the tracking process. For example, because of the issue of thermal cross over issue, the RGB modality is more reliable than infrared modality as shown in Figure 1(a). However, for the frame in Figure 1(c), the poor illumination condition make information from RGB modality less reliable than that in infrared modality. As such, to ensure that more reliable modality play more important roles in multi-modality appearance model while the negative effect of the unreliable modality can be suppressed, how to dynamically evaluate the reliability and incorporate the importance weight for modality fusion should be considered. Besides, only limited labeled tracking samples of different modalities are available during the tracking process, which makes it difficult to online train an power parametric discriminative fusion model for label inference (i.e. target or background) of testing samples (i.e. target candidates) with large appearance variations. Therefore, how to utilize limited tracking samples to infer the labels of target candidates for target position determination is another important issue to address.

To overcome the aforementioned issues, we propose a new discriminative learning model for RGB-infrared tracking. Specifically, we formulates RGB-infrared tracking as transductive semi-supervised graph-based label propagation problem, and propose a multi-modality anchor graph-based label prediction model for inferring the labels of the target candidate. The proposed label prediction model has four advantages. First, transductive learning strategy is exploited to perform the graph-based label prediction, in which the local structural information of both the tracking examples collected in previous frames and the target candidate examples (i.e. testing samples) sampled in current frame are utilized to learn the graphical representations of the relationship among these samples, in which each node denotes the features of examples in one specific modality, and the edges denote the pairwise affinity of two nodes. With more unlabeled examples in the new video frame introduced, the issues of limited samples problem can be alleviated. Such learning strategy also enables the objective of labeling target candidates to be coupled with the tracking model learning, which ensures the optimality of both objectives in our tracking task via an unified model. In addition, instead of optimizing the label of all samples which are in high computational complexity, the anchor label-based prorogation only requires to infer the labels of small number of anchor samples which can be used to infer the label of target candidate under the manifold assumption [14]. Besides, since the graph-based label prediction model can utilize some unlabeled data for learning, by regarding some contaminated examples (e.g. misaligned samples) which may degrade the modal discriminability as unlabeled data, label ambiguity can be avoided. Furthermore, an important weighting scheme is incorporated to dynamically adjust the modality weight according to their discriminative power. In addition to the proposed graph-based label prediction model, we propose to exploit the robust joint sparse representation model for constructing an accurate affinity matrix by simultaneous considering two issues:1) removing sample contamination, and 2) exploiting the correlation among different modalities. A fast relaxation of the sparse model and the optimization algorithm of related models are also derived. In general, the contributions of this work are summarized as follows:

- A multi-modality anchor graph-based label prediction model is proposed to predict the labels of target candidate
- A robust joint sparse representation model is formulated to estimate the affinity matrix among the tracking samples and the target candidates
- Effective model relaxation and learning algorithms are derived to solve the related optimization models.

It should be noted that some anchor-graph label propagation models have been proposed [15]–[17] which also diffuses the labels of anchor samples to unlabeled data. However, their model does not explicitly consider the label propagation for multi-modality samples. This work shares similar merits with them and focuses on label propagation on multi-modality graphs.

The rest of this paper is organized as follows. Section II introduces some related works which include RGB-infrared tracking and graph-based object tracking. Section III presents the proposed model and the related learning algorithms. Section IV describes the implementation details. The experimental evaluation and the conclusion are given in Sections V-B and VI, respectively.

II. Related Works

This section introduces some related works for our proposed method, i.e. RGB-infrared object tracking and graph-based object tracking.

A. RGB-Infrared Object Tracking

Numerous tracking algorithms for single modality video (i.e. RGB) has been developed. For example, to account for the variations existing in data sample, a latent constraint correlation filter is developed in [18]. To model the distribution of correlation response for mitigating the drift issue, an output constraint transfer model is incorporated in the correlation filter framework [19]. To further enhance the tracking robustness, inspired by promising performance of multi-modality/view machine learning in pattern classification [20], [21], several RGB-infrared tracking algorithms have been developed [18]. To simultaneously perform moving object segmentation and tracking, a level set-based framework is proposed [22]. In [23], the results of multiple spatiogram-based trackers which correspond to RGB and infrared modalities are combined within a new fusion-based tracking framework for determining the target position. In [24], a probabilistic background subtraction model is utilized to generate confidence maps of both RGB and infrared modalities, and then the confidence map is combined by using sum rule for determining the target position. There are several sparsity-based tracking algorithms developed for RGB-infrared tracking based on joint sparsity regularization [25], [26], feature concatenation [27], nuclear norm regularization [28], feature template learning [29]. These algorithms exploit sparsity constraint to exploit the correlation among different modalities. Li *et al.* proposed path-based dynamic graph for structure SVM-based tracking [30]. Different from these trackers which focus more on how to perform effective fusion of RGB-infrared modalities, our proposed method further considers the distribution properties of samples of multiple modalities to account for the appearance variations of the track target.

B. Graph-based Object Tracking

Graph-based machine learning has attracted great research interest in recent years [31]. Graph models have been exploited in various tracking algorithms because of its capability in representing structural relationships. In [32], a gragh-based transductive learning model is incorporated for label prediction under local and global constraints. In [33], a graph-based tracking model which considers both the neighborhood and pairwise information among samples are proposed to improve adaptability. In [34], a graph mode-based contextual kernel is introduced for SVM-based tracking. In [35], And-Or Graphs are exploited for simultaneously tracking, learning and Parsing. In [36], a random walk restart algorithm on 8-neighbor graph is developed to estimate the local patch weights within target object bounding box. Different from [36], Li *et al.* proposed patch based dynamic graph learning algorithms to estimate local patch weights in the bounding box, which has been applied in visual tracking [37] and RGB-T tracking [30]. A multi-graph ranking which propagates labels using multiple features is proposed in [38]. Wu *et al.* proposed a landmark-based label prorogation model for tracking [17]. Different from these graph-based tracking methods, the proposed graph model can exploit the heterogeneous data modalities to facilitate the label prediction performance.

III. Proposed Method

This section first give a overview of the problem and the idea of the proposed method, and then introduces the novel aspects of the proposed method: 1) joint sparsity-regularized anchor graph learning, and 2) multi-modality anchor graph-based label prediction.

A. Overview

Assume that we are given N tracking examples of M modalities (M=2 for our case of RGB-infrared tracking), which is composed of labeled and unlabeled examples and denoted as $\{x_n^m|n = 1,\ldots,N, m = 1,\ldots,M\}$. Without the loss of generality, let the first N_1 examples $\{x_n^m|n = 1,\ldots,N_1, m = 1,\ldots,M\}$ be the labeled examples with label vectors $y_n \in \mathbb{R}^{C\times 1}, n = 1,\ldots,N_1$ (C=2 for the tracking problem), and the remaining examples $\{x_n^m|n = N_1+1,\ldots,N_1+N_2, m = 1,\ldots,M\}$ be the unlabeled examples where $N = N_1 + N_2$. We are also given a set of anchor examples of multiple modalities, i.e. $\{d_r^m|m = 1,\ldots,M, r = 1,\ldots,R\}$ which can be obtained by taking cluster centers on $\{x_n^m\}$ after clustering is performed. Under the manifold (smoothness) assumption that the data points close to each other are more likely to share the same label, given unlabeled data $X^m = \{x_n^m\}$, one objective of the proposed model is to exploit the label vector of the nearby anchor points to predict their soft label of different modalities, i.e.

$$[\ell^m(x_1^m),\ldots,\ell^m(x_N^m)]^T = H^m A^m,\ m = 1,\ldots,M \quad (1)$$

where $\ell(x_{n'}^m) \in \mathbb{R}^{C\times 1}$ denotes the output of the labeling function on $x_{n'}^m$, $H^m \in \mathbb{R}^{N\times R}$, $H_{n,r}^m = (\mathcal{K}^m(x_n^m, d_r^m))$, which encodes the pairwise similarity between input example of the function x_n^m and the anchor example d_r^m, $A^m = \left(A_{r,c}^m\right) \in \mathbb{R}^{R\times C}$ and $A_{r,c}^m$ is the confidence value of c-th label of the r-th anchor example. To achieve this objective, several issues should be considered. Since large appearance variation may exist and contaminate the tracking sample during the tracking process, the first issue is how to deal with the contaminated samples and accurately estimate the similarity matrix. Besides, the number of anchor examples is limited, and how to effectively and efficiently propagate the label of anchor examples to large number of unlabeled data is the second issue we need to consider. To address these two issues, inspired by the graph-based semi-supervised scalable learning [15], we will exploit adjacency matrix-based graphical representation to capture the pairwise relationship among all the examples using anchor examples, for which the adjacency matrix would be constructed based on the estimated similarity matrix between the examples and anchor examples.

B. Joint Sparsity Regularized Multi-modality Anchor Graph Learning

a) Formulation of Anchor Graph: Inspired by the idea of locally linear embedding [39], the embedding weights can reflect the similarity between the example and the nearby anchor example. For the example X^m, since large appearance variation may also be encountered during the tracking process,

which would introduce (contaminated) corrupted samples, we further introduce the error term E^m to capture the sample-specific corruption as follows:

$$X^m = D^m(H^m)^T + E^m, H^m_{n,\cdot}\mathbf{1} = 1, H^m \geq \mathbf{0} \quad (2)$$

where $D^m = \{d^m_1, \ldots, d^m_R\}$. To capture the similarity among these examples, H^m should be learned to satisfied two conditions. The first one is the sparsity requirement that H^m should be a sparse matrix because only small number of nearby anchor examples is required to linearly approximate the examples. The second one is the nonnegative requirement that H^m should be nonnegative to make sure that the similarity value is interpretable and comparable.

From (2), we can see that the examples X^m is represented as the embedding weights with respect to the anchor samples D^m, and thus the embedding weights can be regarded as one kind of feature representations of the examples. An intuitive interpretation is that if two examples are similar, they should share similar neighborhood information which means similar anchor examples can be used to approximate these two example. Therefore, the embedding weights should be similar. Therefore, after the H^m is learned, the anchor graph of each modality W^m which reflect the affinity among these examples can be formulated as follows:

$$W^m = H^m(H^m)^T, m = 1, \ldots, M \quad (3)$$

Here we use the inner product (i.e. linear kernel) on the embedding weights to measure the similarity among samples. The formulation of W^m also shows that the sparse and nonnegative property of H^m can enable W^m to be positive and sparse which meet the requirements of adjacent matrix in label propagation. Therefore, imposing proper constraint on the learning of H^m is necessary.

b) Multi-modality Anchor Graph Learning: To ensure that only small number of nearby samples can be used for linear approximation, some sparsity constraints such as ℓ_1 norm can be utilized to enforce the sparsity. In addition, we further consider to exploit the interdependency of different modalities to facilitate the earning of a more accurate similarity matrix using multiple modalities. Therefore, we impose the joint sparsity constraint to enforce modality-consistency sparsity so that the consistency properties of different modalities can be utilized to deduct the same neighborhood information for different modalities. Then the joint sparsity regularized multi-modality anchor graph learning can be formulated as

$$\begin{aligned} \min_{\{E^m, H^m\}} & \lambda_1 \|(E^m)^T\|_{2,1} + \sum_{r=1}^{R} \|H_r\|_{2,1} \quad (4) \\ \text{s.t. } & H^m_{n,\cdot}\mathbf{1} = 1, H^m \geq \mathbf{0}, X^m = D^m(H^m)^T + E^m \\ & H_r = [(H^1_{r,\cdot})^T, \ldots, (H^m_{r,\cdot})^T] \end{aligned}$$

where $\mathbf{1}$ is the all 1's vector, $\|\cdot\|_{2,1}$ is the joint sparsity regularization that $\|A\|_{2,1} = \sum_{m=1}^{M} \sqrt{\sum_{n=1}^{N}(A_{m,n})^2}$ given $A \in \mathbb{R}^{M \times N}$. $H^m_{r,\cdot}$ means the r-th row vector of the matrix H^m, and the matrix H_r can be formed by putting all the transpose of the r-th row vector of of the matrix $H^1, \ldots, H^m$ together. Here we impose the joint sparsity constraint on the error term which aims to detect the sample-specific corruption in the tracking sample. If the n'-th example of m'-th modality is corrupted and cannot be well approximated by the anchor examples, then ℓ_2 norm on the corresponding error term $\|E^{m'}_{\cdot,n'}\|_2$ would be large. The second joint sparsity regularization is exploited to enforce different modalities of the same anchor examples are activated to linearly approximate the example. By inducing the consistent representation of anchor examples in different modalities, similar prorogation patterns in the multi-modality graph can be derived, which facilitates the confidence in label prediction.

However, directly solving (4) would be of high computational complexity because it involves some non-smooth terms. Therefore, we relax the solver to use KNN to (4) as follows: First, we select the nearby anchor examples for each input example using the concatenation of the RGB and infrared modalities by KNN. Then the corresponding nearby anchor examples would be used to approximate each example using LLE [39] for embedding weight estimation. To deal with corrupted samples, we make the following judgement. If the example is the tracking example (not the target candidate) in previous video frame with large approximation error which means the example may be corrupted, we replace the example with the mean feature vector of the tracking samples and then perform LLE again to estimate the embedding weights. In our implementation, the number of nearby anchor samples (i.e. the k in KNN) is set to 10, and the Euclidean distance is used as the criteria to select the nearby samples.

C. Multi-modality Anchor Graph-based Label Prediction

To construct the multi-modality gragh-based label prediction model, two objectives should be considered in constructing the model. Since the pair-wise similarity of different samples in each modality are captured by corresponding graph model, the graph-based label prediction results should be consistent with the similarity pattern. In addition, the label prediction error of the labeled training data should be as low as possible. The proposed label prediction model formulates the aforementioned objectives within the following framework:

$$\begin{aligned} \min_{\{\ell^m, \alpha^m\}} & \sum_{m=1}^{M} \left((\alpha^m)^2 G(\ell^m) + \eta F(\ell^m)\right) \quad (5) \\ \text{s.t. } & \sum_{m=1}^{M} \alpha^m = 1,\ \alpha^m \geq 0,\ m = 1, \ldots, M \end{aligned}$$

where ℓ^m is the prediction function, $G(\ell^m)$ is the graph-based regularization term of each modality, $F(\ell^m)$ is the prediction error term of the labeled training data of each modality, and α is the importance weight of each modality. Based on parametric form of $\ell(\cdot)$ in (1), given the estimated similarity matrix between the anchor examples and the training examples, obtaining the solution of ℓ^m is equivalent to estimate

A^m. Let $D_i^m = \sum_{j=1}^N W_{ij}^m$. Then G^m can be formulated as

$$\begin{aligned} G^m &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\ell^m(x_i^m) - \ell^m(x_j^m)\|_2^2 W_{ij}^m \\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|(A)^T(H_{i,\cdot}^m)^T - (A)^T(H_{j,\cdot}^m)^T\|_2^2 W_{ij}^m \\ &= \sum_{i=1}^{N} H_{i,\cdot}^m AA^T(H_{i,\cdot}^m)^T D_i^m - \sum_{i=1}^{N}\sum_{j=1}^{N} H_{i,\cdot}^m AA^T(H_{j,\cdot}^m)^T W_{ij}^m \\ &= \mathrm{Tr}(A^T(H^m)^T(D^m - W^m)H^m A) \\ &= \mathrm{Tr}(A^T(H^m)^T L^m H^m A) \end{aligned} \tag{6}$$

where L^m is the graph Laplacian regularization of the m-th modality. We can derive $L^m = D^m - W^m = H^m(H^m)^T - \mathrm{diag}\{\left(H^m(H^m)^T\right)\mathbf{1}\}$. We adopt the square loss as the prediction loss on the labeled training data Then (5) is re-formulated as

$$\begin{aligned} \min_{\{A,\alpha^m\}} & \begin{pmatrix} \mathrm{Tr}(A^T(\sum_{m=1}^M(\alpha^m)^2(H^m)^T L^m H^m)A) \\ +\eta\sum_{m=1}^M \|A^T(H_{l,\cdot}^m)^T - Y\|_F^2 \end{pmatrix} \\ \text{s.t.} & \sum_{m=1}^{M}\alpha^m = 1,\ \alpha^m \geq 0,\ m = 1,\ldots,M \end{aligned} \tag{7}$$

where $H_{l,\cdot}^m$ denote the similarity matrix between the labeled examples and the anchor example, and the square loss associated with the tradeoff parameter η corresponds to the label prediction error term for the labeled data $F(\ell^m)$. Here we use the $(\alpha^m)^2$ instead of α^m to avoid the trivial solution that only one important weight is activated to 1 while the others are zero. We can see that in (7), the label prediction model is formulated as minimizing a square loss of perdition with weighted fusion of multiple graph laplacian regularizer. Therefore, the structural information among examples of multiple modalities are jointly exploited through the regularizer for label propagation.

After obtaining the label prediction matrix A, given the sample of m-th modality $\hat{x}^m$, the label prediction results on $\hat{x}^m$ is calculated as

$$S_m = A^T H^m(\hat{x}^m)^T \tag{8}$$

where $H^m(\hat{x}^m)$ is the similarity weight between the sample $\hat{x}$ and the anchor samples of m-th modality. Following [17], we utilize the score of all samples (including the labeled and unlabeled ones) for normalization. The prediction score S_m can be used to infer the tracker position. Traditional graph-based label propagation algorithm usually need a cubic time complexity O(N^3) where N is the number of data point because $N \times N$ matrix inversion is needed for performing label prediction of all data point. For the proposed graph model, label prediction of all data points are based on the prediction label of smaller number of anchor samples, whose data size R is much smaller than N. Thus, the complexity is reduced to O(R^3), which means the label prorogation model is more efficient.

Optimization: To obtain the optimal solution to (7), we alternatively update the optimal variable $\{\alpha^m\}$ and A interatively. $\{A\}$-*subproblem:* With $\{\alpha^m\}$ fixed, by taking the derivatives of the objecting function in (7) with respect to A and setting it to be zero, we can derive

$$A = \left[\sum_{m=1}^{M}\left((\alpha^m)^2(H^m)^T L^m H^m + \eta(H_{l,\cdot}^m)^T(H_{l,\cdot}^m)\right)\right]^{-1} \bullet \left[\sum_{m=1}^{M} H_{l,\cdot}^m\right] \tag{9}$$

$\{\alpha^m\}$-*subproblem:* With A fixed, let $R^m = Tr(A^T(H^m)^T L^m H^m A)$, then α^m can be obtained by solving the following problem:

$$\begin{aligned} \min_{\{\alpha^m\}} & \sum_{m=1}^{M}(\alpha^m)^2 R^m \\ \text{s.t.} & \sum_{m=1}^{M}\alpha^m = 1,\ \alpha^m \geq 0,\ m = 1,\ldots,M \end{aligned} \tag{10}$$

By taking the derivatives of the Lagrange function of (10) i.e. $\mathcal{L}(\{\alpha^m\}) = \sum_{m=1}^M(\alpha^m)^2 R^m + \beta(\sum_{m=1}^M \alpha^m - 1)$, and setting it to be zeros. we can obtain $\alpha^m r^m + \beta = 0$. Based on the equality $\sum_{m=1}^M \alpha^m = 1$, we can derive $\alpha^{m'} = \frac{(R^{m'})^{-1}}{\sum_{m=1}^M (R^m)^{-1}}$. We iteratively update the optimal variables until the ℓ_2 norm of the value difference of optimal variables in consecutive iterations is less than the threshold 10^{-3}. The optimization algorithm is summarized in Algorithm 1.

Algorithm 1: Optimization Algorithm for (7)

Input: Graph Laplacian $\{L^m\}_{m=1}^M$, embedding weights $\{H^m\}_{m=1}^M$, sample number N and modality number M, An

Output: $\{\alpha^m\}_{m=1}^M$, A

Initialization: $i \leftarrow 1, \alpha^{m,i} \leftarrow 0.5$

while *stopping conditions are not satisfied* **do**

- Update A^{i+1} via (9)
- Update $\{\alpha^{m,i+1}\}_{m=1}^M$ via solving (10)
- $i \leftarrow i + 1$
- Check stopping conditions

end

IV. Tracker Implementation Detail

This section mainly describes some important implementation details of the proposed tracker. The proposed tracker is implemented in the same framework with [17].

A. Object Representation

To exploit the local structure information of target appearance for deal with local deformation and occlusion, besides exploiting holistic representation, we further use part-based features for the label prediction model training. Following the feature representation scheme in [17], 5 different image patches are sampled from the target region, which include the whole patch with sub-sampling rate 0.5, 4 local patches from 2-by-2 partitions. For each kind of image patches with

RGB-Infrared modalities, we will train a corresponding multi-modality classification model based on the (7) denoted as $S_{k,m}$, where $k = 1, \ldots, K, m = 1, \ldots, M$, k is the patch index and m is the modality index. By exploiting the sum rule, based on the modality important weight learned in (7) final classification score S_{total} is computed as

$$S_{\text{total}} = \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha^m S_{k,m} \tag{11}$$

B. Model Initialization and Updating

The target location in the first frame is initialized according to the annotation data. Then we shift the tracker location about 1 to 2 pixels to obtain the positive examples, and randomly sample image patches which are far away from the tracker location as negative samples. Then we apply the aforementioned object representation scheme to sample 5 different types of image patches. Standard clustering algorithms (k-means used in our method) are applied on the examples of each type of image patch in RGB-infrared modalities to select the center of each cluster as the anchor examples. Based on the anchor examples and the collected labeled examples, the initial label prediction model can be trained.

To enable the tracker to be adapted to appearance variation and background changes, model updating is essential. In the new video frame, target candidate will be sampled around the target location in the previous frame of RGB-infrared modalities using particle filtering. As mentioned in previous section, features extracted from the target candidates of RGB-infrared modalities would be regarded as unlabeled data of the label prediction model. After obtaining the classification scores of target candidates, the candidate with largest scores is the most likely to be the tracked target. After determining the tracking results, following the way of collecting sample in the first frame, we can collect more examples nearby and far away from the target position. If the classification score is higher than a predefined threshold, then the examples are reliable and is regarded as labeled example. Otherwise, they are regarded as unlabeled examples.

The tracker maintains two kinds of example pools for updating the example and anchor examples. They are training example pool and temporal example pool. The tracking results of RGB-infrared modalities in previous T frames would be put in the temporal example pool. The examples in the temporal example pool would be utilized to update the training example pool. The size of the training example pool is predefined and limited every T frames. If the training example pool is full, then the updating will be performed by being randomly replaced with T examples. For the sake of adaptivity and stability, we always augment the training example pool with the example in the initial frame after the model updating is performed. It should be noted that both the unlabeled target candidate examples and collected examples in the training example pool will be utilized for prediction model learning.

When the training example pool is updated, k-means clustering would be performed on the examples in the training example pool to obtain the cluster centers as the new potential anchor examples. Then k-means clustering are performed with the potential anchor examples with the previous ones again. Since the each examples in the training example pool carries feature representation of RGB-infrared modalities, for simplicity and efficiency while maintaining the correlation between RGB and infrared modalities in the anchor examples, the features of the k-means clustering are the concatenation of features extracted from the RGB and infrared modalities.

C. Target Position Estimation within Paritle Filtering Framework

The target position is estimated within the particle filtering framework. The tracking results at Frame t can be obtained by maximizing a posteriori:

$$\tilde{s}_t = \arg\max_{s_t^i} p(s_t^i | P_t) \tag{12}$$

where $P_t = \{p_j | j = 1, \ldots, t\}$ denote the observation variable set from Frame 1 to Frame t, p_j is the observation variable at Frame j, and s_t^i is the state variable of the i-th particle at Frame t. The particle filtering can be utilized to infer the true posterior by a set of particles with different states s_t^i where the posterior probability $p(s_t^i | P_t)$ is recursively computed as

$$p(s_t | P_T) \propto p(p_t | s_t) \int p(s_t | s_{t-1}) p(s_{t-1} | P_{t-1}) ds_{t-1} \tag{13}$$

where $p(s_t | s_{t-1})$ and $p(p_t | s_t)$ denote the motion model and the observation model, respectively. After obtaining the final classification score, then the observation likelihood can be defined as

$$p(s_t | p_t) \propto S_{\text{total}}(o_t) \tag{14}$$

V. Experiments

This section presents the experimental setting, and then describes the quantitative and qualitative results, respectively.

A. Experimental Setting

c) Testing data and compared methods: Twenty pairs of videos captured by visible and infrared cameras are used for evaluation of the tracking performance. Challenging factors such as occlusion, thermal crossover, poor illumination condition can be found in these testing videos. All the video frames of RGB and infrared modalities have been aligned via registration so that the target position is nearly the same in each video frame of both modalities. We adopt 10 tracking algorithms for comparison, which includes STC [40], CT [41], RPT [42], STUCK [43], MIL [12], L1 [27], JSR [26], CN [44], KCF [45], MEEM [46]. Except the L1 and JSR methods which are designed specifically for RGB-infrared tracking, the other trackers only focus on tracking on RGB videos. The experimental section in [25] provides the way to implement the multi-modality version of these trackers. We can follow them and compare the results with our proposed method. Some of the compared results of these multi-modality trackers on these video data can also be obtained from [25].

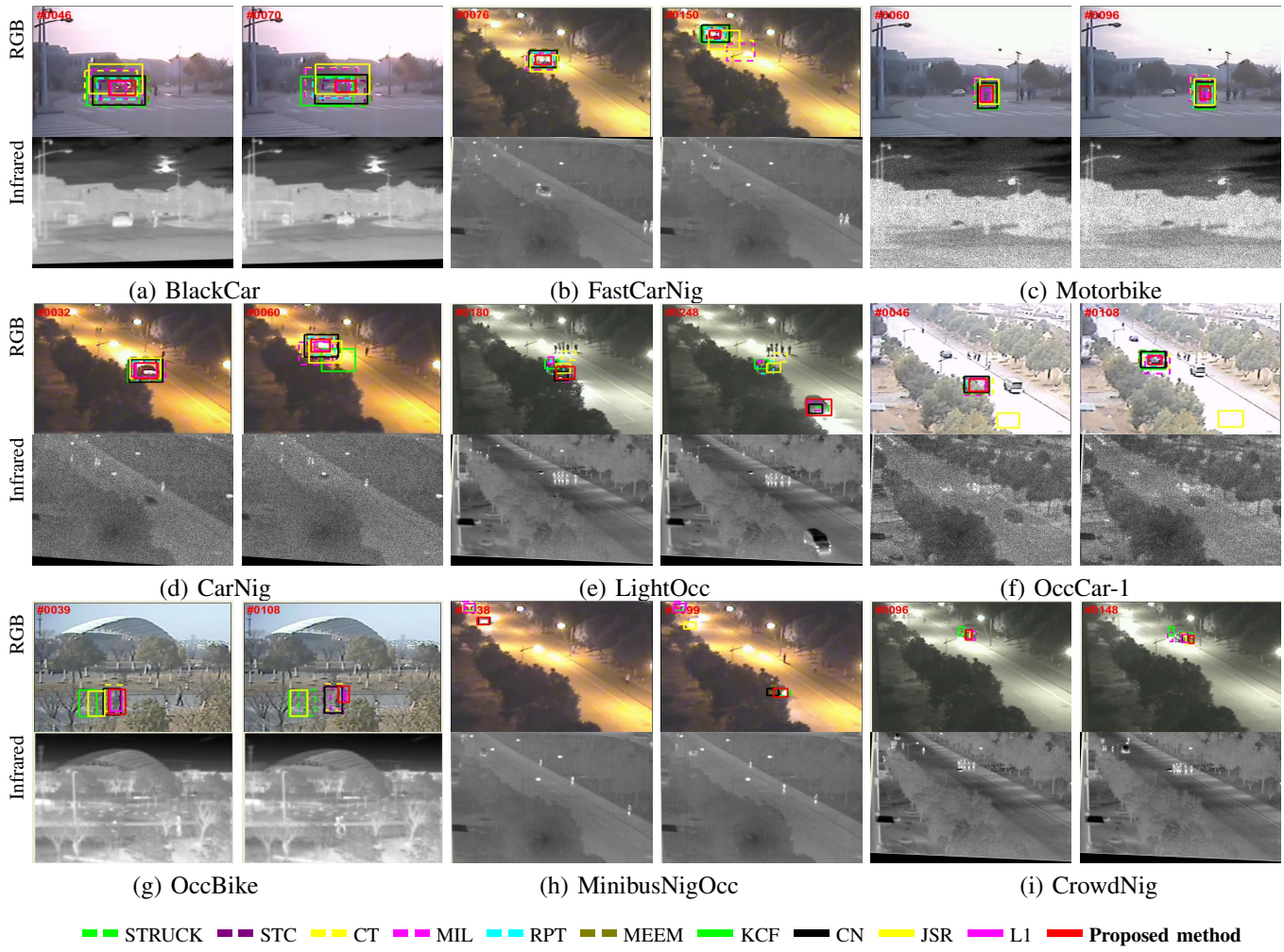


Fig. 2. Qualitative comparison of the 11 trackers on some frames of RGB-infrared videos covering some large variations, such as occlusion (e.g. *OccCar-1*, *MinibusNigOcc*),scale variatons(e.g. *BlackCar*), Thermal crossover (e.g. *CrowdNig*), low illumination (e.g. *CarNig*, *FastCarNig*). For each sub-figure, images of RGB modality are shown in the top row while images of infrared modality are shown in the bottom row.

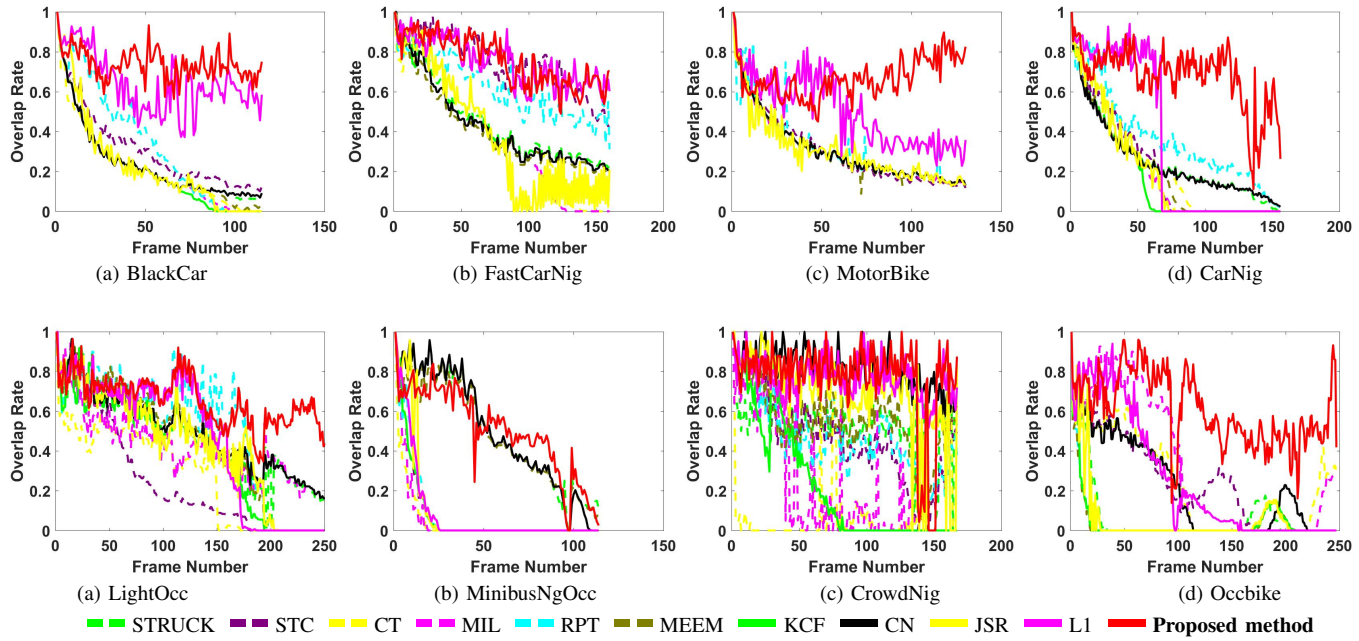


Fig. 3. Quantitative results of 11 trackers on 8 challenging videos in terms of overlapping rate. The frame index is shown in the horizontal axis and the overlapping rate is indicated in the vertical index.

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

	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	JSR	L1	**Proposed Method**
BlackCar	0.24	0.31	0.21	0.22	**0.33**	0.23	0.21	0.24	0.23	**0.64**	**0.74**
BlueCar	0.37	0.27	0.34	0.4	**0.65**	0.47	0.4	0.4	0.4	**0.63**	**0.79**
BusScale	0.47	0.45	0.46	0.49	**0.57**	0.52	0.51	0.51	0.54	**0.72**	**0.81**
Exposure2	0.32	0.37	0.31	0.32	**0.48**	0.3	0.32	0.32	0.35	**0.82**	**0.82**
FastCarNig	0.46	**0.75**	0.36	0.36	0.63	0.41	0.43	0.43	0.38	**0.75**	**0.76**
Motorbike	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	0.3	**0.31**	**0.31**	0.3	**0.5**	**0.67**
CarNig	0.25	0.21	0.2	0.18	**0.36**	0.19	0.16	0.25	0.2	**0.35**	**0.68**
Cycling	0.62	0.47	0.51	**0.64**	0.55	0.03	0.61	**0.63**	0.49	0.36	**0.74**
Minibus1	0.53	0.05	0.52	0.55	0.06	0.38	**0.56**	0.05	0.53	**0.69**	**0.79**
FastMotor	**0.43**	0.24	**0.43**	0.42	0.36	0.37	0.4	0.41	0.41	0.02	**0.46**
LightOcc	0.46	0.25	0.27	0.43	**0.5**	0.45	0.41	**0.5**	0.43	0.46	**0.66**
Minibus	**0.43**	**0.46**	0.42	0.34	**0.43**	0.39	0.41	0.41	0.42	0.37	**0.82**
MinibusNigOcc	**0.51**	0.07	0.04	0.05	0.08	0.48	0.07	**0.5**	0.1	0.08	**0.5**
OccCar-1	0.45	0.46	0.43	0.33	**0.68**	0.41	0.45	0.45	0.07	**0.82**	**0.78**
Otcbvs1	0.63	0.69	0.65	**0.73**	0.72	0.66	0.66	0.68	**0.79**	0.14	**0.82**
Pool	**0.62**	0.06	0.05	0.05	0.04	**0.66**	0.03	0.06	0.05	0.06	**0.16**
RainyCar1	**0.58**	0.5	0.55	0.07	**0.69**	0.49	0.55	0.55	0.05	0.07	**0.72**
Running	0.22	0.32	0.17	0.14	**0.38**	**0.37**	0.3	0.33	0.18	0.36	**0.41**
CrowdNig	0.51	0.46	0.12	0.23	0.46	0.55	0.22	**0.81**	0.69	**0.77**	**0.77**
OccBike	0.07	0.24	0.23	**0.31**	0.04	0.04	0.04	0.21	0.06	**0.26**	**0.61**
Average	**0.42**	0.35	0.33	0.33	**0.42**	0.39	0.35	0.4	0.33	**0.44**	**0.68**

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

	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	JSR	L1	**Proposed Method**
BlackCar	0.12	0.16	0.1	0.12	**0.29**	0.12	0.12	0.12	0.15	**0.83**	**1**
BlueCar	0.33	0.33	0.28	0.38	**0.94**	0.46	0.38	0.38	0.44	**0.68**	**0.97**
BusScale	0.48	0.4	0.46	0.44	**0.61**	0.53	0.5	0.51	0.56	**0.82**	**0.98**
Exposure2	0.2	0.26	0.2	0.2	**0.45**	0.16	0.2	0.2	0.19	**1**	**1**
FastCarNig	0.31	**0.93**	0.28	0.28	0.73	0.26	0.28	0.28	0.39	**1**	**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**
CarNig	0.13	0.19	0.13	0.13	**0.21**	0.13	0.13	0.13	0.17	**0.43**	**0.87**
Cycling	**0.71**	0.43	0.53	**0.71**	0.68	0.02	**0.71**	**0.71**	0.48	0.33	**0.99**
Minibus1	**0.59**	0.04	0.54	0.58	0.05	0.32	0.54	0.04	0.49	**0.69**	**1**
FastMotor	0.32	0.19	**0.33**	**0.33**	0.3	0.2	0.27	0.3	0.27	0.02	**0.45**
LightOcc	0.4	0.23	0.18	0.29	**0.69**	0.56	0.46	0.48	0.46	**0.6**	**0.92**
Minibus	0.27	**0.42**	0.27	0.24	0.25	0.21	0.27	0.27	0.24	**0.32**	**1**
MinibusNigOcc	**0.46**	0.06	0.02	0.04	0.08	0.45	0.06	**0.46**	0.11	0.08	**0.51**
OccCar-1	0.32	0.44	0.27	0.24	**0.89**	0.21	0.32	0.32	0.08	**1**	**1**
Otcbvs1	0.91	0.87	0.98	**1**	0.98	0.94	0.84	0.82	**1**	0.12	**1**
Pool	**0.85**	0.08	0.06	0.06	0.04	**0.83**	0.04	0.08	0.04	0.08	**0.09**
RainyCar1	**0.58**	0.35	0.55	0.08	**0.98**	0.45	0.57	0.57	0.05	0.07	**1**
Running	0.27	0.45	0.17	0.13	**0.51**	**0.47**	0.43	0.43	0.22	0.46	**0.53**
CrowdNig	0.67	0.38	0.16	0.26	0.34	0.67	0.2	**1**	0.92	**1**	**0.93**
OccBike	0.06	0.18	0.2	**0.38**	0.02	0.02	0.03	0.15	0.05	**0.26**	**0.69**
Average	0.41	0.33	0.29	0.3	**0.46**	0.36	0.32	0.37	0.32	**0.51**	**0.85**

d) Parameter settings: The target region is warped and resized to 24-by-24 image patch, and thus all the five image patches for model learning are 12 by 12. For the image patch in RGB modality, we extract HOG features and gray scale intensity features, and concatenate them into a single vector. For infrared modality, we extract the intensity features only. In the initial frame, the number of positive examples and negative examples are 20 and 200, respectively. The threshold for determining the reliability of tracking sample is set to 0.3. After obtaining the tracking results, 2 positive example and 50 negative examples will be collected if the results are reliable and 100 unlabeled data would be collected if the result is not reliable. The limitation of the number of examples in the training example pool is set to 310 which include 50 positive examples, 160 negative examples, and 100 unlabeled examples. The number of anchor samples is set to 30 and the model updating is performed every 10 frames.

B. Experimental Results

To evaluate the tracking performance quantitatively, two criteria are used, i.e. overlapping rate and success rate. The overlapping rate is defined as $\frac{area(A_1 \bigcap A_2)}{area(A_1 \bigcup A_2)}$ where A_1 and A_2 are the bounding box generated by the tracker and the groundtruth. The tracking in each frame is counted as a success if the overlapping rate is greater than 0.5. The percentage of video frames in which a tracking success happens is used to define the success rate. The overlapping rate and the success rate of all the compared methods in the twenty videos are shown in Tables I and II, respectively. We can see that generally, the proposed tracker outperforms other compared methods. It can also be observed that that the proposed tracker ranks in top two on nineteen videos in terms of success rate and overlapping rate, where the top one performance is achieved on sixteen videos in terms of overlapping rate and on seventeen videos in terms of success rate. As shown in Figure 2, the proposed tracker can more able to deal with some large variations such as occlusion (*OccCar-1*#46, *MinibusNigOcc*#99), thermal crossover (*CrowdNig*#46). This is because 1) the proposed tracker can suppress the contaminated features when constructing the anchor graph, which makes it less sensitive to the outliers introduced by occlusion; 2) full utilization of different kinds of tracking examples especially these unlabeled can further enhance the discriminative power of the tracker; 3) dynamically adjusting of the importance weight of different modalities for better discrimination between target and background. The frame-by-frame quantitative results in terms of overlapping rate are shown in Figure 3. It shows that the proposed tracker can achieve a generally higher overlapping rate in these video frames compared with other methods. This shows that the proposed tracker can run more stably.

However, in some videos such as *Pool*, when the target of small size encounters some full body occlusion, tracking loss may be happen as shown in Frame 20 in Figure 4(a). In

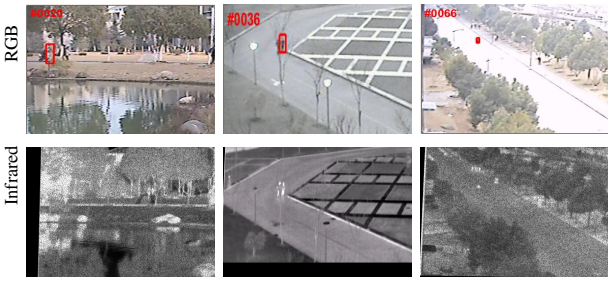


Fig. 4. Results in testing videos which are not excellent: (a) Pool (b) Running (c) FastMotor.

addition, when the tracked target is in low resolution as shown in the video *FastMotor* (e.g. Frame 66 in Figure 4(c)) and the nearby object of similar size and shape exist as shown in video *Running* (e.g. Frame 36 in Figure 4(b)), the tracker may not be able to achieve excellent performance.

Running Speed: Since the proposed algorithm involves in some iterative optimization, it can not run in real time, and the running speed is about 2 frame per second.

VI. Conclusion

In this paper, we propose an new discriminative model for RGB-infrared object tracking. A joint sparsity-regularized multi-modality anchor graph learning model is developed to learn the affinity matrix for constructing the multi-modality anchor graph, and a multi-modality anchor graph-based label prediction model is designed to efficiently propagate limited number of labeled examples to predict the labels of target candidates. Comparison experimental results with other 10 trackers on 20 videos demonstrate the effectiveness of the proposed method.

Since the proposed tracker can not run in real time, one of our future work is to improve the tracking by exploiting or developing more advanced optimization algorithm. In addition, since the similarity graph is constructed by some relaxation with KNN algorithm under the similarity measurement with Euclidean distance which may not be optimal, we will further investigate an optimal way for the relaxation of the similarity graph learning problem.

References

[1] 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. Image Processing*, vol. 28, no. 4, pp. 1993–2007, 2019.

[2] G. Ding, Y. Guo, K. Chen, C. Chu, J. Han, and Q. Dai, "DECODE: deep confidence network for robust image classification," *IEEE Trans. Image Processing*, vol. 28, no. 8, pp. 3752–3765, 2019.

[3] 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.

[4] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Trans. on Information Forensics and Security*, vol. 14, no. 10, pp. 2537–2550, 2019.

[5] X. Peng, C. Lu, Z. Yi, and H. Tang, "Connections between nuclear-norm and frobenius-norm-based representations," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 29, no. 1, pp. 218–224, 2018.

[6] Z. Zhang, W. Jiang, J. Qin, L. Zhang, F. Li, M. Zhang, and S. Yan, "Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 29, no. 8, pp. 3798–3814, 2018.

[7] Z. Zhang, J. ren, W. Jiang, Z. Zhang, R. Hong, S. Yan, and M. Wang, "Multi-manifold positive and unlabeled learning for visual analysis," *IEEE Trans. Circuits and Systems for Video Technology*, 2019, DOI:10.1109/TCSVT.2019.2923007.

[8] 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.

[9] 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*, vol. 21, no. 5, pp. 1183–1194, 2019.

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

[11] 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.*, vol. 41, no. 5, pp. 1116–1130, 2019.

[12] 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.

[13] 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.

[14] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[15] W. Liu, J. He, and S. Chang, "Large graph construction for scalable semi-supervised learning," in *Proc. ICML*, 2010, pp. 679–686.

[16] M. Wang, W. Fu, S. Hao, D. Tao, and X. Wu, "Scalable semi-supervised learning by efficient anchor graph regularization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1864–1877, 2016.

[17] Y. Wu, M. Pei, M. Yang, J. Yuan, and Y. Jia, "Robust discriminative tracking via landmark-based label propagation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1510–1523, 2015.

[18] B. Zhang, S. Luan, C. Chen, J. Han, W. Wang, A. Perina, and L. Shao, "Latent constrained correlation filter," *IEEE Trans. Image Processing*, vol. 27, no. 3, pp. 1038–1048, 2018.

[19] 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.

[20] C. Gong, D. Tao, X. Chang, and J. Yang, "Ensemble teaching for hybrid label propagation," *IEEE Trans. Cybernetics*, vol. 49, no. 2, pp. 388–402, 2019.

[21] C. Chen, H. Qian, W. Chen, Z. Zheng, and H. Zhu, "Auto-weighted multi-view constrained spectral clustering," *Neurocomputing*, 2019, DOI:10.1016/j.neucom.2019.06.098.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. ACM MM*, 2017, pp. 1856–1864.
[31] C. Gong, H. Shi, J. Yang, J. Yang, and J. Yanga, "Multi-manifold positive and unlabeled learning for visual analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, DOI:10.1109/TCSVT.2019.2903563.
[32] Y. Zha, Y. Yang, and D. Bi, "Graph-based transductive learning for robust visual tracking," *Pattern Recognition*, vol. 43, no. 1, pp. 187–196, 2010.
[33] 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.
[34] X. Li, A. R. Dick, H. Wang, C. Shen, and A. van den Hengel, "Graph mode-based contextual kernels for robust SVM tracking," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2011, pp. 1156–1163.
[35] T. Wu, Y. Lu, and S. Zhu, "Online object tracking, learning and parsing with and-or graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2465–2480, 2017.
[36] H. Kim, D. Lee, J. Sim, and C. Kim, "SOWP: spatially ordered and weighted patch descriptor for visual tracking," in *Proc. ICCV*, 2015, pp. 3011–3019.
[37] 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.
[38] X. Yang, M. Wang, and D. Tao, "Robust visual tracking via multi-graph ranking," *Neurocomputing*, vol. 159, pp. 35–43, 2015.
[39] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in *Proc. ECCV*, 2014, pp. 188–203.



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, where he is currently a Research Assistant Professor. His current research interests include intelligent video surveillance and biometric security.



Wei Zhang (S'06-M'11) received the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong in 2010. He is currently a Professor with the School of Control Science and Engineering, Shandong University, China. He has authored over 60 papers in international journals and refereed conferences. His research interests include computer vision, image processing, pattern recognition, and robotics. He served as a program committee member and a reviewer for various international conferences and journals in image processing, computer vision and robotics.



Shengping Zhang (M'13) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He was a Post-Doctoral Research Associate with Brown University and with Hong Kong Baptist University, and a Visiting Student Researcher with the University of California at Berkeley. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai. He has authored or co-authored over 50 research publications in refereed journals and conferences. His research interests include deep learning and its applications in computer vision. He is an Associate Editor of the Signal, Image and Video Processing and the Journal of Electronic Imaging.



Deepak Kumar Jain received the B.Eng. degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, India, in 2010, the M.Tech. 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 Assistant Professor with the Institute of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China. 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. He was an Awardee of the CAS-TWAS Presidential Fellowship, from 2014 to 2018. He was invited as "Foreign Experts" by the Shandong Taian Administration of foreign Expert Affairs.



Huiyu Zhou received the Bachelor's of Engineering degree in radio technology from the Huazhong University of Science and Technology of China, Wuhan, China, and the Master's of Science degree in biomedical engineering from the University of Dundee, Dundee, U.K., respectively. He received the Doctor of Philosophy degree in computer vision from Heriot-Watt University, Edinburgh, U.K. He is currently a Reader with the Department of Informatics, University of Leicester, Leicester, U.K. He has authored more than 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. Dr. Zhou serves as the Editor-in-Chief of Recent Advances in Electrical and Electronic Engineering and an Associate Editor of the IEEE Transaction on Human¨CMachine Systems, Editorial Board Members of several refereed journals.