Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

Robust Joint Discriminative Feature Learning for Visual Tracking

Xiangyuan Lan, Shengping Zhang, Pong C. Yuen
Department of Computer Science, Hong Kong Baptist University
{xylan, csspzhang, pcyuen}@comp.hkbu.edu.hk

Abstract

Because of the complementarity of multiple visual cues (features) in appearance modeling, many tracking algorithms attempt to fuse multiple features to improve the tracking performance from two aspects: increasing the representation accuracy against appearance variations and enhancing the discriminability between the tracked target and its background. Since both these two aspects simultaneously contribute to the success of a visual tracker, how to fully unleash the capabilities of multiple features from these two aspects in appearance modeling is a key issue for feature fusion-based visual tracking. To address this problem, different from other feature fusion-based trackers which consider one of these two aspects only, this paper proposes an unified feature learning framework which simultaneously exploits both the representation capability and the discriminability of multiple features for visual tracking. In particular, the proposed feature learning framework is capable of: 1) learning robust features by separating out corrupted features for accurate feature representation, 2) seamlessly imposing the discriminabiltiy of multiple visual cues into feature learning, and 3) fusing features by exploiting their shared and feature-specific discriminative information. Extensive experiment results on challenging videos show that the the proposed tracker performs favourably against other ten state-of-the-art trackers.

1 Introduction

As an important research topic in the field of computer vision, visual tracking has been extensively studied in last two decades with the development of numerous tracking algorithms. However, it still remains challenging due to large appearance variations of the tracked object caused by illumination, occlusion, etc. To account for these variations as well as their complicated interaction, different kinds of visual cues (features) that describe different characteristics of the object, e.g. color, texture are fused and jointly exploited for more robust appearance modeling [Grabner and Bischof, 2006; Hong *et al.*, 2013; Lan *et al.*, 2014]. Given multiple visual cues (features) of the tracked object, a key problem is how to fuse them properly to obtain both accurate target representation and effective target/background discrimination.

A variety of feature fusion-based tracking methods have been proposed, which can be roughly divided into two categories: discriminative methods and generative methods according to their feature fusion strategies. Discriminative models combine the discriminabilities of different features to facilitate the discrimination between the target and its background. Typical approaches belonging to this category such as [Grabner and Bischof, 2006; Grabner *et al.*, 2008; Babenko *et al.*, 2011] are grounded on online boosting. For example, [Grabner and Bischof, 2006] proposed an online boosting-based tracker in which a large weak classifier pool is learned and updated from various kinds of raw features, and an online boosting algorithm is employed to select and fuse weak classifiers for target/background discrimination. Along this line, more variants of boosting-based approaches such as [Grabner *et al.*, 2008; Babenko *et al.*, 2011] are developed to deal with the drifting problem. Discriminative models can alleviate the background distraction problem to some extent because different features are jointly exploited for foreground/background separation. However, the classifiers which discriminative methods exploit for feature fusion are directly learned and updated with the raw features extracted from the target samples. If the target samples are corrupted/contaminated by external variations such as occlusion, illumination, etc, the extracted features may not well reflect the intrinsic properties of the object appearance. Therefore, learning and updating classifiers using such corrupted features may degrade the fusion performance, which urges the need to remove the corrupted features or learn some uncontaminated features for robust appearance modeling.

Unlike discriminative methods, generative methods directly fuse multiple features to represent the object to increase the representation ability of a tracker. To enhance the tracking robustness to large appearance variations, some strategies are adopted, e.g. using trivial templates [Mei *et al.*, 2015; Hu *et al.*, 2015b] to model the outliers existing in the target's appearance, removing unreliable features [Lan *et al.*, 2015] for robust feature-level fusion, or integrating responses from various Gabor kernels to capture the local appearance changes [Zhang *et al.*, 2016]. Since generative methods directly fuse features for model learning without mapping them

to classification scores, they preserve more information of multiple features, and are more capable of accounting for appearance changes than discriminative methods. However, generative methods do not take advantage of background information for appearance modeling with multiple features, which may make them easily to be distracted by cluttered background and lead to tracking failure.

Generally speaking, discriminative and generative methods have complementary advantages in feature fusion-based appearance modeling, and the success of a visual tracker depends on both its representation ability against appearance variations and its discriminability between the target and its background. As such, the advantages of these two approaches should be exploited jointly for more robust feature fusion, so that multiple features can be employed simultaneously to describe the object appearance accurately and separate the object from background discriminatively. In addition, different features extracted from the same object share some consistency while each feature should also have some specific knowledge in their discriminability. As is pointed out in [Liu *et al.*, 2014], consistency is closely related to agreement while feature specific knowledge provides complementarity and is related to disagreement. While exploring feature-specific discriminative information for feature fusion has been shown to be effective in some methods such as online boosting [Grabner and Bischof, 2006; Grabner *et al.*, 2008; Babenko *et al.*, 2011], the benefits of exploiting shared information among multiple features/modalities/views have also been well demonstrated in some learning and classification tasks recently [Yang *et al.*, 2012; Hu *et al.*, 2015a; Wang *et al.*, 2015]. This motivates us to explore an effective strategy to jointly consider the shared and feature-specific discriminative information for feature fusion.

Based on aforementioned motivations, we propose a novel robust joint discriminative feature learning framework for object tracking using multiple visual cues. Different from other feature fusion-based trackers which directly employ potentially contaminated raw features and utilize the representation ability or discriminability of different visual cues alone, the proposed method aims to learn uncontaminated and discriminative features to jointly exploit the representation and discriminative power of multiple visual cues for visual tracking. Within this unified framework, feature learning is performed by simultaneously and optimally removing corrupted features and learning reliable classifiers. As such, feature learning from multiple visual cues with corrupted feature removal offers uncontaminated features for reliable classifier learning while discriminative classifier learning with multiple visual cues imposes the discriminability to the learned features. Therefore, the limitations of the generative and discriminative approaches can be compensated and the benefits of these approaches can be combined. In addition, we incorporate a novel feature fusion scheme into the feature learning framework to further exploit the shared and feature-specific discriminative information for feature fusion, and the importance of different features in target/background discrimination is also dynamically weighted in this optimal learning framework. By jointly exploiting the learned features and classifiers from multiple visual cues for target representation and target/background classification, the learning framework enhances the tracking performance in term of representation accuracy and discrimination reliability.

It should be noted that some hybrid approaches which attempt to combine the benefits of both the generative and discriminative approaches have been developed, e.g. [Yu *et al.*, 2008]. Their models are developed in the context of using a single feature, while the proposed model is developed for multi-feature appearance model and can be more effectively used for features learning and fusion with multiple visual cues. Although existing single feature-based hybrid approaches may be applied to multiple features by feature concatenation, such an approach ignores different statistical properties of different features and may result in a long feature vector that may degrade the learning efficiency. The proposed method is also different from the recent developed fusion-based tracker [Zhang *et al.*, 2015b] since the proposed model focuses on feature fusion while [Zhang *et al.*, 2015b] focuses on tracker fusion.

The contributions of this paper are listed as follows:

- We propose a novel feature learning model which is able to simultaneously and optimally learn discriminative features and reliable classifiers from potentially contaminated samples to exploit the representation and discriminative power of multiple visual cues for visual tracking.
- We propose a novel feature fusion scheme which simultaneously considers the shared and feature-specific discriminative information from multiple visual cues. Therefore, both the consistency and complementarity of the discriminability of multiple features are jointly exploited for more robust feature fusion.
- We derive a four-step iterative optimization algorithm to effectively solve the proposed robust joint discriminative feature learning model.

2 Related Work

2.1 Feature Learning in Visual Tracking

Recent works on feature learning-based tracking include dictionary learning-based approaches and deep learning-based approaches. Various dictionary learning based-trackers are proposed to update tracking model effectively [Zhang *et al.*, 2015a], enhance the reconstructive and discriminative power [Fan *et al.*, 2014] of the appearance model, etc. Most dictionary learning-based trackers use a single feature, i.e. intensity only, which may not be sufficient to account for large appearance variations. Deep learning-based trackers such as [Li *et al.*, 2014] tune an off-line pre-trained deep neural network online to adapt the appearance variations, which may not be efficient. Different from the aforementioned approaches, this paper aims to exploit multiple visual cues for feature learning without off-line large-scale training samples.

2.2 Shared and Feature-Specific Information among Multiple Features/Modalities/Views for Pattern Classification

Exploiting the shared and feature-specific information among multiple features/modalities/views jointly has been shown to be beneficial for pattern classification. [Yang *et al.*, 2012] proposed a multi-feature collaborative model which simultaneously models the similar and distinctive information among multiple features for image classification. [Hu *et al.*, 2015a] proposed to learn the shared and feature-specific structure of heterogeneous channels for RGB-D activity recognition. A multi-modal sharable and specific feature learning approach is proposed for learning the shared and model-specific properties for RGB-D object recognition [Wang *et al.*, 2015]. These research findings motivate us to exploit the shared and feature-specific discriminative information for feature fusion-based visual tracking.

3 Proposed Model

3.1 Robust Joint Discriminative Feature Learning

In the t-th frame, let $Y_F^k = [y_1^k, \ldots, y_{n_1}^k]$ be the recently obtained target samples of the k-th visual cues, and n_1 denote the number of target samples in the sample set. Since large appearance variations, e.g. occlusion, illumination may occur during tracking, the captured samples may be contaminated/corrupted. To ensure the robustness of the learned features, explicitly seperating out the corrupted samples is essential. Therefore, one objective of the learning model is to learn the uncontaminated features while separating out the corrupted features as follows:

$$Y_F^k = X_F^k + E_F^k, \quad k = 1, \ldots, K \tag{1}$$

where X_F^k and E_F^k are the learned uncontaminated features and the separated corrupted features, respectively, and K is the number of visual cues. The background samples near the target position, also known as local context information in the current frame may share some similarity with the recently obtained target samples, such as illumination conditions. Exploiting such context information has been shown to be beneficial for tracking [Grabner *et al.*, 2010]. Besides, the target samples in recent frames are temporally correlated, and mining the latent structure embedded in the target samples can facilitate revealing the intrinsic characteristic of the target's features [Ross *et al.*, 2008]. To further exploit the spatial and temporal correlation of the background and the target samples for feature learning while separating out the corrupted features from multiple visual cues, we cast the objective discussed above into the following rank and sparsity minimization problem with the sample set of multiple visual cues:

$$\begin{aligned} \min_{\{X^k, E^k\}_{k=1}^K} \quad & \sum_{k=1}^{K} \{rank(X^k) + \lambda_1 \|E^k\|_1\} \\ \text{s.t.} \quad & Y^k = X^k + E^k, \ k = 1, \ldots, K \end{aligned} \tag{2}$$

where $Y_B^k = [y_{n_1+1}^k, \ldots, y_N^k]$ is the nearby background samples in current frame, N is the total number of samples. $Y^k = [Y_F^k, Y_B^k]$, $X^k = [X_F^k, X_B^k]$ and $E^k = [E_F^k, E_B^k]$ is the original feature set, the learned feature set and the separated feature set of the target samples and the background samples in the k-th visual cue, respectively. With the same merit of RPCA [Candès *et al.*, 2011], the rank minimization term is able to uncover the shared latent space embedded in the samples of different visual cues which characterizes intrinsic properties of uncontaminated features of the target and the background, while the sparsity regularization is employed to model the outliers that exist in the corrupted features.

Although the feature learning scheme in (2) is able to learn uncontaminated informative features from multiple visual cues for target representation via joint low-rank and sparse matrix decomposition, it cannot guarantee that the target and background samples can be well discriminated with the learned features. As such, appearance modeling with the learned features may suffer the loss of discriminability, which may lead to the background distraction problem. To strengthen the discriminability of the learned features while modeling such different discriminabilities of different visual cues for robust feature fusion, we impose the discriminability regularization which measures the prediction loss using the learned classifiers to the feature learning process as follows:

$$\begin{aligned} \min_{\{w^k, b^k, \beta^k\}_{k=1}^K} \quad & \sum_{k=1}^{K} ((\beta^k)^2 \|(X^k)^T w^k + \mathbf{1} b^k - L^k\|_2^2 + \lambda_2 \|w^k\|_2^2) \\ \text{s.t.} \quad & \sum_{k=1}^{K} \beta^k = 1, \quad \beta^k \geq 0, \ k = 1, \ldots, K \end{aligned} \tag{3}$$

where $L^k = [L_1^k, ..., L_N^k]^T$ is the label vector, $L_i^k = +1(-1)$ means that the i-th sample of the k-the visual cue x_i^k belongs to the target (background) class, β^k is the importance weight of the prediction loss corresponding to the k-th visual cue, $w^k \in \mathbb{R}^{d_k}$, $\mathbf{1} \in \mathbb{R}^N$ whose elements are all 1s, $b^k \in \mathbb{R}$ and d_k is the dimension of the k-th visual cue. From (3), we can see that imposed discriminability regularization aims to minimize the weighted sum of the prediction loss of the learned features in different visual cues based on multiple linear classifiers $\{w^k, b^k\}_{k=1}^K$ which are learned jointly. Therefore, it ensures the learned features in multiple visual cues for the target and the background samples can be linearly separated as well as possible, which is able to enhance the discriminability of the tracking model and hence alleviates the background distraction problem. Moreover, dynamically learning and updating the importance weights during tracking allows the discriminative powers of different visual cues to be adaptively evaluated, which guarantees that more discriminative features play more important roles in target/background discrimination. Here we use $(\beta^k)^2$ instead of β^k for feature fusion because we want to ensure all the weights are positive which avoids the trivial solution that the weight corresponding to the lowest prediction loss is 1, and 0 otherwise.

To further exploit the shared and feature-specific discriminative information of multiple visual cues, we introduce the following objective function into the proposed feature learning framework:

$$\min_{\{w^k, b^k\}_{k=1}^K, L^*} \quad \sum_{k=1}^{K} \theta^k \|(X^k)^T w^k + \mathbf{1} b^k - L^*\|_2^2 \tag{4}$$

where the $L^* = [L_1^*, \ldots, L_N^*]^T$, and L_i^* is the learned consensus classification score of different visual cues in the i-th training sample, which reflects the consistent discriminative information from different visual cues. Different from

other discriminative feature fusion models which enforce different visual cues to share the same classification score [Yu *et al.*, 2013] or to be with diverse discriminative information [Liu and Yao, 1999], the objective function softly regularizes the classification scores towards the consensus while enabling them to have some disagreement with the consensus. Therefore, both the consistent and feature-specific information among multiple visual cues are explicitly and jointly employed for learning informative features and reliable classifiers. Moreover, the disagreement with the consensus can be controlled by θ^k, and larger (less) θ^k will promote less (larger) disagreement.

Unifying them all together. Based on the aforementioned analysis, we formulate the objectives as mentioned above into an unified joint discriminative feature learning framework in which uncontaminated and corrupted features, classier parameters of multiple visual cues, denoted as $\Omega = \{L^*, X^k, E^k, w^k, b^k, \beta^k | k = 1, \ldots, K\}$ are jointly estimated as follows:

$$\begin{aligned} \min_{\Omega} \quad & \sum_{k=1}^{K} \{\|X^k\|_* + \lambda_1 \|E^k\|_1 + \frac{\lambda_2}{2}\|w^k\|_2^2 \\ & + \frac{\alpha_1 (\beta^k)^2}{2N} \|(X^k)^T w^k + \mathbf{1} b^k - L^k\|_2^2 \\ & + \frac{\alpha_2 \theta^k}{2N} \|(X^k)^T w^k + \mathbf{1} b^k - L^*\|_2^2\} \qquad (5) \\ \text{s.t.} \quad & Y^k = X^k + E^k, \ \sum_{k=1}^{K} \beta^k = 1 \\ & \beta^k \geq 0, \ k = 1, \ldots, K \end{aligned}$$

where α_1 and α_2 are the nonnegative parameters associating with different objective functions, and the constant $\frac{1}{2}$ is used for simplifying deductions. Since the $rank$ minimization in (2) is a NP-hard problem, we employ the standard approach [Candès *et al.*, 2011] and relax this problem by using nuclear norm $\|\cdot\|_*$ instead. The optimization procedure for (5) is derived in Section 3.2.

3.2 Optimization

The objective function in (5) is convex with respect to one of these four blocks $\{X^k, E^k\}_{k=1}^K$, $\{w^k, b^k\}_{k=1}^K$, L^* and $\{\beta^k\}_{k=1}^K$ when the other three blocks are fixed, and it's difficult to derive the analytical solution to (5). Therefore, we derive an iterative optimization algorithm to solve the problem. To make the problem separable, we introduce the auxiliary variables $\{Z^k\}_{k=1}^K$ to replace $\{X^k\}_{k=1}^K$ in the nuclear norm $\|\cdot\|_*$ of (5). Accordingly, $\{\forall k, X^k = Z^k\}$ act as additional constraints. Let $\mathcal{C}$ be the constraint set of (5) on $\{\beta^k\}_{k=1}^K$, and $a_i = \frac{\alpha_i}{N}$ for $i = 1$ *or* 2 . Then the augmented Lagrange function of (5) $\mathcal{L}_{\Omega \in \mathcal{C}}$ is

$$\begin{aligned} & \sum_{k=1}^{K} \{\|Z^k\|_* + \lambda_1 \|E^k\|_1 + \Phi(\Lambda^k, Y^k - X^k - E^k) \\ & \frac{a_1 (\beta^k)^2}{2} \|(X^k)^T w^k + \mathbf{1} b^k - L^k\|_2^2 + \Phi(\Gamma^k, X^k - Z^k) \\ & + \frac{\alpha_2 \theta^k}{2} \|(X^k)^T w^k + \mathbf{1} b^k - L^*\|_2^2 + \frac{\lambda_2}{2}\|w^k\|_2^2\} \qquad (6) \end{aligned}$$

Algorithm 1: Optimization Algorithm for (5)

Input: Sample matrix $\{Y^k\}_{k=1}^K$, label vector $\{L^k\}_{k=1}^K$, sample number N and feature number K

Output: $\{X^{k,i}, E^{k,i}, w^{k,i}, b^{k,i}, \beta^{k,i}\}_{k=1}^K, L^*$

Initialization: $i \leftarrow 1, X^{k,i} \leftarrow Y^k, E^{k,i} \leftarrow \mathbf{0}, \beta^{k,i} \leftarrow \frac{1}{K}, w^{k,i} \leftarrow \mathbf{0}, b^{k,i} \leftarrow 0, k = 1, ...K, a_t \leftarrow \frac{\alpha_t}{N}, t = 1$ or 2

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

- Update $\{X^{k,i+1}, E^{k,i+1}\}_{k=1}^K$ via Algorithm (2)
- Update $\{w^{k,i+1}, b^{k,i+1}\}_{k=1}^K$ via solving (10)
- Update $L^{*,i+1}$ via solving (11)
- Update $\{\beta^{k,i+1}\}_{k=1}^K$ via solving (12)
- $i \leftarrow i + 1$
- Check stopping conditions

end

with the definition $\Phi(A, B) =< A, B > + \frac{\mu}{2}\|B\|_F^2$, where μ is a positive penalty scalar, $< A, B >= trace(A^T B)$ and, $\{\Lambda^k, \Gamma^k\}_{k=1}^K$ are the Lagrangian multipliers. Based on (6), the solutions to (5) can be obtained by iteratively solving the subproblems of (5) in which an inner loop procedure is employed to solve $\{X^k, E^k\}$.

$\{\mathbf{X^k}, \mathbf{E^k}\}$-subproblem: Keeping other variables fixed, we obtain $X^k, E^k, k = 1, \ldots, K$ using Alternating Direction Method of Multipliers (ADMM) [Boyd *et al.*, 2011]. In the $(j+1)$-th step of ADMM, by some algebraic manipulations, $E^{k,j+1}$ and $Z^{k,j+1}$ are obtained as

$$\begin{aligned} Z^{k,j+1} &= \arg\min_{Z^k} \frac{1}{2}\|Z^k - Q^{k,j}\|_F^2 + \frac{1}{\mu}\|Z^k\|_* = \mathcal{J}_{\frac{1}{\mu}}(Q^{k,j}) \\ E^{k,j+1} &= \arg\min_{E^k} \frac{1}{2}\|E^k - P^{k,j}\|_F^2 + \frac{\lambda_1}{\mu}\|E^k\|_1 = \mathcal{S}_{\frac{\lambda_1}{\mu}}(P^{k,j}) \end{aligned} \qquad (7)$$

where $Q^{k,j} = X^{k,j} + \frac{\Gamma^{k,j}}{\mu}$, and $P^{k,j} = Y^{k,j} - X^{k,j} + \frac{\Lambda^{k,j}}{\mu}$. $\mathcal{S}_{(\cdot)}(\cdot)$ is a soft-thresholding operator such that $\mathcal{S}_p(A)_{m,n} = sign(A_{m,n}) \cdot \max(|A_{m,n}| - p, 0)$, and $\mathcal{J}_{(\cdot)}(\cdot)$ is the singular value soft-thresholding operator such that $\mathcal{J}_p(A) = U_A \mathcal{S}_p(\Sigma_A) V_A^T$ where $U_A \Sigma_A V_A^T$ is the singular value decomposition of A. Then by taking partial derivatives of (6) with respect to X^k, we obtain

$$X^{k,j+1} = [G_1^{k,j}]^{-1} G_2^{k,j} \qquad (8)$$

where $G_1^{k,j} = (a_1(\beta^k)^2 + a_2) w^k (w^k)^T + 2\mu I$, and $G_2^{k,j} = a_1(\beta^k)^2 w^k ((L^k)^T - b^k \mathbf{1}^T) + a_2 \theta^k w^k ((L^*)^T - b^k \mathbf{1}^T) + \mu(Y^{k,j+1} - E^{k,j+1} + Z^{k,j+1}) + \Lambda^{k,j} - \Gamma^{k,j}$.

After $\{X^{k,j+1}, Z^{k,j+1}, E^{k,j+1}\}_{k=1}^K$ are obtained, the Lagrangian multipliers are updated as follows:

$$\begin{aligned} \Lambda^{k,j+1} &= \Lambda^{k,j} + \mu(Y^{k,j+1} - X^{k,j+1} - E^{k,j+1}) \\ \Gamma^{k,j+1} &= \Gamma^{k,j} + \mu(X^{k,j+1} - Z^{k,j+1}) \end{aligned} \qquad (9)$$

The ADMM algorithm iteratively updates the optimal variables and the Lagrangian multipliers in the inner loop of the whole optimization algorithm until $\|Y^k - X^k - E^k\|_F^2 < \epsilon \|Y^k\|_F^2$ or the maximum iteration number is reached. Then with $\{X^{k,i}, E^{k,i}\}_{k=1}^K$ in the i-th step of the outer loop, we can solve the subproblem with respect to other variables.

Algorithm 2: Solver for $\{X^k, E^k\}$-subproblem

Input: Sample matrix of k-th visual cue Y^k, label vector L^k, sample number N, other optimal variable with fixed values w^k, b^k, β^k, L^*, initial values $X^{k,i-1}, E^{k,i-1}$ from $(i-1)$-th iteration in Algorithm 1

Output: $X^{k,i}, E^{k,i}$

Initialization: $j \leftarrow 1, X^{k,j} \leftarrow X^{k,i-1}, E^{k,j} \leftarrow E^{k,i-1}, Z^{k,j} \leftarrow X^{k,i-1}, \Gamma^{k,j} \leftarrow \mathbf{0}, \Lambda^{k,j} \leftarrow \mathbf{0}$

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

- Update $Z^{k,j+1}$ and $E^{k,j+1}$ via (7)
- Update $X^{k,j+1}$ via (8)
- Update $\Lambda^{k,j+1}$ and $\Gamma^{k,j+1}$ via (9)
- $j \leftarrow j+1$
- Check stopping conditions

end

$X^{k,i} \leftarrow X^{k,j}, E^{k,i} \leftarrow E^{k,j}$

$\{\mathbf{w^k}, \mathbf{b^k}\}$**-subproblem**: With other variables fixed, it is equivalent to solve the following problem:

$$\min_{w^k, b^k} \quad \frac{a_1(\beta^k)^2}{2}\|(X^k)^T w^k + \mathbf{1}b^k - L^k\|_2^2 + \frac{\alpha_2\theta^k}{2}\|(X^k)^T w^k + \mathbf{1}b^k - L^*\|_2^2 + \frac{\lambda_2}{2}\|w^k\|_2^2 \tag{10}$$

which is an unconstrained quadratic programming problem and can be solved by some standard optimization techniques, e.g. conjugate gradient descent method.

$\mathbf{L^*}$**-subproblem**: Similar to the $\{w^k, b^k\}$-$subproblem$, some standard optimization techniques can be used to solve the following unconstrained quadratic programming problem:

$$\min_{L^*} \quad \sum_{k=1}^{K} \frac{\alpha_2\theta^k}{2}\|(X^k)^T w^k + \mathbf{1}b^k - L^*\|_2^2 \tag{11}$$

$\beta^{\mathbf{k}}$**-subproblem**: Let $R^k = \frac{a^k}{2}\|(X^k)^T w^k + \mathbf{1}b^k - L^k\|_2^2$. Then this subproblem can be rewritten as

$$\min_{\beta^k} \sum_{k=1}^{K} (\beta^k)^2 R^k \quad \text{s.t.} \sum_{k=1}^{K} \beta^k = 1, \quad \beta^k \geq 0,\ k = 1, \ldots, K \tag{12}$$

which is a quadratic programming problem with linear constraints. Based on its Lagrange function, the solution can be derived as $\beta^{k'} = \frac{(R^{k'})^{-1}}{\sum_{k=1}^{K}(R^k)^{-1}}$. We iteratively solve these four subproblems until the relative change of the valuables in the adjacent iterations are less than a predefined threshold or the maximum iteration number is reached. The optimization algorithm for (5) and the solver for $\{X^k, E^k\}$-subproblem are summarized in Algorithm 1 and 2, respectively.

4 Implementation Details

4.1 Target Representation

For the sake of robustness and effectiveness [Zhang *et al.*, 2013a], we adopt the sparse representation scheme [Mei and Ling, 2011] for target representation. To further enhance the adaptivity of the proposed tracker, we augmented the learned features for each visual cue A^k with the recent obtained important target samples R_F^k to construct a template set $D^k = [X_F^k, R_F^k]$ where R_F^k is updated adaptively similar to [Mei and Ling, 2011]. Then we can obtain the sparse representations of the target candidates of each visual cue $u_i^k, i = 1, \ldots, m, k = 1, \ldots, K$ as follows:

$$u_i^k = \arg\min_u \|t_i^k - D^k u\|_2^2 + \lambda\|u\|_1 \tag{13}$$

where λ is the tradeoff between the reconstruction error and the sparseness, t_i^k is the k-th visual cue of the i-th target candidate sampled by particle filter in every frame.

4.2 Observation Likelihood for Particle Filtering

The proposed tracking algorithm is developed in the particle filtering framework. After obtaining the optimal solution in (13), we derive the observation likelihood based on the learned features and classifiers as follows:

$$p(o_t|s_t) \propto \exp(-\sum_{k=1}^{K}\|t^k - D^k u^k\|_2^2 - \rho|\sum_{k=1}^{K}\beta^k((w^k)^T D^k u^k + b^k) - 1|) \tag{14}$$

Here we use the joint decisions based on the reconstruction error and classification reliability of multiple visual cues to find out the true state of the target. This is because a good candidate should have high confidence to be the label of the target (+1) while it should also have low reconstruction error with multiple features. Therefore, both the representation abilities and discriminabilities of multiple visual cues are jointly exploited for target state estimation. It should be noted that the classification reliability is based on the reconstructed samples from the learned features instead of using the original samples. This is because the classifiers are estimated using the learned features, and it is more reliable to perform classification in the learned feature space.

5 Experiments

In this section, we report the experimental results of the proposed tracker quantitatively and qualitatively.

5.1 Experimental Setting

In this section, we evaluate the proposed tracker using fifteen sequences which covers different kinds of challenging factors including cluttered background, illumination variations, partial occlusion, pose variation, etc. We compared the proposed tracking algorithm with other ten state-of-the-art trackers which include discriminative multiple feature trackers: OAB [Grabner and Bischof, 2006], SemiT [Grabner *et al.*, 2008], MIL [Babenko *et al.*, 2011], generative trackers which explicitly model the noise/outliers: L1T [Mei and Ling, 2011], MTT [Zhang *et al.*, 2013b], feature learning-based trackers: IVT [Ross *et al.*, 2008], and other state-of-the-art methods: CT [Zhang *et al.*, 2012], DFT [Sevilla-Lara and Learned-Miller, 2012], LOT [Oron *et al.*, 2012],

Table 1: Center Location Error. The best three results are shown in red, green and blue.

Sequence	STRUCK	LOT	OAB	SemiT	L1T	MTT	MIL	IVT	DFT	CT	Proposed Method
Crossing	**3**	36.6	4.6	3.8	62.9	56.5	**2.9**	**2**	21.8	3.2	**3**
Car11	**1.1**	31.4	2.9	2.5	**1.4**	**1.8**	44	8.9	59	119.3	2.7
Car4	**8.8**	167.4	95.9	152.7	77	22.6	50.2	**2.4**	62.5	86	**9.2**
Animal	**5.3**	64.9	**6.8**	61.6	24.1	19.1	100.2	182.6	98.3	246.2	**7**
DavidIndoor	42.3	24.4	22.2	46.8	13.8	32.5	17.3	**4.6**	42.5	**10.9**	**11.8**
Trellis	**7.1**	48	98.7	69.5	62.3	69	72	120	44.3	**42.2**	**7.5**
Skating1	**83.2**	110.4	**43.2**	288.3	159.2	293.8	139.6	247.6	174.6	150.9	**10.2**
Mountain-bike	9	25.1	12.2	240.7	8.5	**7.3**	72.6	**8.1**	154.7	213.7	**8.1**
Subway	**4.1**	5	113.2	105.6	148.1	166.3	8	131.3	**3.2**	11.6	**4.9**
Faceocc	19.3	35.2	25.1	75.1	**17.8**	21.6	30.1	**18.9**	23.8	26.3	**18.4**
Faceocc2	**6**	15	19.8	50.9	12.9	10.2	13.9	**7.8**	**8.1**	18.9	9.3
Walking2	10.7	64.5	28.7	11.9	**4.5**	**3.5**	60.2	**2**	28.5	58.3	7.3
Bolt	399.3	**12.1**	**256.5**	363.7	409	409.2	394.1	397.5	367.7	364.2	**8.9**
Shaking	30.1	82.7	191.6	275	109.7	97.3	**24**	86.1	**26.3**	79.4	**9.7**
DavidOutdoor	106.7	**9.7**	83.5	237	89.8	341.8	**30**	51.9	51	88.6	**7.2**

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

Sequence	STRUCK	LOT	OAB	SemiT	L1T	MTT	MIL	IVT	DFT	CT	Proposed Method
Crossing	0.96	0.33	0.84	0.88	0.25	0.23	**0.99**	0.24	0.66	**0.99**	**0.99**
Car11	**1**	0.56	0.95	0.93	**1**	**1**	0.18	0.7	0.34	0	**1**
Car4	**0.41**	0.05	0.28	0.25	0.3	0.32	0.28	**1**	0.26	0.28	**0.45**
Animal	**1**	0.06	**0.96**	0.86	0.72	0.87	0.13	0.03	0.31	0.04	**0.99**
DavidIndoor	0.24	0.16	0.15	0.21	**0.7**	0.29	0.25	**0.8**	0.24	0.44	**0.59**
Trellis	**0.78**	0.32	0.18	0.2	0.16	0.2	0.24	0.32	**0.52**	0.36	**0.81**
Skating1	**0.37**	**0.28**	**0.34**	0.08	0.13	0.14	0.1	0.08	0.16	0.11	**0.28**
Mountain-bike	0.86	0.7	0.92	0.29	0.93	**0.97**	0.58	**0.99**	0.35	0.17	**0.97**
Subway	**0.94**	0.7	0.22	0.38	0.23	0.08	0.81	0.21	**0.99**	0.8	**0.87**
Faceocc	**1**	0.3	0.91	0.71	**1**	**1**	0.77	0.98	0.81	0.86	**1**
Faceocc2	**1**	0.35	0.76	0.56	0.81	0.9	0.94	0.92	**1**	0.76	**0.99**
Walking2	0.46	0.39	0.4	0.38	**0.98**	**0.99**	0.39	**1**	0.39	0.39	0.44
Bolt	0.02	**0.55**	0.04	**0.07**	0.01	0.01	0.01	0.01	0.04	0.01	**0.75**
Shaking	0.17	0.08	0.01	0.01	0.04	0.01	**0.23**	0.01	**0.82**	0.04	**0.96**
DavidOutdoor	0.34	**0.95**	0.34	0.18	0.46	0.1	0.69	0.64	**0.75**	0.35	**0.93**

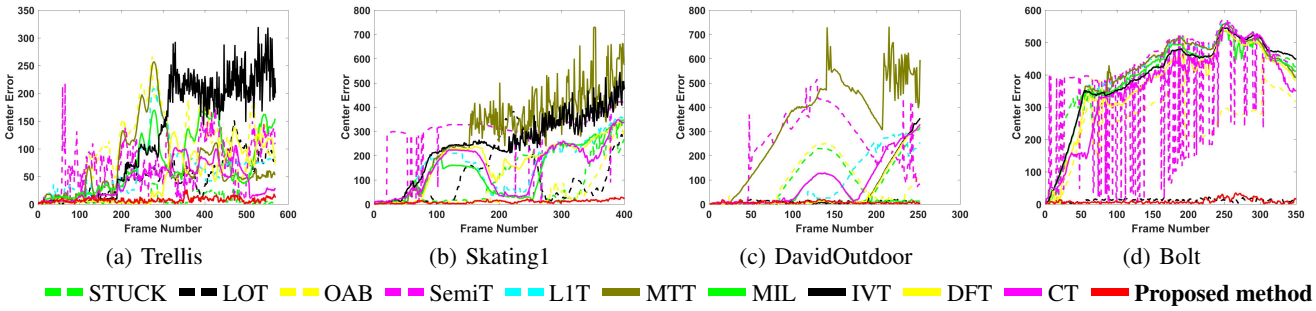

Figure 1: Quantitative frame-by-frame comparison of 11 trackers on 4 Challenging videos in terms of center location error

STRUCK [Hare *et al.*, 2011]. We use the source codes provided by the authors of these papers and set them to be with the same initialization parameters for fair comparison.

We empirically set α_1, α_2, λ_1, and λ_2 to be 0.25, 0.25, 0.1, 0.01, respectively. All $\theta_k(k = 1, ..., K)$ are set to be 1. The training samples $Y^k(k = 1, ..., K)$ consists of the tracking results of the initial 5 and recent 10 frames, and 10 background samples in the current frame. We implement the ℓ_1 tracker [Mei and Ling, 2011] with multiple features to track the target in the initial 15 frames to get target samples for feature learning. We use HOG [Dalal and Triggs, 2005] as global features and covariance descriptors [Tuzel *et al.*, 2006] with log-Euclidean metric [Arsigny *et al.*, 2006] as local features which describe the 2-by-2 non-overlapping parts. So totally 5 kinds of raw visual features are used.

5.2 Experimental results

We adopt two metrics: center location error and success rate for quantitative comparison. The center location error is the Euclidean distance between the center of bounding box and

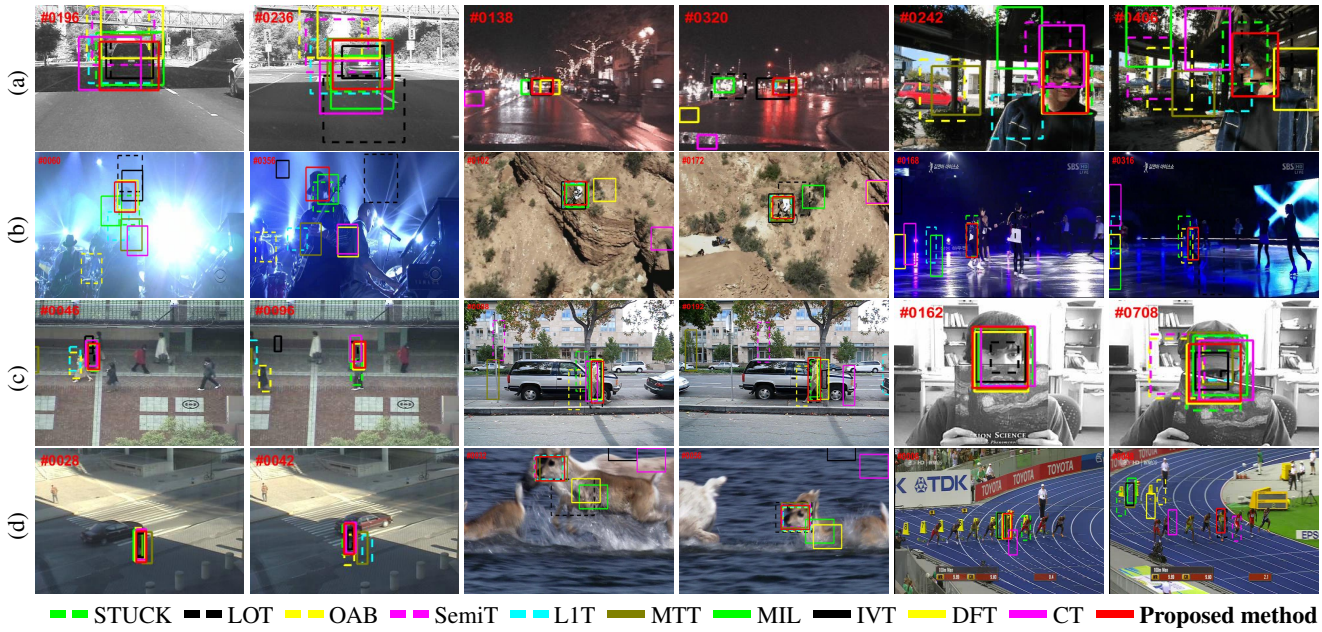

Figure 2: Qualitative results on some typical frames including some challenging factors (video name). (a) Illumination (*Car4, Car11, Trellis*). (b) Pose (*Shaking, Mountain-bike, Skating1*). (c) Occlusion (*Subway, DavidOutdoor, Faceocc2*). (d) Cluttered background (*Crossing, Animal, Bolt*)

the ground truth. The VOC overlapping rate is defined as $\frac{area(ROI_T \bigcap ROI_G)}{area(ROI_T \bigcup ROI_G)}$, where ROI_T and ROI_G are the bounding boxes of the tracker and ground-truth. The tracking result of each frame is considered as a success if the overlapping rate is larger than 0.5.

Tables 1 and 2 record the center location error and the success rate on the 15 videos, respectively. The results show that the proposed tracker outperforms other compared trackers on most videos in terms of both two evaluation metrics such that the center location error of the proposed method rank in top three on 12 videos while the success rate ranks in top three on 14 videos. Particularly, the proposed method demonstrates its superior performance in videos which cover cluttered background (e.g. *Trellis, Crossing, Bolt*), occlusion (e.g. *DavidOutdoor, Faceocc*), and illumination (e.g. *Shaking, Skating1*). This is because the feature learning is performed by simultaneously removing contaminated features and imposing discriminability which enables the appearance model to be less sensitive to target samples contaminated by occlusion and large illumination variation and more discriminative under cluttered background. Figure 1 shows the quantitative frame-by-frame comparison result on some challenging videos, i.e. *Trellis, Skating1, DavidOutdoor* and *Bolt*. We can see that compared with most of other trackers, the proposed tracker can maintain a relatively low tracking error throughout these videos. We can also see that although the proposed tracker is able to track the target throughout the videos *FaceOcc2* and *Walking2* which encounter in-plane rotation and large scale change, respectively, it does not achieve good quantitative results. This is mainly because the rotation and scale state of the tracked target is not well modeled by the particle filter used in the proposed tracker. Since we adopt the sparse representation scheme for target representation and the iterative algorithm for optimization, the proposed tracker can not run in real-time speed. The running time is about 3 frames per second.

Figure 2 illustrates some qualitative results on some typical frames including cluttered background and appearance variations caused by illumination, pose, and occlusion. By properly fusing the learned features from multiple visual cues, the proposed tracker is more stable under some large illumination variations (e.g. *Car4*#196, *Trellis*#242). And explicitly modeling the discriminability of the learned features facilitates the resistance to cluttered background (e.g. *Shaking*#60, *Mountain-bike*#172). Removing contaminated/corrupted feature for appearance modeling enhances the robustness to occlusion (e.g. *DavidOutdoor*#88, *Faceocc2*#162).

6 Conclusion

In this paper, we proposed a novel feature learning framework called robust joint discriminative feature learning for visual tracking with multiple features. By removing the corrupted/contaminated features, introducing the discriminabilities and explicitly modeling the consistency and complementarity in the discriminabilities of multiple visual cues for feature learning in an optimal unified framework, the proposed framework is able to jointly exploit the representation abilities and discriminabilities from multiple visual cues for appearance modeling with multiple features. Extensive comparison experiments with other ten state-of-the-art trackers show its effectiveness and superior performance.

Acknowledgements

This work was supported in part by Hong Kong RGC General Research Fund HKBU 212313 and HKBU 12202514, the Natural Science Foundation of China (No. 61300111) and the Hong Kong Scholars Program (No. XJ2013030). The authors would like to thank the anonymous reviewers for their suggestions on improving the quality of this paper.

References

[Arsigny *et al.*, 2006] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Analysis Applications*, 29(1):328–347, 2006.

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

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

[Candès *et al.*, 2011] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

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

[Fan *et al.*, 2014] B. Fan, Y. Du, H. Gao, and B. Wang. Online discriminative dictionary learning via label information for multi task object tracking. In *Proc. ICME*, pages 1–6, 2014.

[Grabner and Bischof, 2006] H. Grabner and H. Bischof. On-line boosting and vision. In *Proc. CVPR*, pages 260–267, 2006.

[Grabner *et al.*, 2008] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *Proc. ECCV*, pages 234–247, 2008.

[Grabner *et al.*, 2010] H. Grabner, J. Matas, L. J. Van Gool, and P. C. Cattin. Tracking the invisible: Learning where the object might be. In *Proc. CVPR*, pages 1285–1292, 2010.

[Hare *et al.*, 2011] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *Proc. ICCV*, pages 263–270, 2011.

[Hong *et al.*, 2013] Z. Hong, X. Mei, D. Prokhorov, and D. Tao. Tracking via robust multi-task multi-view joint sparse representation. In *Proc. ICCV*, pages 649–656, 2013.

[Hu *et al.*, 2015a] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *Proc. CVPR*, pages 5344–5352, 2015.

[Hu *et al.*, 2015b] W. Hu, W. Li, X. Zhang, and S. Maybank. Single and multiple object tracking using a multi-feature joint sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(4):816–833, 2015.

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

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

[Li *et al.*, 2014] H. Li, Y. Li, and F. Porikli. Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking. In *Proc. BMVC*, 2014.

[Liu and Yao, 1999] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Netw.*, 12(10):1399–1404, 1999.

[Liu *et al.*, 2014] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, and H. Lu. Partially shared latent factor learning with multiview data. *IEEE Trans. Neural. Netw. Learn. Syst.*, 2014.

[Mei and Ling, 2011] X. Mei and H. Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2259–2272, 2011.

[Mei *et al.*, 2015] X. Mei, Z. Hong, D. Prokhorov, and D. Tao. Robust multitask multiview tracking in videos. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(11):2874–2890, Nov 2015.

[Oron *et al.*, 2012] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan. Locally orderless tracking. In *CVPR*, pages 1940–1947, 2012.

[Ross *et al.*, 2008] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *Int. J. Comput. Vis.,*, 77(1-3):125–141, 2008.

[Sevilla-Lara and Learned-Miller, 2012] L. Sevilla-Lara and E. G. Learned-Miller. Distribution fields for tracking. In *Proc. CVPR*, pages 1910–1917, 2012.

[Tuzel *et al.*, 2006] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. ECCV*, pages 589–600, 2006.

[Wang *et al.*, 2015] A. Wang, J. Cai, J. Lu, and T.-J. Cham. MMSS: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *Proc. ICCV*, pages 1125–1133, 2015.

[Yang *et al.*, 2012] M. Yang, D. Zhang, and S. Wang. Relaxed collaborative representation for pattern classification. In *Proc. CVPR*, pages 2224–2231, 2012.

[Yu *et al.*, 2008] Q. Yu, T. B. Dinh, and G. Medioni. Online tracking and reacquisition using co-trained generative and discriminative trackers. In *Proc. ECCV*, pages 678–691. 2008.

[Yu *et al.*, 2013] G.-X. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Zhang. Protein function prediction by integrating multiple kernels. In *Proc. IJCAI*, pages 1869–1875, 2013.

[Zhang *et al.*, 2012] K. Zhang, L. Zhang, and M.-H. Yang. Real compress tracking. In *Proc. ECCV*, pages 864–877, 2012.

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

[Zhang *et al.*, 2013b] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Visual tracking using structured multi-task learning. *Int. J. Comput. Vis.*, 101(2):367–383, 2013.

[Zhang *et al.*, 2015a] S. Zhang, S. Kasiviswanathan, P. C. Yuen, and M. Harandi. Online dictionary learning on symmetric positive definite manifolds with vision applications. In *Proc. AAAI*, pages 3165–3173, 2015.

[Zhang *et al.*, 2015b] X. Zhang, W. Li, M. Fan, D. Wang, and X. Ye. Multi-modality tracker aggregation: From generative to discriminative. In *Proc. IJCAI*, pages 1937–1943, 2015.

[Zhang *et al.*, 2016] S. Zhang, X. Lan, Y. Qi, and P. C. Yuen. Robust visual tracking via basis matching. *IEEE Trans. Circuits Syst. Video Techn.*, 2016. DOI:10.1109/TCSVT.2015.2477936.