Pattern Recognition Letters
journal homepage: www.elsevier.com

Modality-Correlation-Aware Sparse Representation for RGB-Infrared Object Tracking

Xiangyuan Lan[a], Mang Ye[a,**], Shengping Zhang[b], Huiyu Zhou[c], Pong C. Yuen[a]

[a]*Department of Computer Science, Hong Kong Baptist University, Hong Kong, China*
[b]*School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China*
[c]*Department of Informatics, University of Leicester, Leicester, U.K.*

ABSTRACT

To intelligently analyze and understand video content, a key step is to accurately perceive the motion of the interested objects in videos. To this end, the task of object tracking, which aims to determine the position and status of the interested object in consecutive video frames, is very important, and has received great research interest in the last decade. Although numerous algorithms have been proposed for object tracking in RGB videos, most of them may fail to track the object when the information from the RGB video is not reliable (e.g. in dim environment or large illumination change). To address this issue, with the popularity of dual-camera systems for capturing RGB and infrared videos, this paper presents a feature representation and fusion model to combine the feature representation of the object in RGB and infrared modalities for object tracking. Specifically, this proposed model is able to 1) perform feature representation of objects in different modalities by employing the robustness of sparse representation, and 2) combine the representation by exploiting the modality correlation. Extensive experiments demonstrate the effectiveness of the proposed method.

1. Introduction

Developing a reliable object tracker is very important for intelligent video analysis, and it plays the key role in motion perception in videos (Chang et al. (2017b,a); Chang and Yang (2017); Li et al. (2017b); Ma et al. (2018); Wang et al. (2017, 2016b); Luo et al. (2017)). While significant progress in object tracking research has been made and many object tracking algorithms have been developed with promising performance (Ye et al. (2015, 2016, 2017, 2018b); Zhou et al. (2018b,a); Ye et al. (2018a); Liu et al. (2018); Lan et al. (2018a); Zhang et al. (2013b, 2017d,c, 2018c); Song et al. (2017, 2018); Zhang et al. (2017b, 2016, 2018a); Hou et al. (2017); Yang et al. (2016); Zhong et al. (2014); Guo et al. (2017); Ding et al. (2018); Shao et al. (2018); Yang et al. (2018b,a); Pang et al. (2017)), it is worth noting that most of these trackers are designed for tracking objects in RGB image sequences, in which they model the object's appearance via the visual features extracted from RGB video frames. This may limit them to be employed in real applications, such as tracking objects in a dark environment where the lighting condition is poor and the RGB information is not reliable.

Recent years have witnessed an increasing number of vision systems equipped with both RGB and thermal infrared cameras. The development and popularity of multi-spectral imaging techniques further make it effective and efficient for these systems to capture the RGB and thermal infrared videos. Different from RGB cameras which form images using visible light, infrared cameras can capture the infrared radiation of a subject for forming an image, and thus its imaging procedure is not sensitive to lighting conditions. Therefore, to perform more reliable object tracking in more challenging practical scenarios, it is very important to incorporate information from the infrared modality into feature representation and appearance modeling of the tracked object.

To perform RGB-infrared tracking, how to properly fuse the information from RGB and infrared modalities is a key issue which should be considered. It should be noted that RGB images and infrared images are intrinsically different in their visual characteristics (e.g. texture, color) as shown in Figure 1, which leads to the gap between the statistical properties of their features in two different modalities. Therefore, to perform effective fusion of RGB and infrared modalities, mining and ex-

**Corresponding author: mangye@comp.hkbu.edu.hk (Mang Ye)

ploiting the correlation between these heterogeneous modalities is very important. While exploiting the correlation, it is also very important to reflect the complementarity of different modalities (Lan et al. (2018b)), which means the modality-specific properties should also be modeled.

Although several RGB-infrared tracking algorithms have been developed, they do not properly exploit the correlation and the specific properties between heterogeneous modalities for effective modality fusion. One typical approach is to apply some feature combination methods for modality fusion, such as feature concatenation (Wu et al. (2011)), sum rule (Leykin and Hammoud (2010)). These methods may ignore the different statistical properties of different modalities, and do not explicitly model the correlation of different modalities, which means the gap of different modalities still exists when modality fusion is performed. Another kind of approaches such as (Conaire et al. (2008)) regard the tracking of RGB and infrared modalities as two independent tasks and combine the tracking results of these two tasks for target localization. These methods do not model the correlation of different modalities during the tracking process. Although there are some other methods such as joint sparsity-based methods (Liu and Sun (2012); Li et al. (2016)) which exploit the strength of multi-task learning and formulate tracking on two modalities as two correlated representation learning tasks, they may impose a strict constraint on the fusion (i.e. enforcing them to share the same representation pattern), which ignores the modality-specific representation patterns and limits the utilization of the specific properties of different modalities. As such, the complementary properties of two modalities is not well exploited.

To address the aforementioned issues, we have developed a feature fusion model for RGB-infrared object tracking. The developed model utilizes the robustness of sparse representation for appearance modeling while exploiting the correlation between different modalities for effective modality fusion. In addition, different from the existing modality-correlation-aware RGB-infrared object trackers (e.g. joint sparsity based tracker (Liu and Sun (2012); Li et al. (2016)) which may ignore the modality specific properties, by imposing the low rank regularization (Candès and Recht (2009)), our proposed model exploits a way of soft regularization which simultaneously reveals the correlation of different modalities while mining certain modality specific representation for appearance modeling. Therefore, our proposed model jointly exploits the correlation and complementarity of the RGB and infrared modalities for representation learning in an unified optimization framework, which fully unleashes the representation capability of different modalities to deal with the appearance variations during the tracking process. Moreover, to guarantee the optimality of the developed model, we derive an iterative optimization algorithm to learn the feature fusion model. Experiments show that the developed model and the derived learning algorithm are both effective for RGB-infrared tracking.

In summary, the contribution of this paper are listed as follows:

- A modality-correlation-aware sparse representation model is developed to fuse multiple sparse representations for appearance modeling in RGB-infrared tracking.
- A effective optimization algorithm is derived to solve the modality-correlation-aware sparse representation problem.
- Extensive experiments are performed to show the effectiveness of the developed model.

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

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

The rest of this paper is organized as follows. In Section 2, we first review some related works on RGB-Infrared object tracking and sparse representation-based visual tracking, and then some works about multi-modality classification and recognition methods will also be introduced. In Section 3, we present our developed tracking model and its corresponding learning algorithm. We describe the implementation details in Section 4. Experimental analysis and conclusion are given in Sections 5 and 6, respectively.

2. Related Work

Since the developed model is related to sparse representation and modality fusion, besides the discussion of some related work in RGB-infrared tracking, we further introduce some related work on sparse representation-based visual tracking and multi-modality recognition. For more comprehensive literature review of visual tracking research, interested readers can refer to (Zhang et al. (2013a); Wu et al. (2015); Smeulders et al. (2014); Li et al. (2013); Salti et al. (2012)).

2.1. RGB-Infrared Object Tracking

Several methods have been developed for object tracking in RGB-infrared videos. In (Bunyak et al. (2007)), a level set-based RGB-infrared moving object segmentation and tracking framework is developed (Bunyak et al. (2007)). (Conaire et al. (2008)) proposed to fuse the results of a multiple spatiogram tracker on RGB and infrared modalities for determining the final target position. In (Leykin and Hammoud (2010)), sum rule was utilized to aggregate the confidence maps from RGB and infrared modalities based on a probabilistic background model, and the fused confidence map was used to determine the position of the tracked pedestrian. Several sparsity-based tracking algorithms were also proposed among which representation in different modalities are fused via concatenation (Wu et al.

(2011)), joint sparsity regularization (Liu and Sun (2012); Li et al. (2016)). These methods can exploit the correlation of different modalities to some extent. However, most of them utilized strict regularization and ignore the modality specific properties.

2.2. Sparse Representation-based Visual Tracking

Sparse Tracker Using A Single Feature. Lots of tracking algorithms have been proposed for tracking in RGB videos and they achieved promising performance. For example, Zhang et al. (2018b) developed a laten constrained correlation filter framework to deal with distorted tracking samples. In (Zhang et al. (2017a)), an output constraint transfer (OCT)-based correlation filter method which models the distribution of correlation response in a Bayesian optimization framework was developed to mitigate the drifting problem. In Zhang et al. (2016), multiple Gaussian uncertainty theory deeply investigate the tracking elements from a new viewpoint. Inspired by the success of sparse and low rank representation in pattern classification (Wright et al. (2010); Peng et al. (2018, 2017, 2016)), to enhance the robustness of trackers to large appearance change, sparse representation was exploited in appearance modeling for visual tracking. Mei and Ling (2011) proposed the ℓ_1 tracker which sparsely represents the target's appearance using feature templates collected from previous frames. Along this line, much efforts have been done to improve the robustness and the efficiency of sparse representation-based trackers by removing contaminated samples (Lan et al. (2016)), exploiting local structural information (Ma et al. (2015, 2016); Wang et al. (2015b); Zhang et al. (2015b)), improving computation complexity (Zhang et al. (2015a)), exploiting the correlation among particles (Zhang et al. (2013c)), imposing discriminative information (Lan et al. (2017)), learning dictionary (Liu et al. (2016)), performing subspace learning (Sui et al. (2015)) and so on. However, these trackers only exploit one single feature extracted from RGB modality (e.g. grey intensity) for appearance modeling, and they may not be able to handle the irregular and complicated appearance changes with a single feature.

Sparse Tracker Using Multiple Features. The use of multiple visual cues/features has been shown to be beneficial for visual tracking (Wang et al. (2014); Yuan et al. (2014)), and there are several works which fuse multiple sparse representations from multiple features for appearance modeling. Inspired by the multi-task joint sparse representation (Yuan et al. (2010)), Hu et al. (2015b) proposed a joint sparse representation-based tracking algorithm which combines multiple features based on joint sparsity constraint for appearance modeling. To exploit the relationship among different features of different particles in a particle filtering-based tracking framework, Hong et al. (2013) formulates visual tracking as a multi-task multi-view sparse learning model which combines multiple features while detecting outlier particles. In (Lan et al. (2014)), a robust joint sparse representation model was proposed to perform feature-level fusion of multiple visual cues and remove unreliable features. Along this line, Lan et al. (2015) further developed a kernelized joint sparse representation model which utilizes the non-linearity of features and combine features from different kernel space. In (Lan et al. (2018b)), a multiple sparse representation framework with appearance proximity constraint was developed which fuses multiple features and exploits the multi-feature similarity of object appearance for appearance modeling. Although most of these trackers can exploit the correlation of multiple features for feature fusion-based appearance modeling, they exploit a strict regularization (e.g. joint sparsity constraint) which leads to a inflexible representations of different features and limits the exploitation of feature-specific properties.

2.3. Multi-Modality Classification and Recognition

To improve the performance of image classification and object recognition, with the increasing number of multi-modal sensors and the available multi-modality big data, multi-modality classification and recognition has been received great research interests in recent years (Han et al. (2012, 2013)). For example, Hu et al. (2015a) proposed a new learning model which mines the shared and modality-specific structures of the RGB-D modality for heterogeneous features learning. Wang et al. (2015a) proposed a multimodal sharable and specific feature learning algorithm to obtain features which reflect the shared and modal-specific properties for RGB-D object recognition. In (Wang et al. (2016a)), RGB-D scene classification was performed by learning and fusing modality and component aware features. However, some of these algorithms need some off-line training data for model learning which may not be able to be applied in online multi-modality tracking.

3. Proposed Method

This section mainly introduces two key aspects of the proposed algorithm which includes: 1) modality-correlation-aware sparse representation, and 2) the optimization algorithm for model learning.

3.1. Modality-Correlation-Aware Sparse Representation

In our tracking framework, we have object template sets of different modalities, denoted as $X^m = [x_1^m, \ldots, x_N^m], m = 1, \ldots, M$, where m is the index of the modalities, M is the number of the modalities, and N is the number of templates in the template set. Our tracking model is to use the linear combination of feature templates to represent the tracking result of different modalities, denoted as $y^m, m = 1, \ldots, M$, which is shown as follows:

$$y^m = X^m w^m + \epsilon^m, m = 1, \ldots, M \tag{1}$$

where ϵ^m is the vector which characterizes the representation error in the m-th modality, and $w^m \in \mathbb{R}^N$ is the coefficient vector of the m-th modality which are used for linear combination of the object templates.

How to determine the coefficient vectors of different modalities for accurate object representation is the key problem which should be considered for constructing the tracking model. The object templates of different modalities should be able to handle different appearance variations of the tracked target during

the tracking process, which means each feature template characterizes one distinct aspect of the object appearance. Therefore, it is very important to adaptively select representative and informative feature templates to deal with different appearance changes. In addition, to performance effective modality fusion of different representation, it is also crucial to bridge the gap among different modalities by mining the correlation of different modalities. Based on the aforementioned considerations, we formulate the multi-modality tracking problem as learning model-correlation-aware sparse representations:

$$\min_{\{w^m\}_{m=1}^M} \sum_{m=1}^{M} \frac{1}{2}\|y^m - X^m w^m\|_2^2 + \lambda_1 \Omega(\{w^k\}_{m=1}^M) + \lambda_2 \sum_{m=1}^{M} \|w^m\|_1 \quad (2)$$

where the first term $\sum_{m=1}^{M} \|\cdot\|_2^2$ represents the total reconstruction error of the tracked target using the feature templates in different modalities, the second term $\Omega(\cdot)$ is the regularization function which aims to discover the correlation of different modality representations, the third term $\sum_{m=1}^{M} \|\cdot\|$ is the sparsity regularization which aims to select representative templates in different modalities for appearance modeling, and λ_1 and λ_2 denote the trade-off parameters of different terms.

To facilitate bridging the gap between different modalities, the learning model in Eq. (2) should be able to learn representation vectors of different modalities which are correlated with each other as highly as possible. Let $W = [w^1, \ldots, w^M]$, maximizing the correlation among the representation vectors of different modalities $w^1, \ldots, w^M$ can be achieved by minimizing the rank of W. Therefore, the second term $\Omega(\cdot)$ in Eq. (2) is defined as

$$\Omega(\{w^k\}_{m=1}^M) = \text{rank}(W), W = [w^1, \ldots, w^M] \quad (3)$$

Different from joint sparsity regularization (Liu and Sun (2012); Li et al. (2016)) which exploits the modality correlation by strictly enforcing the representation of different modalities to share the same sparsity pattern, the low rank regularization allow the sparsity pattern of different presentation to be different while maintaining the correlation. Therefore, the modality-specific pattern can be also discovered which may help to utilize the modality-specific properties, which is useful to exploit the complementarity of different modalities for appearance modeling. With the formulation in Equation (3), our tracking model in Eq. (2) can be re-written as

$$\begin{aligned} &\min_{\{w^m\}_{m=1}^M} \sum_{m=1}^{M} \frac{1}{2}\|y^m - X^m w^m\|_2^2 + \lambda_1 \text{rank}(W) + \lambda_2 \sum_{m=1}^{M} \|w^m\|_1 \\ &\text{s.t. } W = [w^1, \ldots, w^M] \end{aligned} \quad (4)$$

Since minimizing the low rank regularization is a NP-hard problem, for the tractability of optimization, we follow the standard relaxation (Candès and Recht (2009)) and approximate the rank function as the nuclear norm $\|\cdot\|_*$, which is the sum of all the singular values of a matrix. Then Eq.(4) can be reformulated as

$$\begin{aligned} &\min_{\{w^m\}_{m=1}^M} \sum_{m=1}^{M} \frac{1}{2}\|y^m - X^m w^m\|_2^2 + \lambda_1 \|W\|_* + \lambda_2 \sum_{m=1}^{M} \|w^m\|_1 \\ &\text{s.t. } W = [w^1, \ldots, w^M] \end{aligned} \quad (5)$$

To solve the optimization problem Eq. (5), we drive an optimization algorithm to obtain the optimal solution. The detailed derivation can be found in Section (3.2).

3.2. Optimization

Since the objective function in Eq. (5) is composed of a differential function (i.e. square of ℓ_2 norm) and non-differential ones (nuclear norm and ℓ_1 norm), for tractable optimization, we introduce two block of auxiliary variables $R = [r^1, \ldots, r^M]$ and $Z = [z^1, \ldots, z^M]$, and replace the variables in ℓ_1 norm and nuclear norm, which separate the original problem into several subproblems that can be solved more effectively. Accordingly, $R = W$ and $Z = W$ serve as additional constraints for the transformed problem:

$$\begin{aligned} &\min_{\{w^m\}_{m=1}^M} \sum_{m=1}^{M} \frac{1}{2}\|y^m - X^m w^m\|_2^2 + \lambda_1 \|Z\|_* + \lambda_2 \sum_{m=1}^{M} \|r^m\|_1 \\ &\text{s.t. } W = [w^1, \ldots, w^M], Z = W, R = W \end{aligned} \quad (6)$$

Then the augmented Lagrange function $\mathcal{L}$ for Eq. (6) is

$$\begin{aligned} \mathcal{L} = &\sum_{m=1}^{M} \frac{1}{2}\|y^m - X^m w^m\|_2^2 + \lambda_1 \|Z\|_* + \lambda_2 \sum_{m=1}^{M} \|r^m\|_1 \\ &+ \Phi(\Gamma, Z - W) + \Phi(\Lambda, R - W) \end{aligned} \quad (7)$$

where the function $\Phi(\cdot)$ is defined as $\Phi(A, B) = \text{Trace}(A^T B) + \frac{\mu}{2}\|B\|_F^2$, μ is a positive penalty constant, $\Gamma = [\Gamma^1, \ldots, \Gamma^M]$ and $\Lambda = [\Lambda^1, \ldots, \Lambda^M]$ are the Lagrange multipliers. Based on Eq. (7), the solution to Eq. (6) can be obtained by the Alternative Direction Method of Multipliers (ADMM) (Boyd et al. (2011)). We derive the algorithm based on ADMM which iteratively updates three blocks of variables in three subproblems: $\{R, Z\}$-subproblem, $\{W\}$-subproblem, and $\{\Lambda, \Gamma\}$-subproblem.

$\{R, Z\}$-*subproblem*: Keeping other variables fixed, by some mathematical manipulation, solving $\{R, Z\}$-subproblem is equivalent to solving the following problems:

$$\min_R \frac{1}{2}\|R - P\|_F^2 + \frac{\lambda_2}{\mu}\|R\|_1 \quad (8)$$

$$\min_Z \frac{1}{2}\|Z - Q\|_F^2 + \frac{\lambda_1}{\mu}\|Z\|_* \quad (9)$$

where $P = W - \frac{\Lambda}{\mu}$, $Q = W - \frac{\Gamma}{\mu}$. The optimums of problems Eqs. (8) and (9) can be solved by computing $R = \mathcal{S}_{\frac{\lambda_2}{\mu}}(P)$ and $Z = \mathcal{T}_{\frac{\lambda_1}{\mu}}(Q)$ where $\mathcal{S}_{(\cdot)}(\cdot)$ is the soft-thresholding operator such that $\mathcal{S}_\alpha(A)_{m,n} = \text{sign}(A_{m,n}) \cdot \max(|A_{m,n}| - \alpha, 0)$, and $\mathcal{T}_{(\cdot)}(\cdot)$ is the singular-value thresholding operator such that $\mathcal{T}_\beta(B) = U_B \mathcal{S}_\beta(\Sigma_B) V_B^T$ where $U_B \Sigma_B V_B^T$ is the singular value decomposition of B

$\{W\}$-*subproblem*: With other variables fixed, by taking the derivative of Eq. (7) and setting it to be zero, it is equivalent to solve the linear equations:

$$\begin{aligned} &\left[(X^m)^T (X^m) + 2\mu I\right] w^m = (X^m)^T y^m + \mu(z^m + r^m) + \Lambda^m + \Gamma^m \\ &m = 1, \ldots, M \end{aligned} \quad (10)$$

Algorithm 1 Optimization Procedure for Problem (6)

Require: template set $\{X^m\}_{m=1}^M$, target candidate sample $\{y^m\}_{m=1}^M$, regularization parameters λ_1 and λ_2

1: **Initializaiton**: $Z \leftarrow \mathbf{0}$, $R \leftarrow \mathbf{0}$, $\Gamma \leftarrow \mathbf{0}$, $\Lambda \leftarrow \mathbf{0}$, $W \leftarrow \mathbf{0}$, $\mu \leftarrow 10^{-6}$,$Z \leftarrow \mathbf{0}$, $\rho \leftarrow 1.5$, $\epsilon \leftarrow 10^{-5}$, $\mu_{\max} \leftarrow 10^7$
2: **repeat**
3: Updating R and Z via (8) and (9).
4: Updating W via (10).
5: Updating Γ and Λ via (11)
6: $\mu \leftarrow \min(\rho\mu, \mu_{\max})$
7: **until** $\|Z - W\|_\infty < \epsilon\|W\|_\infty \&\& \|R - W\|_\infty < \epsilon\|W\|_\infty$

Ensure: W

$\{\Lambda, \Gamma\}$-*subproblem*: With the updated optimal variables, the lagrange multipliers are updated as follows:

$$\Gamma \leftarrow \Gamma + \mu(Z - W) \tag{11}$$

$$\Lambda \leftarrow \Lambda + \mu(R - W) \tag{12}$$

The derived optimization algorithm iteratively updates the optimal variables and the Lagrangian multipliers until the norm of the primal residual converges to zero (i.e. $\|Z - W\|_\infty < \epsilon\|W\|_\infty$ and $\|R - W\|_\infty < \epsilon\|W\|_\infty$). The optimization problem (6) is a convex optimization problem with linear constraints in which the optimal variables can be separated into two blocks R, Z and W respectively and therefore we can exploit the alternating direction method of multipliers (ADMM) to estimate the global optimal solution. The convergence of ADMM in solving structured optimization problem with linear constraints has been guaranteed by the Theorem 4.1 in (Eckstein and Bertsekas (1992)).

4. Implementation Details

This section mainly introduces some implementation details of the proposed multi-modality tracker which include the target position decision criteria.

4.1. Particle Filtering Framework for Target Position Decision

Our tracking algorithm is performed within a particle filtering framework. Let o_t and s_t denote the observation and state variable at frame t. Provided with an observation variable set up to Frame t, i.e. $O_t == \{o_k | k = 1, \ldots, t\}$, the true posterior $p(s_t^i|O_t)$ is approximated by a set of particles with states $s_t^i, i = 1, \ldots, n$. the tracking result at Frame t can be estimated by maximizing a posteriori:

$$\tilde{s}_t = \arg\max_{s_t^i} p(s_t^i|O_t) \tag{13}$$

Within the particle filtering framework, the posterior probability $p(s_t^i|O_t)$ is recursively estimated as

$$p(s_t|O_T) \propto p(o_t|s_t) \int p(s_t|s_{t-1}) p(s_{t-1}|O_{t-1}) ds_{t-1} \tag{14}$$

where $P(o_t|s_t)$ and $P(s_t|s_{t-1})$ denote the observation model and the state transition model,respectively. To model the target motion across two consecutive frames, we define the $s_t = [a_1, a_2, a_3, a_4, a_5, a_6]^T$ which denote the vertical and horizontal translation, rotation angle, scale, aspect ratio, and skew respectively. The transition model $P(s_t|s_{t-1})$ is model as $P(s_t|s_{t-1}) = N(s_t|s_{t-1}, \Sigma)$ where Σ is a diagonal matrix.With the learned representation of each target candidate in different modalities, the observation likelihood function can be defined as follows:

$$p(o_t^i|s_t^i) \propto \exp(-\sum_{m=1}^{M} \|y_t^{m,i} - X_t^m w_t^{m,i}\|_2^2) \tag{15}$$

where $y_t^{m,i}$ and $w_t^{m,i}$ denote the m-th modality feature of the i-th particle and its sparse coefficients at Frame t, and X_t^m is the object template of m-th modality at Frame t. The right hand side of Eq. (5) is defined based on the total reconstruction error using the object templates of different modalities.

5. Experiments

In this section, we first introduce the setting of the experiments, and then we demonstrate and analyze the comparison experimental results with the other trackers.

5.1. Experimental Setting

We adopt fifteen RGB-infrared video pairs[1] in which video pairs are captured by visible and infrared cameras to evaluate the RGB-infrared tracking performance. The tracked objects in these video pairs encountered some large appearance variations such as occlusion, poor illumination conditions, large scale changes, etc.. Alignment and registration have been performed on these video pairs accurately. Therefore, the tracked targets of each video pair is almost in the same position in each video frame of RGB and infrared modalities. We use nine baseline methods for comparison. They are MEEM (Zhang et al. (2014a)), STRUCK (Hare et al. (2016)), STC (Zhang et al. (2014b)), CN Danelljan et al. (2014), RPT (Li et al. (2017a)), KCF (Henriques et al. (2015)), CT (Zhang et al. (2014c)), MIL Babenko et al. (2011), and JSR (Liu and Sun (2012)) methods. Among these comparison method, only the JSR methods is proposed for RGB-infrared tracking. For the others, they are originally designed for RGB object tracking. Following the setting used in (Li et al. (2016)), the multi-modality version of these trackers can be implemented. Tracking results of these trackers on these RGB-infrared videos can be obtained from (Li et al. (2016)).

The tradeoff parameters λ_1, λ_2 in Eq. (5) are set as 0.01 and 0.001, respectively. Given the image patch in the bounding box, we transform its RGB image to be in grey scale and extract the HOG features Dalal and Triggs (2005) as representation from the RGB modality, and extract the intensity feature from the infrared image as representation from the infrared modality. The cell number of the HOG feature is 4 by 4 while the number of orientation is 9. We follow the setting in Zou et al. (2013) where the image patch inside the bounding box is resized to a

[1] http://hcp.sysu.edu.cn/resources/
http://vcipl-okstate.org/pbvs/bench/index.html

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

	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	JSR	**Proposed Method**
BusScale	0.47	0.45	0.46	0.49	**0.57**	0.52	0.51	0.51	**0.54**	**0.75**
Exposure2	0.32	**0.37**	0.31	0.32	**0.48**	0.3	0.32	0.32	0.35	**0.62**
FastCar2	**0.57**	0.53	0.43	0.48	0.51	0.49	0.5	0.54	**0.56**	**0.6**
FastCarNig	0.46	**0.75**	0.36	0.36	**0.63**	0.41	0.43	0.43	0.38	**0.53**
MinibusNig	0.54	0.55	0.54	0.55	**0.68**	0.55	0.57	**0.59**	0.33	**0.65**
Motorbike	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	0.3	**0.31**	**0.31**	0.3	**0.48**
CarNig	**0.25**	0.21	0.2	0.18	**0.36**	0.19	0.16	**0.25**	0.2	**0.32**
Minibus1	0.53	0.05	0.52	**0.55**	0.06	0.38	**0.56**	0.05	0.53	**0.65**
Motorbike1	**0.68**	0.66	**0.69**	0.57	0.65	0.65	0.56	0.59	**0.69**	**0.68**
BusScale1	0.4	0.41	0.43	0.39	**0.67**	0.44	0.43	0.42	**0.47**	**0.58**
Football	0.65	0.6	**0.73**	0.67	0.56	0.65	**0.69**	0.59	0.62	**0.7**
OccCar-1	0.45	**0.46**	0.43	0.33	**0.68**	0.41	0.45	0.45	0.07	**0.64**
Otcbvs1	0.63	0.69	0.65	**0.73**	**0.72**	0.66	0.66	0.68	**0.79**	0.68
Walking	0.13	0.3	**0.46**	0.36	0.21	**0.5**	0.22	0.05	0.27	**0.4**
RainyCar2	**0.55**	0.46	0.35	0.44	**0.58**	**0.55**	0.4	**0.55**	0.52	0.5
Average	0.46	0.45	0.46	0.45	**0.51**	**0.47**	0.45	0.42	0.44	**0.59**

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

	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	JSR	**Proposed Method**
BusScale	0.48	0.4	0.46	0.44	**0.61**	0.53	0.5	0.51	**0.56**	**0.97**
Exposure2	0.2	**0.26**	0.2	0.2	**0.45**	0.16	0.2	0.2	0.19	**0.99**
FastCar2	**0.55**	0.48	0.35	0.43	0.48	0.5	0.53	**0.55**	**0.57**	**0.83**
FastCarNig	0.31	**0.93**	0.28	0.28	**0.73**	0.26	0.28	0.28	0.39	**0.55**
MinibusNig	0.51	0.49	**0.55**	0.51	**0.92**	0.51	0.54	**0.55**	0.36	**0.91**
Motorbike	**0.14**	**0.16**	**0.14**	0.13	0.13	0.12	**0.14**	**0.14**	0.12	**0.3**
CarNig	0.13	**0.19**	0.13	0.13	**0.21**	0.13	0.13	0.13	**0.17**	**0.17**
Minibus1	**0.59**	0.04	0.54	**0.58**	0.05	0.32	0.54	0.04	0.49	**0.91**
Motorbike1	0.96	0.85	**1**	0.82	0.92	0.85	0.64	0.7	**0.97**	**0.99**
BusScale1	0.33	0.34	0.36	0.27	**0.87**	0.45	0.36	0.36	**0.47**	**0.66**
Football	**0.9**	0.81	**0.96**	0.83	0.64	0.87	**0.97**	0.76	0.76	**0.9**
OccCar-1	0.32	**0.44**	0.27	0.24	**0.89**	0.21	0.32	0.32	0.08	**0.96**
Otcbvs1	0.91	0.87	0.98	**1**	0.98	0.94	0.84	0.82	**1**	**1**
Walking	0.13	0.28	**0.62**	**0.45**	0.06	**0.69**	0.25	0.04	0.29	0.1
RainyCar2	0.55	0.52	0.3	0.43	**0.76**	0.62	0.54	**0.65**	0.62	**0.72**
Average	0.47	0.47	**0.48**	0.45	**0.58**	**0.48**	0.45	0.4	0.47	**0.73**

grey-scale image whose size is 32 by 32. Therefore, the size is the same for every template, which means the feature dimension is the same for every template.

5.2. Experimental Results

5.2.1. Quantitative analysis

We adopt two metrics: VOC overlapping rate and success rate to quantitatively measure the tracking accuracy. The VOC 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 of the ground-truth and the tracker. We consider it as a tracking success if the overlapping rate for the result in a video frame is larger than 0.5. The success rate is defined as the percentage of video frames in which the track success is achieved. The success rate and the overlapping rate of all the compared tracker on these fifteen videos are recorded on Tables 1 and 2. The quantitative results from these two tables show that the proposed tracker performs better than the other nine standard trackers on most videos in terms of overlapping rates and success rates, and the proposed method achieves the best overall performance with the highest mean value in terms of both two metrics. The proposed tracker ranks in top three on fifteen videos in terms of success rates and overlapping rates. The proposed tracker utilizes the sparse and low ranking regularization to combine information from multiple modalities, in which the sparsity regularization helps to adaptively select representative and informative feature templates in each modality to handle different appearance changes, and low rank regularization helps to exploit the correlation of different modalities for more effective modality fusion. As illustrated in Fig. 2, it demonstrates good performance to some large appearance variations, such as occlusion (e.g. *BusScale1*, *Motorbike*), poor illumination conditions (e.g. *FastCarNig*, *Otcbvs1*), thermal crossover (e.g.*Football*, *FastCar2*), etc..

Fig.3 shows the frame-by-frame quantitative comparison of overlapping rates on some videos. we can see that the proposed tracker maintains a high overlapping rate throughout the video, which validates the stability of the proposed tracker. We can see that the proposed method can run throughout the videos under dim environment (e.g. *MinibusNig*), partial occlusion (e.g. *Exposure2*), thermal crossover (e.g. *MotorBike*, *FastCar2*), low resolution (e.g. *Motorbike1*).

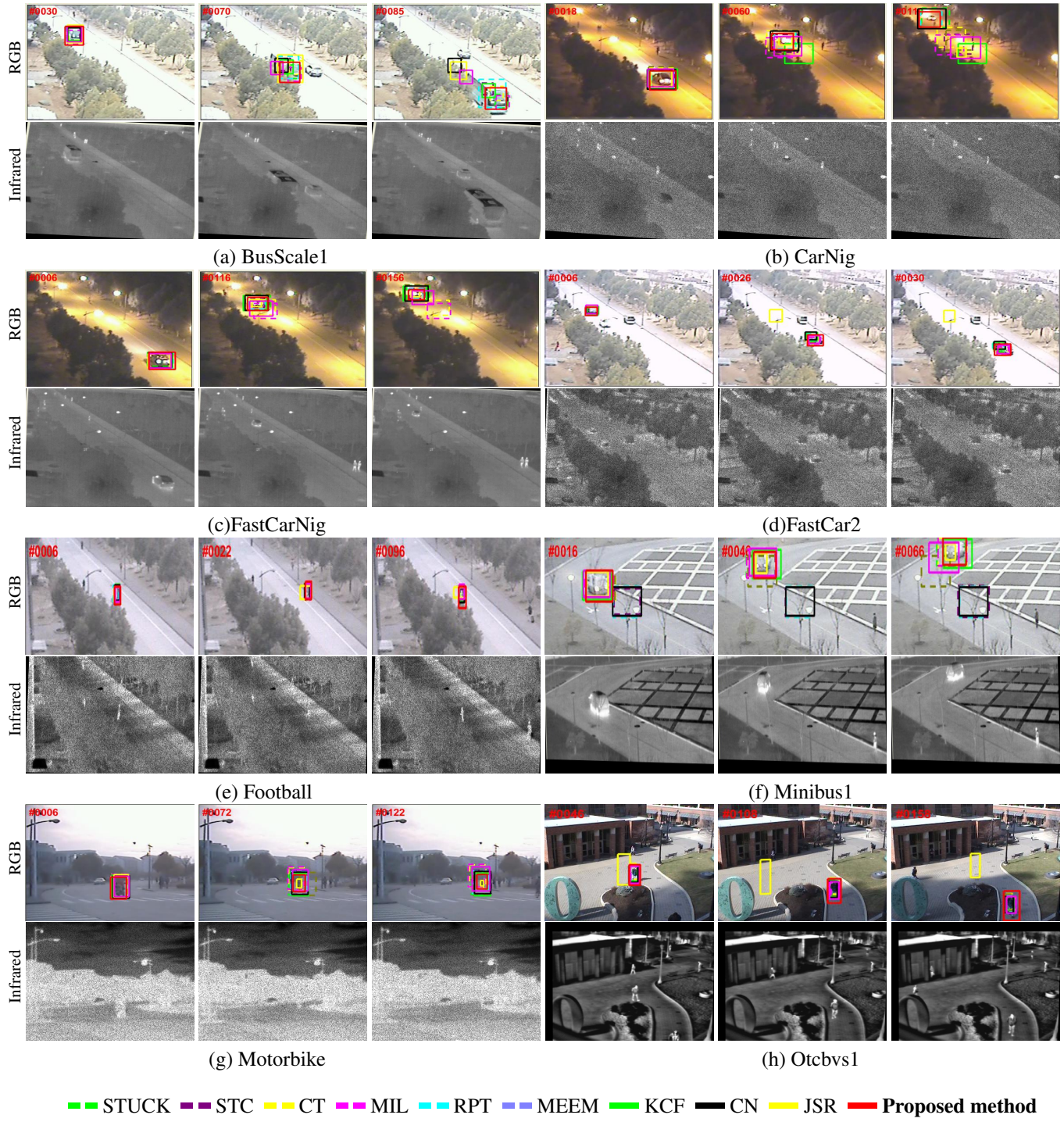

Fig. 2. Qualitative results on some frames of RGB and infrared modality with challenging factors, which includes Thermal crossover (e.g. *Motorbike*, *FastCar2*), scale changes (e.g. *BusScale1*), poor illumination (e.g. *CarNig*), occlusion (e.g. *Football*). For each sub-figure, the top row shows images of RGB modality while the bottom one shows images of infrared modality.

Running time. Because the learning algorithm is performed in an iterative way, the proposed tracker cannot achieve real-time speed and it is about 3 frames/sec.

5.2.2. *Qualitative analysis*

Thermal crossover. Videos such as *Motorbike1* and *FastCar2* undergos thermal crossover which means the infrared modality is not reliable. As shown in Fig. 3(g), the motorbike is hardly seen in the video frames. Because of the proper fusion of the RGB and infrared modality, the proposed tracker can track the motorbike stably. However, methods such as the MIL method which improperly uses feature concatenation for fusion cannot achieve stable performance and have small drift during the tracking.

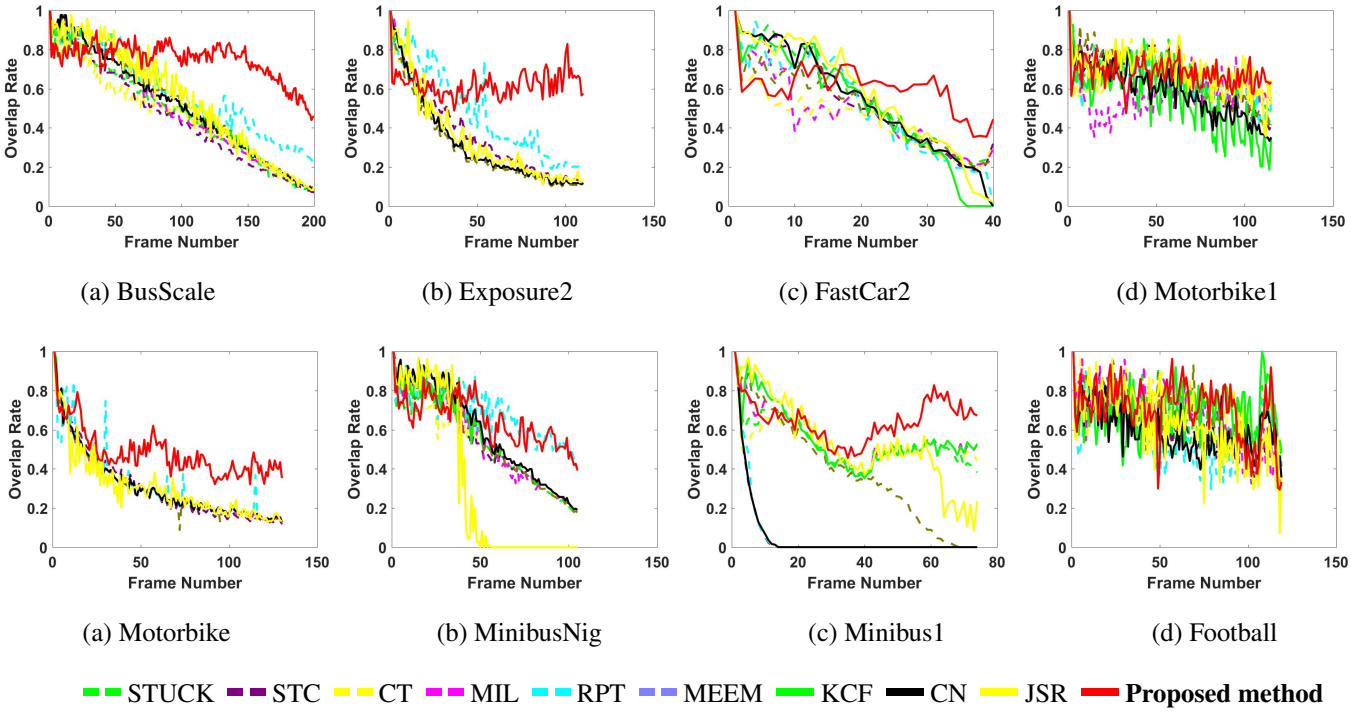

Fig. 3. Quantitative comparison of 10 trackers on 8 challenging videos in terms of overlapping rate. The horizontal axis is the frame index and the vertical axis indicates the overlapping rate.

Poor illumination. The lighting conditions of some testing videos such as *CarNig* is not good. As shown in Fig. 3(b), because of the overexposure at night, the car is ambiguous around the 60th frame in the RGB modality of *CarNig* while it is clearly shown in the infrared modality. By exploit the correlation of the rgb and infrared modalities, the proposed tracker can perform effective fusion of the reliable information from rgb and infrared modalities, which enables them to track the target under poor lighting condition.

Occlusion. The fusion of two modalities also enables the tracker to run stably under occlusion. As shown in the Fig.3(h)where the car is occluded by the trees in the initial stage of the video, the high contrast of between the trees and the cars in infrared modality can facilitate the robustness to the occlusion from the tree. Therefore, the effective integration of the infrared modality help the proposed tracker can run through the occlusion at the beginning of the video. However, some trackers such as the CN methods are distracted by the occlusion, which further illustrate the importance of the integration of the rgb and infrared modalities in dealing with occlusion.

6. Conclusion

In this paper, we designed a modality-correlation-aware sparse representation model for RGB-Infrared object tracking. The proposed model exploits the correlation of different modalities via the low rank regularization and adaptively select representative templates to deal with appearance changes via the sparsity regularization, which makes it more able to perform effective modality fusion and handle large appearance changes. An effective and efficient learning algorithm is also derived to optimize the learning model. Experimental results demonstrate its effectiveness.

Due to some appearance changes, features extracted from some modalities may be corrupted, and not informative for the appearance model. Therefore, one of our future work is to develop some learning model to remove these outlier features. In addition, properly evaluating the reliability of different modalities is also very important. We will also study an effective algorithm to adaptively determine the reliability weight of different modalities.

7. Acknowledgement

This work is partially supported by Hong Kong RGC General Research Fund HKBU 12254316. The work of H. Zhou was supported in part by UK EPSRC under Grant EP/N508664/1, Grant EP/R007187/1, and Grant EP/N011074/1 and in part by the Royal Society-Newton Advanced Fellowship under Grant NA160342.

References

Babenko, B., Yang, M., Belongie, S., 2011. Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1619–1632.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning 3.

Bunyak, F., Palaniappan, K., Nath, S.K., Seetharaman, G., 2007. Geodesic active contour based fusion of visible and infrared video for persistent object tracking, in: Proc. WACV.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Foundations of Computational Mathematics 9, 717–772.

Chang, X., Ma, Z., Yang, Y., Zeng, Z., Hauptmann, A.G., 2017a. Bi-level semantic representation analysis for multimedia event detection. IEEE Trans. Cybernetics 47, 1180–1197.

Chang, X., Yang, Y., 2017. Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Trans. Neural Netw. Learning Syst. 28, 2294–2305.

Chang, X., Yu, Y., Yang, Y., Xing, E.P., 2017b. Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1617–1632.

Conaire, C.Ó., O'Connor, N.E., Smeaton, A.F., 2008. Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. Mach. Vis. Appl. 19, 483–494.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Proc. CVPR, pp. 886–893.

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

Ding, G., Chen, W., Zhao, S., Han, J., Liu, Q., 2018. Real-time scalable visual tracking via quadrangle kernelized correlation filters. IEEE Trans. Intelligent Transportation Systems 19, 140–150.

Eckstein, J., Bertsekas, D.P., 1992. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. 55, 293–318.

Guo, Y., Ding, G., Liu, L., Han, J., Shao, L., 2017. Learning to hash with optimized anchor embedding for scalable retrieval. IEEE Trans. Image Processing 26, 1344–1354.

Han, J., Pauwels, E.J., de Zeeuw, P.M., de With, P.H.N., 2012. Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment. IEEE Trans. Consumer Electronics 58, 255–263.

Han, J., Shao, L., Xu, D., Shotton, J., 2013. Enhanced computer vision with microsoft kinect sensor: A review. IEEE Trans. Cybernetics 43, 1318–1334.

Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M., Hicks, S.L., Torr, P.H.S., 2016. Struck: Structured output tracking with kernels. IEEE Trans. Pattern Anal. Mach. Intell. 38, 2096–2109.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. 37, 583–596.

Hong, Z., Mei, X., Prokhorov, D., Tao, D., 2013. Tracking via robust multi-task multi-view joint sparse representation, in: Proc. ICCV, pp. 649–656.

Hou, R., Chen, C., Shah, M., 2017. Tube convolutional neural network (T-CNN) for action detection in videos, in: Proc. ICCV, pp. 5823–5832.

Hu, J.F., Zheng, W.S., Lai, J., Zhang, J., 2015a. Jointly learning heterogeneous features for rgb-d activity recognition, in: Proc. CVPR, pp. 5344–5352.

Hu, W., Li, W., Zhang, X., Maybank, S., 2015b. Single and multiple object tracking using a multi-feature joint sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 37, 816–833.

Lan, X., Ma, A., Yuen, P., Chellappa, R., 2015. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. IEEE Trans. Image Process. 24, 5826–5841.

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

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

Lan, X., Yuen, P.C., Chellappa, R., 2017. Robust mil-based feature template learning for object tracking, in: Proc. AAAI, pp. 4118–4125.

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

Lan, X., Zhang, S., Yuen, P.C., Chellappa, R., 2018b. Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker. IEEE Trans. Image Processing 27, 2022–2037.

Leykin, A., Hammoud, R.I., 2010. Pedestrian tracking by fusion of thermal-visible surveillance videos. Mach. Vis. Appl. 21, 587–595.

Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L., 2016. Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Trans. Image Processing 25, 5743–5756.

Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., Hengel, A.v.d., 2013. A survey of appearance models in visual object tracking. ACM Trans. Intell. Syst. Technol. 4, 58:1–58:48.

Li, Y., Zhu, J., Hoi, S.C., 2017a. Reliable patch trackers: Robust visual tracking by exploiting reliable patches, in: Proc. CVPR, pp. 353–361.

Li, Z., Nie, F., Chang, X., Yang, Y., 2017b. Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis. IEEE Trans. Knowl. Data Eng. 29, 2100–2110.

Liu, H., Sun, F., 2012. Fusion tracking in color and infrared images using joint sparse representation. Sci. China Inf. Sci. 55, 590–599.

Liu, R., Lan, X., Yuen, P.C., Feng, G., 2016. Robust visual tracking using dynamic feature weighting based on multiple dictionary learning, in: Proc. EUSIPCO, pp. 2166–2170.

Liu, S.Q., Lan, X., Yuen, P.C., 2018. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection, in: Proc. ECCV, pp. 558–573.

Luo, M., Chang, X., Li, Z., Nie, L., Hauptmann, A.G., Zheng, Q., 2017. Simple to complex cross-modal learning to rank. Computer Vision and Image Understanding 163, 67–77.

Ma, B., Hu, H., Shen, J., Liu, Y., Shao, L., 2016. Generalized pooling for robust object tracking. IEEE Trans. Image Processing 25, 4199–4208.

Ma, B., Shen, J., Liu, Y., Hu, H., Shao, L., Li, X., 2015. Visual tracking using strong classifier and structural local sparse descriptors. IEEE Trans. Multimedia 17, 1818–1828.

Ma, Z., Chang, X., Xu, Z., Sebe, N., Hauptmann, A.G., 2018. Joint attributes and event analysis for multimedia event detection. IEEE Trans. Neural Netw. Learning Syst. 29, 2921–2930.

Mei, X., Ling, H., 2011. Robust visual tracking and vehicle classification via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 33, 2259–2272.

Pang, M., Wang, B., Cheung, Y., Lin, C., 2017. Discriminant manifold learning via sparse coding for robust feature extraction. IEEE Access 5, 13978–13991.

Peng, X., Lu, C., Zhang, Y., Tang, H., 2018. Connections between nuclear norm and frobenius norm based representation. IEEE Trans. Neural Netw. Learn. Syst. 29, 218–224.

Peng, X., Lu, J., Yi, Z., Rui, Y., 2017. Automatic subspace learning via principal coefficients embedding. IEEE Trans. Cybern. 47, 3583–3596.

Peng, X., Tang, H., Zhang, L., Yi, Z., Xiao, S., 2016. A unified framework for representation-based subspace clustering of out-of-sample and large-scale data. IEEE Trans. Neural Netw. Learn. Syst. 27, 2499–2512.

Salti, S., Cavallaro, A., di Stefano, L., 2012. Adaptive appearance modeling for video tracking: Survey and evaluation. IEEE Trans. Image Process. 21, 4334–4348.

Shao, R., Lan, X., Yuen, P.C., 2018. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. IEEE Trans. Inf. Forensics Security DOI:10.1109/TIFS.2018.2868230.

Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M., 2014. Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1442–1468.

Song, Y., Ma, C., Gong, L., Zhang, J., Lau, R.W.H., Yang, M., 2017. CREST: convolutional residual learning for visual tracking, in: Proc. ICCV, pp. 2574–2583.

Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Rynson, L., Yang, M.H., 2018. Vital: Visual tracking via adversarial learning, in: IEEE Conference on Computer Vision and Pattern Recognition.

Sui, Y., Tang, Y., Zhang, L., 2015. Discriminative low-rank tracking, in: Proc. ICCV, pp. 3002–3010.

Wang, A., Cai, J., Lu, J., Cham, T., 2016a. Modality and component aware feature fusion for RGB-D scene classification, in: Proc. CVPR, pp. 5995–6004.

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

Wang, D., Lu, H., Bo, C., 2015b. Visual tracking via weighted local cosine similarity. IEEE Trans. Cybernetics 45, 1838–1850.

Wang, Q., Fang, J., Yuan, Y., 2014. Multi-cue based tracking. Neurocomputing 131, 227–236.

Wang, S., Chang, X., Li, X., Long, G., Yao, L., Sheng, Q.Z., 2016b. Diagnosis code assignment using sparsity-based disease correlation embedding. IEEE Trans. Knowl. Data Eng. 28, 3191–3202.

Wang, S., Li, X., Chang, X., Yao, L., Sheng, Q.Z., Long, G., 2017. Learning multiple diagnosis codes for ICU patients with local disease correlation mining. TKDD 11, 31:1–31:21.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S., 2010. Sparse representation for computer vision and pattern recognition. Proceedings of

the IEEE 98, 1031–1044.
Wu, Y., Blasch, E., Chen, G., Bai, L., Ling, H., 2011. Multiple source data fusion via sparse representation for robust visual tracking, in: Proc. Int. Conf. Inf. Fusion,, pp. 1–8.
Wu, Y., Lim, J., Yang, M., 2015. Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. 37, 1834–1848.
Yang, B., Ma, A.J., Yuen, P.C., 2018a. Body parts synthesis for cross-quality pose estimation. IEEE Transactions on Circuits and Systems for Video Technology DOI=10.1109/TCSVT.2017.2789224.
Yang, B., Ma, A.J., Yuen, P.C., 2018b. Learning domain-shared group-sparse representation for unsupervised domain adaptation. Pattern Recognit. 81, 615–632.
Yang, L., Chen, C., Wang, H., Zhang, B., Han, J., 2016. Adaptive multi-class correlation filters, in: Proc. PCM, pp. 680–688.
Ye, M., Lan, X., Yuen, P.C., 2018a. Robust anchor embedding for unsupervised video person re-identification in the wild, in: Proc. ECCV, pp. 2651–2664.
Ye, M., Liang, C., Wang, Z., Leng, Q., Chen, J., 2015. Ranking optimization for person re-identification via similarity and dissimilarity, in: ACM MM, pp. 1239–1242.
Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., Chen, J., Hu, R., 2016. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Trans. Multimedia 18, 2553–2566.
Ye, M., Ma, A.J., Zheng, L., Li, J., Yuen, P.C., 2017. Dynamic label graph matching for unsupervised video re-identification, in: ICCV, pp. 5142–5150.
Ye, M., Wang, Z., Lan, X., Yuen, P.C., 2018b. Visible thermal person re-identification via dual-constrained top-ranking, in: Proc. IJCAI, pp. 1092–1099.
Yuan, X.T., Liu, X., Yan, S., 2010. Visual classification with multitask joint sparse representation, in: Proc. CVPR, pp. 3493–3500.
Yuan, Y., Fang, J., Wang, Q., 2014. Robust superpixel tracking via depth fusion. IEEE Trans. Circuits Syst. Video Techn. 24, 15–26.
Zhang, B., Gu, J., Chen, C., Han, J., Su, X., Cao, X., Liu, J., 2018a. One-two-one networks for compression artifacts reduction in remote sensing. Isprs Journal of Photogrammetry and Remote Sensing .
Zhang, B., Li, Z., Cao, X., Ye, Q., Chen, C., Shen, L., Perina, A., Ji, R., 2017a. Output constraint transfer for kernelized correlation filter in tracking. IEEE Trans. Systems, Man, and Cybernetics: Systems 47, 693–703.
Zhang, B., Luan, S., Chen, C., Han, J., Wang, W., Perina, A., Shao, L., 2018b. Latent constrained correlation filter. IEEE Trans. Image Processing 27, 1038–1048.
Zhang, B., Perina, A., Li, Z., Murino, V., Liu, J., Ji, R., 2016. Bounding multiple gaussians uncertainty with application to object tracking. Int J. Comput. Vis. 118, 364–379.
Zhang, B., Yang, Y., Chen, C., Yang, L., Han, J., Shao, L., 2017b. Action recognition using 3d histograms of texture and a multi-class boosting classifier. IEEE Trans. Image Process. 26, 4648–4660.
Zhang, J., Ma, S., Sclaroff, S., 2014a. Meem: Robust tracking via multiple experts using entropy minimization, in: Proc. ECCV, pp. 188–203.
Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.H., 2014b. Fast visual tracking via dense spatio-temporal context learning, in: Proc. ECCV, pp. 127–141.
Zhang, K., Zhang, L., Yang, M.H., 2014c. Fast compressive tracking. IEEE Trans. Pattern Anal. Mach. Intell. 36, 2002–2015.
Zhang, S., Lan, X., Qi, Y., Yuen, P.C., 2017c. Robust visual tracking via basis matching. IEEE Trans. Circuits Syst. Video Techn. 27, 421–430.
Zhang, S., Lan, X., Yao, H., Zhou, H., Tao, D., Li, X., 2017d. A biologically inspired appearance model for robust visual tracking. IEEE Trans. Neural Netw. Learn. Syst. 28, 2357–2370.
Zhang, S., Qi, Y., Jiang, F., Lan, X., Yuen, P.C., Zhou, H., 2018c. Point-to-set distance metric learning on deep representations for visual tracking. IEEE Trans. Intelligent Transportation Systems 19, 187–198.
Zhang, S., Yao, H., Sun, X., Lu, X., 2013a. Sparse coding based visual tracking: Review and experimental comparison. Pattern Recognit. 46, 1772–1788.
Zhang, S., Yao, H., Zhou, H., Sun, X., Liu, S., 2013b. Robust visual tracking based on online learning sparse representation. Neurocomputing 100, 31–40.
Zhang, S., Zhou, H., Jiang, F., Li, X., 2015a. Robust visual tracking using structurally random projection and weighted least squares. IEEE Trans. Circuits Syst. Video Techn. 25, 1749–1760.
Zhang, T., Ghanem, B., Liu, S., Ahuja, N., 2013c. Robust visual tracking via structured multi-task sparse learning. Int. J. Comput. Vis. 101, 367–383.
Zhang, T., Liu, S., Xu, C., Yan, S., Ghanem, B., Ahuja, N., Yang, M.H., 2015b. Structural sparse tracking, in: Proc. CVPR, pp. 150–158.
Zhong, B., Yao, H., Chen, S., Ji, R., Chin, T., Wang, H., 2014. Visual tracking via weakly supervised learning from multiple imperfect oracles. Pattern Recognit. 47, 1395–1410.
Zhou, J.T., Tsang, I.W., Ho, S.s., Müller, K.R., 2018a. N-ary decomposition for multi-class classification. Machine Learning .
Zhou, J.T., Zhao, H., Peng, X., Fang, M., Qin, Z., Goh, R.S.M., 2018b. Transfer hashing: From shallow to deep. IEEE Trans. Neural Netw. Learn. Syst. DOI: 10.1109/TNNLS.2018.2827036.
Zou, W.W., Yuen, P.C., Chellappa, R., 2013. Low-resolution face tracker robust to illumination variations. IEEE Trans. Image Process. 22, 1726–1739.