The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Robust Collaborative Discriminative Learning for RGB-Infrared Tracking

Xiangyuan Lan,[†] Mang Ye,[†] Shengping Zhang,[‡] Pong C. Yuen[†]*
[†]Department of Computer Science, Hong Kong Baptist University
[‡]School of Computer Science and Technology, Harbin Institute of Technology
xiangyuanlan@life.hkbu.edu.hk, mangye@comp.hkbu.edu.hk, s.zhang@hit.edu.cn, pcyuen@comp.hkbu.edu.hk

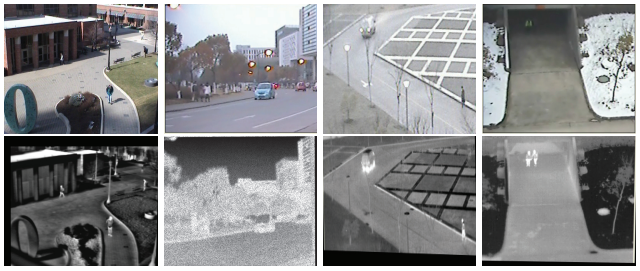
Abstract

Tracking target of interests is an important step for motion perception in intelligent video surveillance systems. While most recently developed tracking algorithms are grounded in RGB image sequences, it should be noted that information from RGB modality is not always reliable (e.g. in a dark environment with poor lighting condition), which urges the need to integrate information from infrared modality for effective tracking because of the insensitivity to illumination condition of infrared thermal camera. However, several issues encountered during the tracking process limit the fusing performance of these heterogeneous modalities: 1) the cross-modality discrepancy of visual and motion characteristics, 2) the uncertainty of degree of reliability in different modalities, and 3) large target appearance variations and background distractions within each modality. To address these issues, this paper proposes a novel and optimal discriminative learning framework for multi-modality tracking. In particular, the proposed discriminative learning framework is able to: 1) jointly eliminate outlier samples caused by large variations and learn discriminability-consistent features from heterogeneous modalities, and 2) collaboratively perform modality reliability measurement and target-background separation. Extensive experiments on RGB-infrared image sequences demonstrate the effectiveness of the proposed method.

1 Introduction

As a key component for intelligent motion perception in intelligent video surveillance systems (Ye et al. 2015; 2016; 2017; Wang et al. 2016b), tracking target of interests has received great research interests and significant progress has been achieved recently (Zhang et al. 2013a; 2013b; 2015; 2017a; 2017b). Most recently developed tracking algorithms are grounded on RGB image sequences captured by visible spectrum cameras, and they construct the appearance model using visual cues from RGB information (Liu et al. 2016; Lan, Yuen, and Chellappa 2017), which may disable them to be applied in some practical scenarios, especially when information from RGB imaging is not reliable (e.g.in a dark environment with poor lighting conditions).

With the development of multispectral imaging techniques, more and more vision systems of robotics and video surveillance are equipped with dual cameras, including one RGB camera and one thermal camera. Infrared thermal cameras can capture the infrared radiation of a subject for imaging, and they are more effective than visible spectrum cameras under poor lighting conditions. Therefore, to develop a robust object tracking system for practical applications, it is very necessary to integrate information from infrared modality with that of RGB modality for effective tracking.



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

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

However, several key issues encountered during tracking process may limit the modality fusing performance, which should be addressed for robust multi-modality tracking. Firstly, images from RGB and infrared modalities are intrinsically distinct in their visual characteristic (e.g. intensity, texture), as shown in Fig. 1(a). Such cross-modality discrepancy may lead to significant difference between the statistical properties of features from different modalities even if the features represent the same subject. Therefore, traditional homogeneous feature fusion methods (e.g. concatenation (Wu et al. 2011), multiple kernel learning (Xu, Wang, and Lu 2012), etc.), which do not explicitly consider the discrepancy issue, is unsuitable for multi-modality tracking. As such, bridging the gap between heterogeneous modalities during modality fusion process is essential. In addition, not all modalities are reliable all the time, and reliability of different modalities are erratically changed under different scenarios. As shown in Fig. 1(b), the blue car can be differentiated from the background based on color information of RGB modality while it is ambiguous in infrared modal-

*Corresponding author

ity because of the 'thermal crossover' issue[1]. Contrarily, in Fig. 1(c), the vehicle of infrared modality is easier to distinguished than the RGB-modality. Therefore, dynamically and properly determining the reliability of different modalities is required for effective modality fusion. Moreover, large appearance changes with different background distraction within each modality would be encountered during tracking process. As illustrated in Fig. 1(d), the target in green is difficult to be distinguished in RGB modality, while he is clearly shown in infrared image. However, another pedestrian with similar appearance which is only shown in infrared image is close to the target as a background distraction. Therefore, how to exploit the complementarity of different modalities to deal with appearance changes and background distraction is another issue to be addressed.

Although several RGB-infrared tracking algorithms have been developed, most of them may not effectively handle either or all aforementioned issues. One kind of approaches regard the RGB and infrared modalities as different feature channels and exploit some homogeneous feature fusion techniques such as feature concatenation (Wu et al. 2011), joint sparsity regularization (Liu and Sun 2012; Li et al. 2016), sum rule (Leykin and Hammoud 2010) to combine multiple modalities for tracking. Such kind of approaches do not explicitly consider the cross-modality discrepancy issue, which may limit the fusion performance. Another kind of approaches such as (Conaire, O'Connor, and Smeaton 2008) perform tracking on each modality independently and then fuse the results from each modality to determine the target position. Although the modality fusion is performed on tracker level, the complementarity of different modalities is not fully exploited during the tracking process, and the reliability of different modalities are not properly evaluated for fusion.

To address aforementioned issues, we proposes a novel and optimal discriminative learning framework for RGB-infrared object tracking. The proposed learning framework is capable of adaptively and collaboratively performing two discriminative learning tasks: learning classifiers of each modality for target/background separation and the reliability weights of different modalities for modality fusion. Through the way of collaborative learning, the classifier learning from each modality provides discriminability measurement of each modality to the process of reliability weight determination, while reliability weight learning offers importance weight for classifier fusion which ensures more discriminative modalities provide higher impact in target/background separation. In addition, to alleviate the cross-modality discrepancy issue and bridge the gap among different heterogeneous modalities, a discriminability-consistent regularization is imposed on the learning model which enforces all the learned heterogeneous feature space share consistent discriminability for more effective modality fusion. To deal with contaminated samples caused by large appearance variations (e.g. occlusion, illumination), a feature decontamination scheme is further incorporated into the learning framework, which produces uncontaminated feature to facilitate reliable classifier learning. Moreover, an efficient iterative optimization algorithm is derived to learn the optimal model parameters which guarantee the optimality of the proposed learning framework.

The contributions of this paper are listed as follows:

- A novel discriminative learning framework is proposed to adaptively and collaboratively learn classifiers and reliability weights of different modalities for RGB-infrared tracking.
- A new feature learning scheme is incorporated the learning framework to learn dsicriminability-consistent uncontaminated features from heterogeneous modalities.
- An efficient optimization algorithm is derived to solve the learning model.

2 Related Work

RGB-Infrared Object Tracking Several algorithms have been proposed for RGB-Infrared Tracking. (Bunyak et al. 2007) developed RGB-thermal moving object segmentation and tracking based on level set framework. (Conaire, O'Connor, and Smeaton 2008) propose a RGB-thermal tracking framework in which multiple spatiogram trackers are run on each modality and the results from different modalities are fused for final decision. In (Leykin and Hammoud 2010), a pedestrian tracker is developed based on background model in which the probabilistic background map is aggregated from the RGB and infrared maps using sum rule. To more effectively deal with outliers introduced by large appearance variations, several sparse representation-based RGB-Infrared trackers have been proposed in which modality fusion is performed by feature concatenation (Wu et al. 2011), group sparsity regularization (Liu and Sun 2012; Li et al. 2016). These methods do not explicitly consider the cross-modality discrepancy issue, which may limit the fusion performance.

Multi-Modality Classification and Recognition To improve the performance of classification and recognition, many algorithms have been developed to exploited multiple sources of data from heterogeneous modalities (Han et al. 2012; 2013; Wang, Fang, and Yuan 2014; Yuan, Fang, and Wang 2014). For example, (Hu et al. 2015) proposed to jointly learn heterogeneous features from RGB-D modalities by mining their shared and modality-specific structures. (Wang et al. 2016a) proposed to learn and fuse modality and component aware features for RGB-D scene classification. However,these algorithms do not consider the cases when some features are contaminated. Additionally, some of them may require large amount of off-line training data. All these issues make it difficult to employ these algorithms into online multi-modality tracking.

3 Proposed Model

3.1 Robust Collaborative Discriminative Learning

Jointly learning heterogeneous features and classifiers with discriminabilty-consistency constraint Let $Y^m =$

[1]means the target has similar temperature with background.

$[Y_1^m, \ldots, Y_{N_0}^m, Y_{N_0+1}^m, \ldots, Y_N^m]$ denote the recently obtained target and background samples with corresponding labels $L = [L_1, \ldots, L_{N_0+1}, L_{N_0+1}, \ldots, L_N^m]^T$, $m = 1, \ldots, M$ denote the index of modality, and $1, \ldots, N_0$ denote indexes of target samples while $N_0 + 1, \ldots, N$ denote the indexes of background samples. The first step of the learning framework is to learn multiple classifiers from these training samples of different modalities for target/background separation, i.e.

$$L_n = f^m(Y_n^m), m = 1, \ldots, M; n = 1, \ldots, N \quad (1)$$

where $f^m(\cdot)$ denote the classifier of the m-th modality which would be used to predict the label of a given sample in m-th modality. Once the classifier of each modality is obtained, classifier fusion can be performed to determine the target position. Traditional discriminative learning algorithms (e.g. SVM (Hare et al. 2016), correlation filter (Henriques et al. 2015), etc) may not be applicable to determine the classifiers f^m of each modality because of the following issues. First, large appearance variation in each modality (e.g. large illumination changes) during the tracking process usually introduce some contamination into the collected samples. Discriminability of the tracking model may be degraded if the classifiers are learned from these contaminated training samples. Second, cross-modality discrepancy may lead to inconsistent discriminability among heterogenous feature space (Zhang, Patel, and Chellappa 2017). Fusing heterogenous modality without considering discrepancy issue may not achieve good fusion performance.

To deal with these two problems, inspired by robust joint discriminative feature learning (Lan, Zhang, and Yuen 2016), we propose to jointly learn heterogeneous features and classifiers for multi-modality tracking under discriminabilty-consistency constraint where feature learning is performed to eliminate contaminated features from large appearance changes and discriminabilty-consistency constraint is imposed to reduce the cross-modality decrepancy. Let $X^m = [X_1^m, \ldots, X_N^m]$ denote the learned uncontaminated features of Y^m in m-th modality, E^m denote the removed outliers from contaminated features of m-th modality, $[(w^m)^T, b^m]^T \in \mathbb{R}^{d^m+1}$ denote the classifier parameters, and $[A, B]_+$ denote the maximum operator for two numbers, i.e. $[A, B]_+ = \max(A, B)$. Then the uncontaminated features and the classifiers can be learned by solving the following optimization problem:

$$\begin{aligned}
&\min_{\Omega} \alpha_1 g_1(\{X^m, L, w^m, b^m\}) + \frac{\lambda_2}{2} \sum_{m=1}^{M} \|w^m\|_2^2 \\
&+ \alpha_2 g_2(\{X^m, E^m\}) + \alpha_3 g_3(\{X^m, L^0, w^m, b^m\}) \quad (2) \\
&\text{s.t. } Y^m = X^m + E^m, m = 1, \ldots, M;\ g_1(\{X^m, L, w^m, b^m\}) \\
&= \sum_{m=1}^{M} \sum_{n=1}^{N} [0, 1 - L_n((X_n^m)^T w^m + b^m)]_+ \\
&g_2(\{X^m, E^m\}) = \sum_{m=1}^{M} \text{rank}\{X^m\} + \lambda_1 \|E^m\|_1 \\
&g_3(\{X^m, L^0, w^m, b^m\}) = \sum_{m=1}^{M} \|(X^m)^T w^m + \mathbf{1} b^m - L^*\|_2^2
\end{aligned}$$

where $\Omega = \{X^m, E^m, w^m, b^m, L^0\}$ denote the set of parameters, $\mathbf{1} \in \mathbb{R}^N$ is the all-one vector, $\alpha_1, \alpha_2, \lambda_1, \lambda_2$ control the tradeoff between different terms and $L_n = +1(-1)$ means the n-th sample belong to the class of target (background). The objective function in (2) consists of three major components: $g_1(\cdot)$, $g_2(\cdot)$ and $g_3(\cdot)$. In the following, we discuss these components in detail one by one.

• $g_2(\{X^m, E^m\})$: this component intends to separate out outliers E^m and facilitate uncontaminated feature learning of X^m via low rank and sparse modeling. Since target samples of different modalities in recent frames are temporally correlative and the target/background samples in the same frame of each modality owns some similar characteristic (e.g. lighting condition), this component exploit such kinds of temporal and spatial correlation to separate out the outliers and mine the latent feature space for feature representation. The sparsity regularization is imposed on the contaminated feature to model the outliers while the rank minimization aims to reveal the shared intrinsic subspace among tracking samples from different modality. Since rank minimization problem is NP-hard problem, we relax the problem as a nuclear norm $\|\cdot\|_*$ minimization problem when deriving the optimization algorithm for this problem.

• $g_1(\{X^m, L, w^m, b^m\})$: this component aims to minimize the prediction losses of different modalities based on the learned features and the classifier parameters. To fully unleash the discriminative power of the tracking model and ensure that the learned features of target and background samples in different modalities can be linearly separated as well as possible, we employ the margin maximization principle and adopt the sum of hinge loss functions for optimization. By jointly optimizing $g_1(\cdot)$ and $g_2(\cdot)$, uncontaminated features could be provided for reliable classifier learning in each modality by removing outliers while classifier learning enhance discriminability of the learned feature. Therefore, this joint learning strategy enables the feature learning and classifier training to benefit from each other, which enhance the representation power and discriminability of the tracking mdoel.

• $g_3(\{X^m, L^0, w^m, b^m\})$: this component aims to alleviate the cross-modality discrepancy. Here $L^0 = [L_1^0, \ldots, L_N^0]^T$ denotes the consensus vector of the classification scores from different modalities, and it encodes the consistent discriminative information of different modalities. In order to constrain the heterogeneous feature space share some consistent discriminability, this regularization term enforces the classification score of each sample of different modalities to be similar and close to the consensus. Considering that consistency is related to the concept of agreement while complementarity can be reflected by disagreement (Liu et al. 2015), we do not enforce the scores of different modalities to be the same. Instead, we adopt the soft regularization strategy and allow some small disagreement among different modalities in order to exploit their complementarity in their discriminability.

Large-margin reliability weight learning Since not all modalities are reliable all the time under different scenarios during the tracking process, the reliability of different modalities should be adaptively evaluated. A straightforward solution to incorporate the reliability weights is to associate the hinge loss of each modality in $g_1(\cdot)$ with a

reliability weight so that $g_1(\cdot)$ is formulated by a linear weighted sum of the prediction loss of different modalities. However, such strategy may not be applicable. This is because if the value of some hinge loss function is zero (e.g. $L_n^{m'}((X_n^{m'})^T w^{m'} + b^{m'}) > 1$ for $n = 1, \ldots, N$ in m'-th modality), it is intractable to learn the associated weight. Therefore, it is not appropriate to jointly optimizing the reliability weight in the learning framework (2), and thereby we propose another new model to learn the reliability weight adaptively for modality fusion. After obtaining the classifier parameters and learned features of each modality $\{X^m, w^m, b^m\}$ by solving (2), the classification sores of the training sample using the learned classifiers can be predicted as $S^m = (X^m)^T w^m + \mathbf{1}b^m,\ m = 1, \ldots, M$, where $S^m = [S_1^m, \ldots, S_N^m]$ and S_n^m is the classification score of the n-th sample in m-th modality. Let $\beta = [\beta^1, \ldots, \beta^M]$ and β^m denote the reliability weight of the m-th modality. Based on LPBoost (Demiriz, Bennett, and Shawe-Taylor 2002), the reliability weights of different modalities can be learned under max-margin principle by solving the following problem:

$$\begin{aligned} &\min_{p,\{\beta^m\}} -p + C_1 \sum_{n=1}^{N} [0, p - L_n \sum_{m=1}^{M} \beta^m S_n^m]_+ + C_2 \|\beta - \beta_0\|_2^2 \\ &\text{s.t. } \sum_{m=1}^{M} \beta^m = 1, \beta^m \geq 0, m = 1, \ldots, M. \end{aligned} \tag{3}$$

where C_1, C_2 are the tradeoff parameters, β_0 is the reliability weight in previous video frame, and p is margin parameter to be learned. The objective function in (3) intends to maximize the margin parameter and imposes that the constraint $L^n \sum_{m=1}^{M} \beta^m S_n^m > p,\ n = 1, \ldots, N$ are satisfied as well as possible, which ensures that the incorporated reliability weight would facilitate the separation between target and background by a large margin as far as possible. In addition, based on the intuition that the target share more similarity with the tracking results in recent frames, we further exploit temporal consistency to determine the reliability weight based on the weights learned from previous frame.

Putting them all together. Based on all above derivation, the proposed learning framework can be summarized as follows:

$$\begin{aligned} &\min_{p,\{\beta^m\}} -p + C_1 \sum_{n=1}^{N} [0, p - L^n \sum_{m=1}^{M} \beta^m S_n^m]_+ + C_2 \|\beta - \beta_0\|_2^2 \\ &\text{s.t. } \sum_{m=1}^{M} \beta^m = 1, \beta^m \geq 0, m = 1, \ldots, M, \\ &\{w^m, b^m, X^m, E^m\} = \arg\min \lambda_2 \|w^2\|_2^2 + \alpha_2 g_2(\{X^m, E^m\}) \\ &+ \alpha_1 g_1(\{X^m, L, w^m, b^m\}) + \alpha_3 g_3(\{X^m, L^0, w^m, b^m\}) \\ &Y^m = X^m + E^m, m = 1, \ldots, M \end{aligned} \tag{4}$$

The learning framework collaboratively performs feature learning, classifier learning, and reliability weight determination of multiple modalities within a unified optimal framework, which enables these three learning tasks to benefit from each other and achieve a better performance. The optimization algorithm for solving (4) will present in the following section.

3.2 Optimization

To construct the learning model in (4), two optimization problems, i.e. problem (2) and (3) need to be solved. For problem (3), it is equivalent to solve the following problem by introducing some slack variables $\{\xi_n\}$:

$$\begin{aligned} &\min_{p,\{\beta^m\},\{\xi_n\}} -p + C_1 \sum_{n=1}^{N} \xi_n + C_2 \|\beta - \beta_0\|_2^2 \\ &\text{s.t. } \sum_{m=1}^{M} \beta^m = 1, \beta^m \geq 0, m = 1, \ldots, M \\ &L_n \sum_{m=1}^{M} \beta^m S_n^m > p - \xi_n, \xi_n \geq 0, n = 1, \cdots, N \end{aligned} \tag{5}$$

This is a quadratic programming with linear constraint, which can be solved by standard optimization toolbox. Now we focus on how to solve problem (2).

The objective function in (2) involves three non-smooth functions, which are hinge loss function, regularization function from nuclear norm and ℓ_1 norm. Let $u_n^m = 1 - L_n((X_n^m)^T w^m + b^m)$. For the sake of simplicity and efficiency, according to (Nesterov 2005; Xu, Tao, and Xu 2015), the hinge loss function can be approximated by its smooth version with smooth parameter $\sigma > 0$, and the smooth version with respect to w^m and X_n^m, denoted as $h_\sigma^{(1)}$ and $h_\sigma^{(2)}$, can be defined as follows:

$$h_\sigma^{(1)}(\Omega) = \begin{cases} 0 & u_n^m > 1 \\ 1 - u_n^m - \frac{\sigma}{2}\|X_n^m\|_\infty & u_n^m < 1 - \sigma \\ \frac{(u_n^m)^2}{2\sigma\|X_n^m\|_\infty} & else \end{cases}$$

$$h_\sigma^{(2)}(\Omega) = \begin{cases} 0 & u_n^m > 1 \\ 1 - u_n^m - \frac{\sigma}{2}\|w^m\|_\infty & u_n^m < 1 - \sigma \\ \frac{(u_n^m)^2}{2\sigma\|w^m\|_\infty} & else \end{cases}$$

where $\Omega = \{L_n, X_n^m, w^m, b^m\}$ denote the input of the function. We employ the Alternating Direction Method of Multipliers (ADMM)(Boyd et al. 2011) to solve problem (2) with the approximation of hinge loss function. The objective function in (2) is not jointly convex in all optimal variables, but is convex with one of these three blocks $\{X^m, E^m\}$, $\{w^m, b^m\}$ and L^0 when the other three blocks are fixed. Because of the non-smooth functions in (2), it is not tractable to derive an analytical solution to (2). Therefore, an iterative optimization algorithm based on ADMM is derived to obtained the optimal solution. To make the problem separable, $\{Z^m\}$ are introduced as auxiliary variables to replace X^m in the nuclear norm $\|\cdot\|_*$ of (2),and thus $\{\forall m, X^m = Z^m\}$ is introduced as additional constraints. Then the augmented Lagrange function $\mathcal{L}$ is

$$\begin{aligned} &\sum_{m=1}^{M} \{\Phi(\Lambda^m, Y^m - X^m - E^m) + \Phi(\Gamma^m, X^m - Z^m) \\ &+ \frac{\lambda_2}{2}\|w^m\|_2^2\} + \alpha_1 g_1(\{X^m, L, w^m, b^m\}) \\ &+ \alpha_2 g_2(\{X^m, E^m\}) + \alpha_3 g_3(\{X^m, L^0, w^m, b^m\}) \end{aligned} \tag{6}$$

where $\Phi(A, B) = \frac{\mu}{2}\|B\|_F^2 + \mathrm{trace}(A^T B)$, μ is the positive penalty parameter, and $\{\Lambda^m, \Gamma^m\}$ are the lagrange multipliers. The optimization algorithm iteratively updates one

block of variables or lagrange multipliers of (6) by fixing the other variables , which are shown as follows:

Updating $\mathbf{L^0}$:With other variable fixed, L^0 is updated by solving the following problem:

$$\min_{L^0} \sum_{m=1}^{M} \|(X^m)^T w^m + \mathbf{1}b^m - L^0\|_F^2 \tag{7}$$

which has the close-form solution:

$$L^0 = \frac{1}{M} \sum_{m=1}^{M} [(X^m)^T w^m + b^m] \tag{8}$$

Updating $\{\mathbf{X^m}, \mathbf{E^m}\}$:By some manipulations, Z^m and E^m are updated as

$$\begin{aligned} \hat{E}^m &= \arg\min_{E^m} \frac{1}{2}\|E^m - A^m\|_F^2 + \frac{\alpha_2\lambda_1}{\mu}\|E^m\|_1 = \mathcal{S}_{\frac{\alpha_2\lambda_1}{\mu}}(A^m) \\ \hat{Z}^m &= \arg\min_{Z^m} \frac{1}{2}\|Z^m - B^m\|_F^2 + \frac{\alpha_2}{\mu}\|Z^m\|_* = \mathcal{T}_{\frac{\alpha_2}{\mu}}(B^m) \end{aligned} \tag{9}$$

where $A^m = Y^m - X^m + \frac{\Lambda^m}{\mu}$, and $B^m = X^m + \frac{\Gamma^m}{\mu}$. $\mathcal{S}_{(\cdot)(\cdot)}$ is the solf-thresholding operator and $\mathcal{S}_a(A)_{r,c} = \text{sign}(A_{r,c}) \cdot \max(0, |A_{r,c}| - a)$. $\mathcal{T}_{(\cdot)}(\cdot)$ is the singular value soft-thresholding operator, and $\mathcal{T}_a(A) = U_A\mathcal{S}_a(\Sigma_A)V_A^T$ where $U_A\Sigma_A V_A^T$ is the singular value decomposition of A. After updating X^m and Z^m, by employing the smooth version of hinge loss function with respect to X_n^m, X_n^m is updated as follows:

$$\hat{X}_n^m = \tilde{X}_n^m - \tau\nabla_{X_n^m}\mathcal{L}(\tilde{X}_n^m) \tag{10}$$

where $\nabla_{X_n^m}\mathcal{L}$ is the gradient of $\mathcal{L}$ with respect to X_n^m, τ is the step size, and $\nabla_{X_n^m}\mathcal{L}(\tilde{X}_n^m) = \alpha_1\nabla_{X_n^m}h_\sigma^{(2)}(\tilde{X}_n^m) + 2\alpha_3 w^m[(w^m)^T\tilde{X}_n^m + b^m - L^0] - \Lambda_n^m + \Gamma_n^m + \mu(2\tilde{X}_n^m + \hat{E}_n^m - Y_n^m - \hat{Z}_n^m)$.$\tilde{X}_n^m$ is the value of X_n^m before updating. We adopt the proximal gradient method similar to the one in (Lan, Yuen, and Chellappa 2017) to update the variables.

Updating $\{\mathbf{w^m}, \mathbf{b^m}\}$: With other variable fixed, by employing the smooth version of hinge loss function $h_\sigma^{(1)}(\cdot)$, $\{w^m, b^m\}$ are updated by solving the following problem:

$$\begin{aligned} \min_{\{w^m,b^m\}} \frac{\lambda_2}{2}\|w^m\|_2^2 + \alpha_1 \sum_{n=1}^{N}\sum_{m=1}^{M} h_\sigma^{(1)}(L_n, X_n^m, w^m, b^m) \\ + \alpha_3 g_3(\{X^m, L^0, w^m, b^m\}) \end{aligned} \tag{11}$$

which is an unconstrained quadratic problem. We employ the gradient decent to update $\{w^m, b^m\}$, i.e.

$$[(\hat{w}^m)^T, \hat{b}^m] = [(\tilde{w}^m)^T, \tilde{b}^m] - \tau[\nabla_{w^m}^T\mathcal{L}(\tilde{w}^m), \nabla_{b^m}^T\mathcal{L}(\tilde{b}^m)] \tag{12}$$

where $\nabla_{b^m}^T\mathcal{L}(\tilde{b}^m) = \alpha_1\sum_{n=1}^{N}\nabla_{b^m}h_\sigma^{(1)}(\tilde{b}^m) + 2\alpha_3\mathbf{1}^T[(X^m)^T w^m + \mathbf{1}\tilde{b}^m - L^0]$, and $\nabla_{w^m}^T\mathcal{L}(\tilde{w}^m) = \alpha_1\sum_{n=1}^{N}\nabla_{w^m}h_\sigma^{(1)}(\tilde{w}^m) + 2\alpha_3\tilde{w}^m[(\tilde{w}^m)^T\tilde{X}^m + \mathbf{1}b^m - L^0] + \lambda_2\tilde{w}^m$

Algorithm 1: Optimization Algorithm for (4)

Input: Sample number N, modality number M, sample matrix $\{Y^m\}_{m=1}^M$, and label vector $\{L^m\}_{m=1}^M$,
Output: $\{X^{m,i}, E^{m,i}, w^{m,i}, b^{m,i}, \beta^{m,i}\}, L^0$
Initialization:
$i \leftarrow 1, X^{m,i} \leftarrow Y^k, E^{m,i} \leftarrow \mathbf{0}, w^{k,i} \leftarrow \mathbf{0}, b^{m,i} \leftarrow 0$
while *stopping conditions are not satisfied* **do**
  Update $L^{0,i+1}$ via solving (7)
  Update $\{X^{m,i+1}, E^{m,i+1}\}$ via (9) and (10)
  Update $\{w^{m,i+1}, b^{m,i+1}\}$ via (11)
  Update $\{\Lambda^{m,i+1}, \Gamma^{m,i+1}\}$ via (13)
  $\mu^{i+1} \leftarrow \max(\mu_{\max}, \rho\mu^i)$
  $i \leftarrow i + 1$
  Check stopping conditions
end
Obtain β^m via solving (5)

Updating $\{\mathbf{\Lambda^m}, \mathbf{\Gamma^m}\}$:The multipliers are updated as follows:

$$\begin{aligned} \hat{\Gamma}^m &= \tilde{\Gamma}^m + \mu(\hat{X}^m - \hat{Z}^m) \\ \hat{\Lambda}^m &= \tilde{\Lambda}^m + \mu(\hat{Y}^m - \hat{X}^m - \hat{E}^m) \end{aligned} \tag{13}$$

The optimization algorithm iteratively update the optimal valuables and the multipliers until $\|Y^m - X^m - E^m\| < \gamma\|Y^m\|, m = 1, \ldots, M$. In each iteration, the penalty parameter μ is updated as $\hat{\mu} = \max(\mu_{\max}, \rho\mu)$. We set γ as 10^{-5}, ρ as 1.5, $\mu_{\max}$ as 10^6 and the initial value of μ as 10^{-6}. The overall procedure is shown in Algorithm 1.

4 Implementation Details

4.1 Appearance Modeling and Target Decision

After the learned features are obtained by solving problem (3), another issue is how to utilize these feature for appearance modeling. Typical approaches such as SVM (Hare et al. 2016) or sparse representation (Lan, Ma, and Yuen 2014; Lan et al. 2015) can be utilized for appearance modeling. For the sake of robustness, we adopt the sparse representation for appearance modeling. Based on the learned features of different modalities X^m, $m = 1, \ldots, M$, we construct feature sets of different modalities which are denoted as D^m, and also includes some recently obtained important samples. The sparse representations $\{a_i^m\}$ of the target candidates $\{C_i^m\}, i = 1, \ldots, P, m = 1, \ldots, M$ which are sampled by a particle filter can be learned as follows:

$$a_i^m = \arg\min_a \eta\|a\|_1 + \|C_i^m - D^m a\|_2^2 \tag{14}$$

where η controls the tradeoff between the reconstruction error and sparse regularization. Then we define the decision function for target state decision as follows:

$$\begin{aligned} F(\{a^m, C^m\}) = \sum_{m=1}^{M} \|C^m - D^m a^m\|_2^2 \\ + \nu|\sum_{m=1}^{M} \beta^m((w^m)^T D^m a^m + b^m) - 1| \end{aligned} \tag{15}$$

where ν is the tradeoff parameters. The decision function consists of two components: the reconstruction error using

the learned features and the prediction loss with respect to the target label (+1), which simultaneously exploit the reconstruction ability and discriminability of the learned feature for final target decision. Since the classifier parameters are estimated using the learned features, it is more suitable to perform classification in the same feature space. As such, we use the reconstructed samples to calculate the prediction loss for classification. We choose the sate of target candidate which achieve the lowest value of the decision function as the target state.

5 Experiment

5.1 Experimental Setting

Sixteen video pairs[2] which include videos of RGB and infrared modality under different scenarios and conditions are used to evaluate the RGB-infrared tracking performance. These video pairs cover various challenging factors such as occlusion, poor illumination conditions, large scale changes, etc.. They are aligned accurately, which makes the tracked targets of each video pair locate in almost the same position in each video frame from RGB and infrared modalities. Ten baseline methods are used for comparison, which include the STRUCK (Hare et al. 2016), RPT (Li, Zhu, and Hoi 2017), KCF (Henriques et al. 2015), MEEM (Zhang, Ma, and Sclaroff 2014), STC (Zhang et al. 2014), CT (Zhang, Zhang, and Yang 2014), MIL (Babenko, Yang, and Belongie 2011), CN (Danelljan et al. 2014), L1 (Wu et al. 2011), and JSR (Liu and Sun 2012) methods. The L1 and JSR methods are proposed for RGB-infrared tracking. For the remaining trackers, they are originally proposed for tracking objects in RGB modality. Following the setting used in (Li et al. 2016), features from RGB and infrared modalities are concatenated as the input of these trackers so that they can perform RGB-infrared tracking on these fifteen video pairs. Tracking results of these trackers on these RGB-infrared videos can be obtained from (Li et al. 2016).

We empirically set the λ_1, λ_2, α_1, α_2 and α_3 in (2), the C_1 and C_2 in (3), η in (14) and ν in (15) to be 0.1, 0.01, 0.1, 1, 0.02, 0.1, 0.001, 0.01, 0.1, respectively. Considering the tradeoff between stability and adaptivity, in addition to the target and background samples in recent frames, the target sample from the first few frames from the beginning are also included in the training samples. For image patch in RGB modality, we transform it to be in grey scale and extract the HOG feature (Dalal and Triggs 2005) in order to capture the gradient information. For the one in infrared modality, we extract the intensity feature.

5.2 Experimental Results

Two metrics are used to quantitatively evaluate the proposed tracker: VOC overlapping rate and success rate. The VOC overlapping rate is defined as $\frac{area(S_1 \bigcap S_2)}{area(S_1 \bigcup S_2)}$ where S_1 and S_2 are the bounding box of the ground-truth and the tracker. If the overlapping rate of a tracking results in a video frame is larger than 0.5, we regard it as a track success. The success rate is defined as the percentage of video frames in which the track success happen. Tables 1 and 2 record the success rate and the overlapping rate of all the compared tracker on these 16 videos. The quantitative results from these two tables show that the proposed tracker performs better than other ten compared trackers on most videos in terms of overlapping rate and success rate with the best mean performance in terms of both metrics. The proposed tracker ranks in top three on fifteen videos in terms of success rate and ranks in top three on fourteen videos. In particular, it achieves excellent performance on some videos which cover occlusion (e.g. *Minibus1*, *Tricycle*), poor illumination conditions (e.g. *MinibusNig*, *BusScale*), thermal crossover (e.g.*RainyCar1*, *RainyCar2*), etc.. This is because the proposed learning framework can adaptively perform feature decontamination and learning reliable classifiers, which enable it to deal with contaminated samples caused by large appearance changes (e.g. occlusion, illumination change). In addition, by learning classifiers of each modality and performing weight determination under dscirminabity-consistency constrain, cross-modality discrepancy can be reduced and reliable modality can be guaranteed to play more important role in target/background separation. Even if some unreliable modality exists under some scenarios (e.g. infrared modality under the case of thermal crossover), the impact of such kind of modality would be suppressed, and the importance of reliable modality would be enhanced for more effective modality fusion.

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

6 Conclusion

In this paper, we propose a novel discriminative learning framework to fuse RGB-Infrared modality for object tracking. By explicitly imposing the discriminability-consistent constrain, removing outliers and learning uncontaminated feature, collaboratively estimating classifiers and reliability weight of different modalities in an optimal learning framework, the proposed method could alleviate the cross-modality discrepancy and perform effective fusion of multiple modalities to handle large appearance variation more robustly and differentiate the target from background more discriminatively. Extensive comparison experiments with other ten baseline methods demonstrate its effectiveness and excellent performance.

Acknowledgements

This work was supported in part by Hong Kong RGC General Research Fund HKBU12254316 and the National Natural Science Foundation of China under Grant 61672188. The authors would like to thank Dr. Chenglong Li and Dr. Xiao Wang for providing the videos and some results for comparison, and the anonymous reviewers for their suggestions on improving the paper quality.

References

Babenko, B.; Yang, M.; and Belongie, S. 2011. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(8):1619–1632.

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

Sequence	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	L1	JSR	Proposed Method
BlueCar	0.33	0.33	0.28	0.38	**0.94**	0.46	0.38	0.38	**0.68**	0.44	**0.97**
BusScale	0.48	0.4	0.46	0.44	**0.61**	0.53	0.5	0.51	**0.82**	0.56	**1**
BusScale1	0.33	0.34	0.36	0.27	**0.87**	0.45	0.36	0.36	**0.76**	0.47	**0.69**
Exposure2	0.2	0.26	0.2	0.2	**0.45**	0.16	0.2	0.2	**1**	0.19	**0.88**
fastCar2	**0.55**	0.48	0.35	0.43	0.48	0.5	0.53	**0.55**	0.45	**0.57**	**0.78**
Football	**0.9**	0.81	**0.96**	0.83	0.64	0.87	**0.97**	0.76	0.17	0.76	0.84
Cycling	**0.71**	0.43	0.53	**0.71**	0.68	0.02	**0.71**	**0.71**	0.33	0.48	**0.88**
MotorNig	0.37	0.64	0.66	0.66	**0.85**	0.74	0.68	0.79	**1**	0.59	**0.85**
MinibusNig	0.51	0.49	0.55	0.51	**0.92**	0.51	0.54	0.55	**1**	0.36	**0.99**
Minibus1	**0.59**	0.04	0.54	0.58	0.05	0.32	0.54	0.04	**0.69**	0.49	**0.91**
RainyMotor1	0.38	0.21	0.08	0.17	0.03	0.65	0.05	0.56	**0.98**	**0.75**	**0.91**
Tricycle	0.98	0.72	0.99	**1**	**1**	0.98	0.85	0.75	0.56	0.93	**1**
Otcbvs1	0.91	0.87	0.98	**1**	0.98	0.94	0.84	0.82	0.12	**1**	**1**
RainyCar1	**0.58**	0.35	0.55	0.08	**0.98**	0.45	0.57	0.57	0.07	0.05	**0.9**
RainyCar2	0.55	0.52	0.3	0.43	**0.76**	0.62	0.54	0.65	**0.76**	0.62	**0.8**
Jogging	0.92	0.97	0.7	0.82	**1**	0.99	0.99	**1**	0.24	0.77	**1**
Average	0.58	0.49	0.53	0.53	**0.7**	0.57	0.58	0.58	**0.6**	0.56	**0.9**

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

Sequence	STRUCK	STC	CT	MIL	RPT	MEEM	KCF	CN	L1	JSR	Proposed Method
BlueCar	0.37	0.27	0.34	0.4	**0.65**	0.47	0.4	0.4	**0.63**	0.4	**0.77**
BusScale	0.47	0.45	0.46	0.49	**0.57**	0.52	0.51	0.51	**0.72**	0.54	**0.72**
BusScale1	0.4	0.41	0.43	0.39	**0.67**	0.44	0.43	0.42	**0.66**	0.47	**0.61**
Exposure2	0.32	0.37	0.31	0.32	**0.48**	0.3	0.32	0.32	**0.82**	0.35	**0.58**
fastCar2	**0.57**	0.53	0.43	0.48	0.51	0.49	0.5	0.54	0.34	**0.56**	**0.55**
Football	0.65	0.6	**0.73**	**0.67**	0.56	0.65	**0.69**	0.59	0.26	0.62	0.64
Cycling	**0.62**	0.47	0.51	**0.64**	0.55	0.03	0.61	**0.63**	0.36	0.49	**0.62**
MotorNig	0.46	0.61	**0.63**	0.6	**0.67**	**0.63**	0.61	**0.63**	**0.73**	0.6	0.58
MinibusNig	0.54	0.55	0.54	0.55	**0.68**	0.55	0.57	0.59	**0.74**	0.33	**0.76**
Minibus1	0.53	0.05	0.52	0.55	0.06	0.38	**0.56**	0.05	**0.69**	0.53	**0.65**
RainyMotor1	0.47	0.39	0.29	0.38	0.05	0.56	0.06	0.56	**0.66**	**0.63**	**0.64**
Tricycle	0.68	0.64	0.62	0.71	**0.72**	**0.73**	0.64	0.64	0.57	0.67	**0.76**
Otcbvs1	0.63	0.69	0.65	**0.73**	0.72	0.66	0.66	0.68	0.14	**0.79**	**0.74**
RainyCar1	**0.58**	0.5	0.55	0.07	**0.69**	0.49	0.55	0.55	0.07	0.05	**0.58**
RainyCar2	0.55	0.46	0.35	0.44	**0.58**	0.55	0.4	0.55	**0.63**	0.52	**0.59**
Jogging	0.61	0.66	0.58	0.62	**0.77**	0.7	0.72	**0.77**	0.27	0.64	**0.73**
Average	**0.53**	0.48	0.5	0.5	**0.56**	0.51	0.51	**0.53**	0.52	0.51	**0.66**

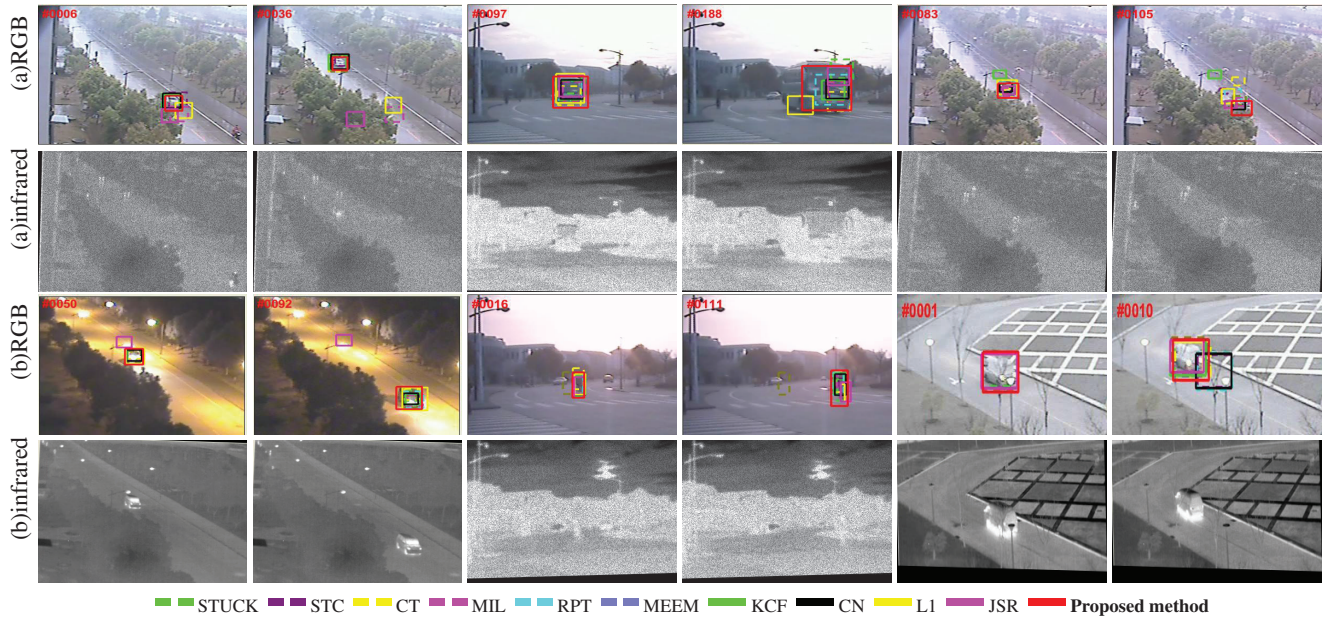


Figure 2: Qualitative results on some frames of RGB and infrared modality with challenging factors. (a) Thermal crossover and scale changes(*RainyCar1, BusScale, RainyCar2*). (b) Low illumination conditions and occlusion(*MinibusNig, Cycling, Minibus1*). The RGB infrared modality are shown in the top and bottom rows of each sub-figure, respectively.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3(1).

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

Conaire, C. Ó.; O'Connor, N. E.; and Smeaton, A. F. 2008. Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. *Mach. Vis. Appl.* 19(5-6):483–494.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 886–893.

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

Demiriz, A.; Bennett, K. P.; and Shawe-Taylor, J. 2002. Linear programming boosting via column generation. *Mach. Learn.* 46(1):225–254.

Han, J.; Pauwels, E. J.; de Zeeuw, P. M.; and 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(2):255–263.

Han, J.; Shao, L.; Xu, D.; and Shotton, J. 2013. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybernetics* 43(5):1318–1334.

Hare, S.; Golodetz, S.; Saffari, A.; Vineet, V.; Cheng, M.; Hicks, S. L.; and Torr, P. H. S. 2016. Struck: Structured output tracking with kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(10):2096–2109.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2015. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(3):583–596.

Hu, J.-F.; Zheng, W.-S.; Lai, J.; and Zhang, J. 2015. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proc. CVPR*, 5344–5352.

Lan, X.; Ma, A.; Yuen, P.; and Chellappa, R. 2015. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. *IEEE Trans. Image Process.* 24(12):5826–5841.

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

Lan, X.; Yuen, P. C.; and Chellappa, R. 2017. Robust mil-based feature template learning for object tracking. In *Proc. AAAI*, 4118–4125.

Lan, X.; Zhang, S.; and Yuen, P. C. 2016. Robust joint discriminative feature learning for visual tracking. In *Proc. IJCAI*, 3403–3410.

Leykin, A., and Hammoud, R. I. 2010. Pedestrian tracking by fusion of thermal-visible surveillance videos. *Mach. Vis. Appl.* 21(4):587–595.

Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; and Lin, L. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Trans. Image Processing* 25(12):5743–5756.

Li, Y.; Zhu, J.; and Hoi, S. C. 2017. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *Proc. CVPR*, 353–361.

Liu, H., and Sun, F. 2012. Fusion tracking in color and infrared images using joint sparse representation. *Sci. China Inf. Sci.* 55(3):590–599.

Liu, J.; Jiang, Y.; Li, Z.; Zhou, Z.-H.; and Lu, H. 2015. Partially shared latent factor learning with multiview data. *IEEE Trans. Neural. Netw. Learn. Syst.* 26(6):1233–1246.

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

Nesterov, Y. 2005. Smooth minimization of non-smooth functions. *Math. Prog.* 103(1):127–152.

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

Wang, Z.; Hu, R.; Liang, C.; Yu, Y.; Jiang, J.; Ye, M.; Chen, J.; and Leng, Q. 2016b. Zero-shot person re-identification via cross-view consistency. *IEEE Trans. Multimedia* 18(2):260–272.

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

Wu, Y.; Blasch, E.; Chen, G.; Bai, L.; and Ling, H. 2011. Multiple source data fusion via sparse representation for robust visual tracking. In *Proc. Int. Conf. Inf. Fusion,*, 1–8.

Xu, C.; Tao, D.; and Xu, C. 2015. Large-margin multi-label causal feature learning. In *Proc. AAAI*, 1924–1930.

Xu, J.; Wang, D.; and Lu, H. 2012. Fragment-based tracking using online multiple kernel learning. In *Proc. ICIP*, 393–396.

Ye, M.; Liang, C.; Wang, Z.; Leng, Q.; and Chen, J. 2015. Ranking optimization for person re-identification via similarity and dissimilarity. In *ACM MM*, 1239–1242.

Ye, M.; Liang, C.; Yu, Y.; Wang, Z.; Leng, Q.; Xiao, C.; Chen, J.; and Hu, R. 2016. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Trans. Multimedia* 18(12):2553–2566.

Ye, M.; Ma, A. J.; Zheng, L.; Li, J.; and Yuen, P. C. 2017. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, 5142–5150.

Yuan, Y.; Fang, J.; and Wang, Q. 2014. Robust superpixel tracking via depth fusion. *IEEE Trans. Circuits Syst. Video Techn.* 24(1):15–26.

Zhang, S.; Yao, H.; Sun, X.; and Lu, X. 2013a. Sparse coding based visual tracking: Review and experimental comparison. *Pattern Recognit.* 46(7):1772–1788.

Zhang, S.; Yao, H.; Zhou, H.; Sun, X.; and Liu, S. 2013b. Robust visual tracking based on online learning sparse representation. *Neurocomputing* 100:31–40.

Zhang, K.; Zhang, L.; Liu, Q.; Zhang, D.; and Yang, M.-H. 2014. Fast visual tracking via dense spatio-temporal context learning. In *Proc. ECCV*, 127–141.

Zhang, S.; Zhou, H.; Jiang, F.; and Li, X. 2015. Robust visual tracking using structurally random projection and weighted least squares. *IEEE Trans. Circuits Syst. Video Techn.* 25(11):1749–1760.

Zhang, S.; Lan, X.; Yao, H.; Zhou, H.; Tao, D.; and Li, X. 2017a. A biologically inspired appearance model for robust visual tracking. *IEEE Trans. Neural Netw. Learn. Syst.* 28(10):2357–2370.

Zhang, S.; Lan, X.; Qi, Y.; and Yuen, P. C. 2017b. Robust visual tracking via basis matching. *IEEE Trans. Circuits Syst. Video Techn.* 27(3):421–430.

Zhang, J.; Ma, S.; and Sclaroff, S. 2014. Meem: Robust tracking via multiple experts using entropy minimization. In *Proc. ECCV*, 188–203.

Zhang, H.; Patel, V. M.; and Chellappa, R. 2017. Hierarchical multimodal metric learning for multimodal classification. In *Proc. CVPR*, 3057–3065.

Zhang, K.; Zhang, L.; and Yang, M.-H. 2014. Fast compressive tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(10):2002–2015.