Adaptive Weight Collaborative Complementary Learning for Robust Visual Tracking

Benxuan Wang[1], Jun Kong[1,2*], Min Jiang[1], Jianyu Shen[1], Tianshan Liu[1], Xiaofeng Gu[1]
[1] Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, 214122, P. R. China
[e-mail: kongjun@jiangnan.edu.cn]
[2] College of Electrical Engineering, Xinjiang University
Urumqi, 830047, P. R. China
*Corresponding author: Jun Kong

Received December 7, 2017; revised April 26, 2018; accepted September 17, 2018; published January 31, 2019

Abstract

Discriminative correlation filter (DCF) based tracking algorithms have recently shown impressive performance on benchmark datasets. However, amount of recent researches are vulnerable to heavy occlusions, irregular deformations and so on. In this paper, we intend to solve these problems and handle the contradiction between accuracy and real-time in the framework of tracking-by-detection. Firstly, we propose an innovative strategy to combine the template and color-based models instead of a simple linear superposition and rely on the strengths of both to promote the accuracy. Secondly, to enhance the discriminative power of the learned template model, the spatial regularization is introduced in the learning stage to penalize the objective boundary information corresponding to features in the background. Thirdly, we utilize a discriminative multi-scale estimate method to solve the problem of scale variations. Finally, we research strategies to limit the computational complexity of our tracker. Abundant experiments demonstrate that our tracker performs superiorly against several advanced algorithms on both the OTB2013 and OTB2015 datasets while maintaining the high frame rates.

Keywords: Visual tracking, correlation filter, complementary learning, adaptive weight, collaborative model

This work was partially supported by the National Natural Science Foundation of China (61362030, 61201429), China Postdoctoral Science Foundation (2015M581720, 2016M600360), Jiangsu Postdoctoral Science Foundation (1601216C), Scientific and Technological Aid Program of Xinjiang(2017E0279).

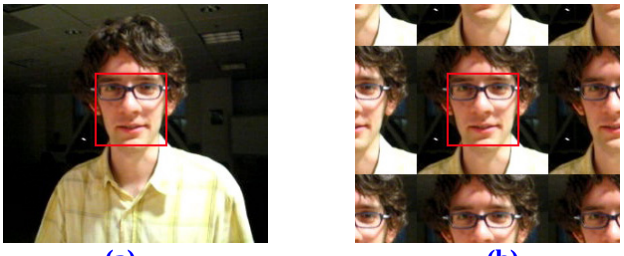
http://doi.org/10.3837/tiis.2019.01.018 **ISSN : 1976-7277**

1. Introduction

Visual object tracking remains a classical problem in computer vision and enjoys a wide popularity recently with numerous applications, such as driverless vehicles, robotics, surveillance, human motion analyses and human-machine interactions [1]. The most general scenario of visual tracking is to figure out the general trajectory of a single target throughout the image sequences, with its location obtained from the first frame. Despite substantial progress in recent years, visual object tracking still remains a challenging problem in computer vision due to several factors which from both the background and the target itself, such as occlusions, fast motion, deformations and illumination variations [2]. At the same time, the increasing number of background patches can drastically degrade the robustness of such trackers against occlusions, and eventually increases the risk of tracking drift specifically when the target and background possess similar visual cues. Moreover, for the online nature of tracking, an ideal tracker should be accurate and robust under the demanding computational constraints of real-time vision systems.

In recent years, a group of correlation filter (CF) based trackers have attracted extensive attention due to continuous performance improvements on tracking benchmarks. These methods obtain the approximate dense sampling by performing the circular sliding window operation on a set of training images and utilize the fast Fourier transform (FFT) to insure the computational efficiency at both learning and detection stages [3]. And that is the major reasons behind the success of this tracking paradigm. Initially, Bolme *et al.* [4] introduced the correlation filter into the tracking application, and their extension achieved state-of-the-art performance on the benchmark videos.

Despite substantial progress in recent years, CF trackers still have several drawbacks. The CF based trackers struggle against kinds of adversity, e.g. fast target motion, target deformations and occlusions. These harsh environmental factors may cause the lack of real negative training examples and lead to train an over-fitted filter. Also, the standard CF formula is based on the periodic assumption by utilizing a circular correlation. Due to the periodic assumption of the training and detection samples, one deficiency of CF is the harmful boundary effects (see **Fig. 1**). Such situation brings about an endogenous inclusion of substantial background information within the target. It can severely reduce the discrimination of the trackers, resulting in inferior results.



(a) **(b)**

Fig. 1. Original image (a) and the periodic assumption (b) under the frame of standard DCF methods

In this work, we consider the problems mentioned above and the main contributions of our work can be summed up as follows. First, we propose a novel collaborative method to take advantage of the strengths of both DCF based model and color-based model. The robust DCF based model we use for obtaining template scores is aimed to distinguish the target from the background with effect. The color-based model we use for obtaining histogram scores is aimed to better cope with occlusion and deformation. To combine these two models adaptively, two criterions and high-confidence adaptation of weights are utilized. Then we can obtain the final translation response. Second, in the learning stage of the DCF-based tracker, we formulate a spatial regularization component which is used to penalize the background information and promote the discriminative power of the tracker. Third, to surmount the deficiency that DCFs tend to be useless when the size of the target is changing, we utilize the discriminative multi-scale estimate method and also take computational complexity into account. The flowchart of our work is shown in **Fig. 2**. To better verify the performance of our tracker, we perform a comprehensive evaluation on the Online Tracking Benchmark (OTB) datasets [5, 6]. Our approach improves the baseline DCF both in the aspect of accuracy and speed.

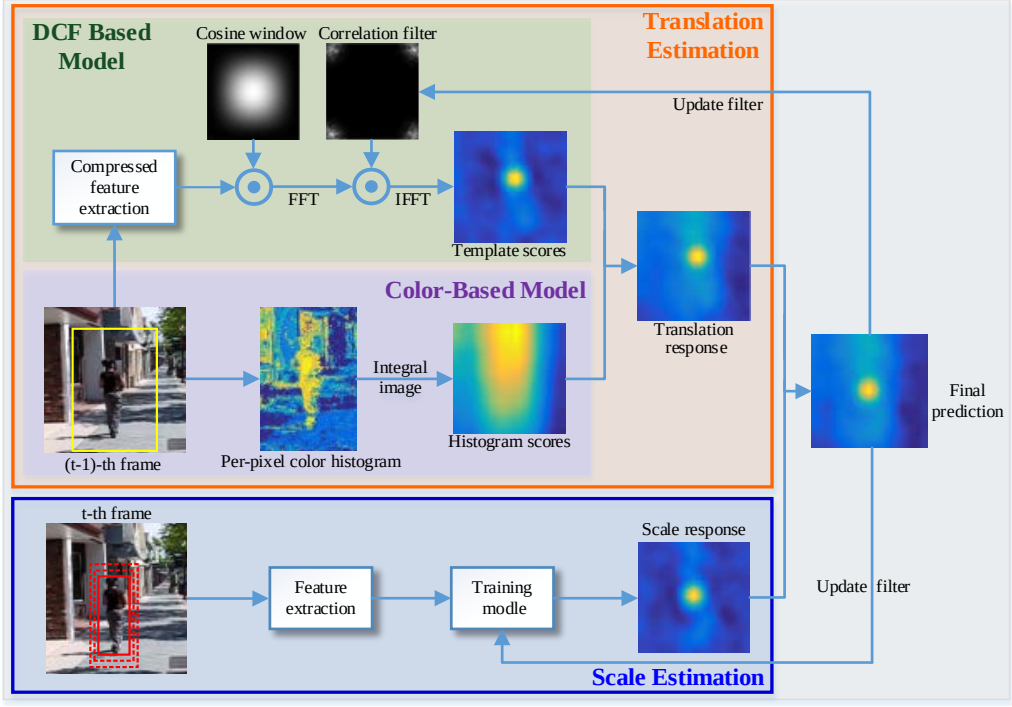


Fig. 2. Flowchart of our tracking algorithm

The remainder of this paper is organized as follows. In Section 2, we give an introduction of the related work and an overview of our work. In Section 3, we minutely introduce our method from two main parts: collaborative translation estimate and fast

discriminative multi-scale estimate. The explanations and results of the comparison experiments is given in Section 4. And conclusions are finally provided in Section 5.

2. Related work

2.1 Current Tracking Algorithms

Existing tracking approaches tend to explore an effective algorithm which can be designed to be either generative [7] or discriminative [8, 9] models. The generative appearance models are aimed at representing the target with statistical models or templates and perform tracking by searching the best-matching windows. The discriminative methods employ machine learning techniques to design a robust classifier or filter to detect the target from backgrounds, and establish an optimal mechanism to update the model at each frame.

Correlation filters, as a discriminative method, can generate the high response for the interested target with the low response to the background. These DCF based trackers are adequate for the tasks of target location, but the tough challenging environment for online tracking is still an open question. To perform better on online object tracking task, many recent advancements in DCF based trackers are driven by the use of non-linear kernels [10], multi-dimensional features [11], robust scale estimation [12, 13], long-term memory components [14], complicated training models [15, 16] and reducing boundary effects [17]. Henriques *et al.* [10] proposed kernel correlation filter (KCF) and multi-channel features by solving a simple rigid regression problem over training data in the dual form. Multi-dimensional features, such as HOG [18], Color-Names [11] and the attentional features [19], depend much on harsh approximations of the standard loss function and lead to a suboptimal solution. Danelljan *et al.* [12] investigated robust scale estimation problem by learning discriminative correlation filter (DCF) based on incorporating a multi-scale template. Bertinetto *et al.* [20] combined two image representations to learn a tracking model which is robust to both deformations and illumination changes. Zhang *et al.* [21] proposed a new model to complement the strength of multi-task correlation filter and particle filter, which get favorable performance on multiple sequences. Danelljan *et al.* [17] introduced the method which penalizes CF coefficients in the learning depending on their spatial locations and achieves excellent tracking accuracy. Besides, Danelljan *et al.* [22] tackled the key causes behind the problems of computational complexity and over-fitting to improve both speed and performance of the tracker. Others introduced deep leaning method into DCF based trackers and investigated sophisticated training models [23, 24] which might lead to the heavy computation cost for online tracking applications.

2.2 Standard DCF Tracker

Our tracking approach is built upon the discriminative correlation filters (DCF). In this part, more details can be found in [10]. Here, we introduce a multi-channel correlation filter f learning from a set of training examples. Each training sample x contains a

d-dimensional feature map extracted from the sample region. Thus we obtain a feature vector $x(m,n) \in \mathbb{R}^d$ with dimension d. x^l denotes x with the feature layer $l \in \{1, \cdots, d\}$. The desired output y is a 2D Gaussian function. The desired filter f is composed of a $m \times n$ convolution filter f^l for every feature layer. We obtain the filter function by minimizing the squared error of the correlation responses to y on the training samples x,

$$\varepsilon = \left\| \sum_{l=1}^{d} x^l \star f^l - y \right\|^2 + \lambda \sum_{l=1}^{d} \left\| f^l \right\|^2 \quad (1)$$

Here, λ is the weight parameter of the regularization term and $\star$ denotes circular convolution. Then the convolution response of the filter f on sample x can be given by

$$\sum_{l=1}^{d} x^l \star f^l \quad (2)$$

As a linear function problem, Eq. (1) can be transformed to the Fourier domain by utilizing discrete Fourier Transformed (DFT) under the frame of Parseval's formula. We use z to represent the new feature map extracted from an image region. The full detection response over all locations is given by the convolution properties of the DFT,

$$\mathcal{F}^{-1}\left(\sum_{l=1}^{d} \hat{z}^l \cdot \hat{f}^l \right) \quad (3)$$

Here, the hat denotes the DFT of a term and $\mathcal{F}^{-1}$ denotes the inverse DFT. The $\cdot$ symbol represents element-wise multiplication computation. The computation complexity may decline to $\mathcal{O}(dmn \log mn)$ from $\mathcal{O}(dm^2n^2)$ with the help of FFT. To be more efficient, a sliding-window-like method, used in the learning stage of DCF, is adopted to collect many translated samples around the target by cyclic shifts without much extra computation. But, it is noteworthy that the calculation in Eq. (2) corresponds to generate the periodic extension of the sample x and brings in unwanted periodic boundary effects (see **Fig. 1**).

3. Proposed approach

3.1 Robust DCF Based Tracking Model

3.1.1 Feature dimensionality reduction

As all other DCF based trackers, the computational cost of our algorithm mainly caused by the FFT operations. Since the training and detection stages require one FFT operation on every feature layer, the computation scales of FFT have a significant linear correlation with the feature dimension. To simplify the computations, we introduce a feature dimensionality reduction method based on principal component analysis (PCA) to speed up the optimization process. The update process of target template is $u_t = (1 - \eta)u_{t-1} + \eta u_t$, where η is a learning rate parameter and u_t denotes the target template of t-th frame. The learned template u_t is used to construct a projection matrix M_t which defines the low-dimensional feature subspace. The low-dimensional training sample $\hat{x}_t' = \mathcal{F}(M_t x_t)$ and the low-dimensional target template $\hat{u}_t' = \mathcal{F}(M_t u_t)$ can be

used for updating the filter. The detection response of the test sample z_t is acquired similar to Eq. (3), with the filter on the low-dimensional sample $\hat{z}_t' = \mathcal{F}(M_{t-1}z_t)$.

3.1.2 Train model with information regularization

For the unwanted boundary effects [17], rectangular initialization bounding boxes in learning stages always include some background information captured from the model. And in the best of circumstances, we can learn a filter that has a high response for the target patch and a near-zero response for other patches. Therefore, we tend to achieve this by adding the boundary patches as a spatial regularization component to the standard formulation. In every frame, we sample boundary patches $c_{p(p=1,\cdots,q)}$ consists of the feature map extracted from the background. The optimization problem is expressed as,

$$\varepsilon_f = \left\|\sum_{l=1}^{d} x^l \star f^l - y\right\|^2 + \lambda \sum_{l=1}^{d} \left\|f^l\right\|^2 + \gamma \left\|\sum_{l=1}^{d} c^l \star f^l\right\|^2 \tag{4}$$

Here, the third term in Eq. (4) is the spatial penalty term with a weight parameter γ. As a result, the target patch is regressed to y like in the standard formulation Eq. (1), while the negative boundary patches are regressed to zeros controlled by the parameter γ. The primal objective function ε_f can be rewritten by stacking the boundary patches below the target image patch to form a new training data matrix $g \in \mathbb{R}^{(q+1)m \times n}$:

$$\varepsilon_f = \|g \star f - y'\|^2 + \lambda \|f\|^2 \tag{5}$$

where $g = \left\{x; \sqrt{\gamma}c_1; \cdots; \sqrt{\gamma}c_q\right\}$ and the new regression target $y' = \{y; 0; \cdots; 0\}$. Since ε_f is convex, it can be minimized by setting the gradient to zero. We can obtain:

$$f = (g^T g + \lambda I)^{-1} g^T y' \tag{6}$$

Similar to all the DCF tracker, we use the identity for circulant matrices to obtain the following closed-form solution in the Fourier domain.

$$\hat{f} = \frac{\hat{x}^* \cdot \hat{y}}{\hat{x}^* \cdot \hat{x} + \lambda + \gamma \sum_{p=1}^{q} \hat{c}_p^* \cdot \hat{c}_p} \tag{7}$$

Here, the symbol $*$ denotes complex conjugation. Note that the detection formula is just like the standard formulation in Eq. (3). And the solution in the primal domain in Eq. (6) has the same form with the solution of the standard ridge regression problem [10]. We use α to denote the dual conjugate of f with $f = \sum_i \alpha_i x_i$. The solution is given by:

$$\alpha = (gg^T + \lambda I)^{-1} y', \text{ where } \alpha \in \mathbb{R}^{(q+1)m} \tag{8}$$

Note that the detection formula is just like the standard formulation, but g contains the boundary image patches in addition to the target. As a result, the detection formula finally can be rewritten as follows:

$$S_f(z) = \mathcal{F}^{-1}\left(\hat{z} \cdot \hat{x}^* \cdot \hat{\alpha}_0 + \sqrt{\gamma} \sum_{p=1}^{q} \hat{z} \cdot \hat{c}^* \cdot \hat{\alpha}_p\right) \tag{9}$$

3.2 Schemes to Establish Collaborative Translation Model

3.2.1 Color probability distribution based model

One inherent problem of correlation filters is that the rigid template would not adapt to the shape deformation of the target in the course of a sequence. To achieve robustness to deformation, color-histogram based methods [25, 26] were used in plenty of previous object tracking algorithms for the insensitiveness to shape variation. But it is not well enough to distinguish the interested target from the background. In recent years, the only successful method we know named Distractor-Aware Tracker (DAT) [27], which identified distracting regions with similar colors compared to the target in advance to prevent the drifting and preformed competitively in modern benchmarks.

Ideally, to differentiate object feature pixels from similar background feature, we adopt a color-histogram based Bayes classifier h on the sample images. The histogram score is computed from the histogram feature I defined on a finite region Ω:

$$S_h(x) = \beta^T \left(\frac{1}{|\Omega|} \sum_{r \in \Omega} I[r] \right) \tag{10}$$

Here, r denotes image pixels and histogram weight vector is β. The final histogram score can be considered as the average vote by utilizing a single integral image and is invariant to spatial arrangement of its feature image. We obtain the training example from sample windows and the regression target is y. The loss function for color-based model is

$$\varepsilon_h = \frac{1}{|\Omega|} \sum_{r \in \Omega} (\beta^T I[r] - y)^2 \tag{11}$$

We intend to apply linear regression over object regions O and background regions B independently, with the positive example $(v, 1)$ and negative example $(\overline{v}, 0)$.

$$\varepsilon_h = \frac{1}{|O|} \sum_{r \in O} (\beta^T I[r] - 1)^2 + \frac{1}{|B|} \sum_{r \in B} (\beta^T I[r])^2 \tag{12}$$

So, the solution of the ridge regression problem is as follows.

$$\beta_t^l = \frac{\rho^l(O)}{\rho^l(O) + \rho^l(B) + \lambda_h} \tag{13}$$

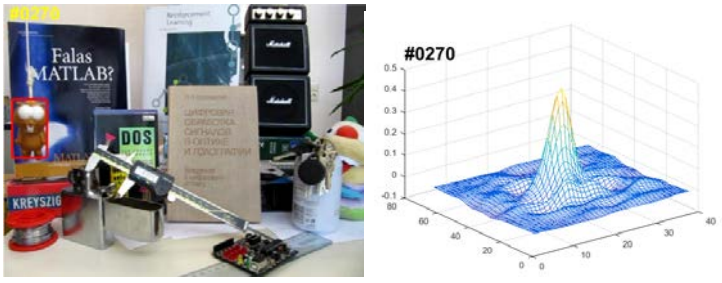
For each feature dimension $l \in \{1, \cdots, d\}$, $N^l(\mathcal{H}) = |\{r \in \mathcal{H} : I[r] \neq 0\}|$ is the number of pixels in the region $\mathcal{H}$ of Ω with the non-zero feature $I[r]$ and $\rho^l(\mathcal{H}) = N^l(\mathcal{H})/|\mathcal{H}|$ is the proportion of the non-zero feature in this region.

3.2.2 Combining multiple estimates

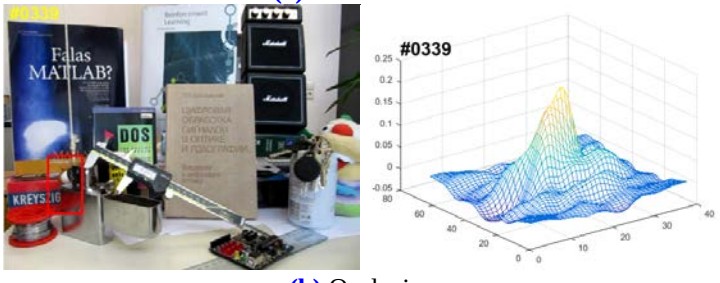
Sometimes color-based model is not enough to distinguish the target from the surrounding background. On the contrary, CF trackers, as a kind of template model, depend on the spatial construction of the target and often fail when the appearance of the target changes rapidly. We tend to search a tracker which can take advantage of both color-based and template models. For that, we propose a score function to combine the correlation filter scores and histogram scores:

$$S(x) = \omega_f S_f(x) + \omega_h S_h(x) \tag{14}$$

Here, we combine the two scores with setting $\omega_f = 1 - \omega_h$, depending on how much we believe in them. In general, most of presented visual tracking algorithms obtain the final score to locate the position of target by searching the response map. The response map reveals the degree of confidence about the tracking results to some extent. The response map should have only one sharp peak and be smooth in all other areas when the detected target in the current frame is extremely matched to the correct target. The sharper the correlation peaks are, the better the location accuracy is. But, if the object is occluded severely or even missing, the whole response map will fluctuate intensely, resulting in a pattern that is significantly different from the normal response map as shown in **Fig. 3**.



(a) No occlusion



(b) Occlusion

Fig. 3. A part of the sequence and their corresponding response maps

So, to better take advantage of the response map pattern and the color histogram, we propose a novel high confidence associative mechanism with two criteria. The first one is the maximum response score P_{max} of the response map. ω_f and ω_h is related to the second criterion called average-peak energy (APE) which is defined as:

$$APE = \frac{|P_{max} - P_{mean}|^2}{\sum_r (P_r - P_{mean})^2} \tag{15}$$

Here, P_{max}, P_{mean} and P_r denote the maximum, mean and the r-th pixel of the correlation response map. APE indicates the fluctuated degree of response maps and the confidence level of the maximum response score which represents detected targets. For sharper peaks and fewer noise, i.e., the target apparently appearing in the detection scope,

APE will close to 1 and far above a predefined threshold. In this situation, the template response map will become relatively smooth except for only one sharp peak and the correlation filter scores would be more confident. ω_f is set to a high value with the decrease of ω_h. Whereas, if the object is deformed or occluded, APE will significantly decrease and approach to 0. In that case, the APE is lower than the predefined threshold and the confidence of correlation filter scores would be decreased. We need more reliance on the color-based models to fine-tune the final tracking results and ω_h will be set to a higher value. And the value of the predefined threshold is based on the historical observations of APE.

Fig. 4 illustrates the comparison of two different associative strategies. Compared with the simple linear superposition, the proposed adaptive associative scheme provides a more flexible method to choose the better combination. For the characteristic of APE, the algorithm can realize the adaptive strength of the two response scores. The experimental results presented in Section 4 demonstrate the effectiveness of the proposed collaborative complementary scheme.

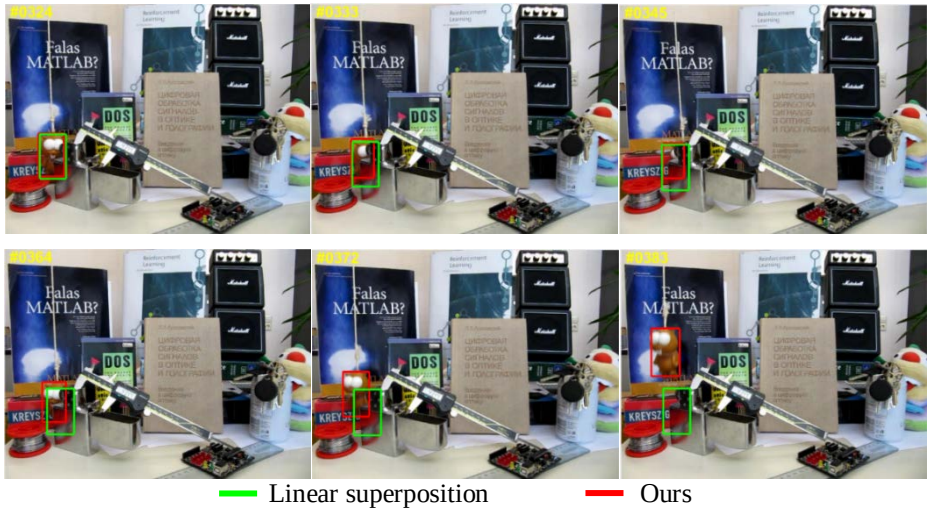


Fig. 4. The comparison of two associative strategies

3.3 Fast Discriminative Multi-Scale Estimate Method

3.3.1 Adaptive multi-scale correlation filter

When tracking the target in a series of images, the size of the target is varying all the time with the change of relative distance. But most of the DCF based trackers cannot deal with that scale variation problem. To solve the problem, we utilize a separate 1- dimensional correlation filter to calculate correlation scores at different scale dimensions and estimates the best scale of target in an image.

We use $Q \times R$ as the target size in the previous frame and $U \times 1$ as the size of the scale filter. At each scale level $e \in \left\{\left\lfloor -\frac{U-1}{2} \right\rfloor, \cdots, \left\lfloor \frac{U}{2} \right\rfloor\right\}$, we extract several image patches of size $a^e Q \times a^e R$ centered around the target in the training sample x_{scale} with the

scale factor a. In this case, the desired correlation output y_{scale} is a 1-dimensional Gaussian. The optimal scale filter can be calculated by minimizing the sum of squared errors, just similar to Eq. (1). The test sample z_{scale} can be obtained same as the training samples x_{scale}. The correlation scores S_{scale} can be computed as:

$$S_{scale}(z_{scale}) = \mathcal{F}^{-1}\left(\frac{\sum_l \hat{v}^* \cdot \hat{z}_{scale}}{\hat{w} + \lambda_{scale}}\right).\ l = 1, \cdots, d \tag{16}$$

Here, v and w are the numerator and denominator of the formula of scale filter f_{scale} which can be updated by Eq. (17).

$$\begin{aligned} \hat{v}_t &= (1-\tau)\hat{v}_{t-1} + \tau\, \hat{y}^*_{scale} \cdot x_{t;scale} \\ \hat{w}_t &= (1-\tau)\hat{w}_{t-1} + \tau \sum_l \hat{x}^*_{t;scale} \cdot x_{t;scale} \end{aligned} \tag{17}$$

where t denotes the t-th frame and τ is a scale learning rate parameter. Typically, the translation filter will be applied in the new frame first, and then the scale filter is applied. For more details in this part, readers can refer to [12].

3.3.2 Dimensionality reduction on scale filter

Because of the feature dimensionality reduction method we used in translation correlation filter, the feature dimensionality of scale filter is larger than the size of low-dimensional training samples (see Section 4.2 for more details). Moreover, the number of scales is more than or equal to the row of the correlation matrix. So that, the template of scale filter can be compressed without too much loss of target information. To compress the scale filter, we need to construct a projection matrix $M_{t;scale}$, and the compressed templates will be used in Eq. (19) to update the scale filter. This process is very similar to the one we noted in Section 2.1. With compression, the computational cost can be effectively decreased by reducing the size of the performed FFTs in the training and detection stages.

3.3.3 Sub-grid interpolation of scale response

For the compressed scale filter we mentioned above, we employ the sub-grid interpolation strategy to make sure that we can use the coarser features for the training and detection samples. The response map are efficiently interpolated with trigonometric polynomial which is suitable for the computed DFT coefficients. On each scale level, we use the sub-grid method respectively. And the scale level with the maximum response will be applied to the update of target location and scale.

3.4 The Framework of the Proposed Tracker

An overview of our proposed method is summarized in **Algorithm 1**.

Algorithm 1

When t-th frame arrives

Inputs:

Target location S_{t-1} and scale $S_{t-1;\text{scale}}$.

Output:

Estimated target location S_t based on the response and scale $S_{t;scale}$.

Translation estimate:

1: Extract training samples x from the last location S_{t-1} and extract its feature map.
2: Calculate model scores $S_{t;f}$ and $S_{t;h}$ via Eq. (9) and Eq. (10), and APE with Eq. (15).
3: Judge $\omega_{t;f}$ and $\omega_{t;h}$ in Eq. (14) by $S_{t;f}$, $S_{t;h}$ and APE, then get the target location S_t from $S(x)$.

Scale estimate:

1: Extract scale samples of different sizes z_{scale} from S_{t-1} and $S_{t-1;\text{scale}}$.
2: Use Eq. (16) to compute the scale correlation responses.
3: Set $S_{t;scale}$ to the target scale that maximize the response.

4. Experiments

4.1 Parameter Setup

Our proposed algorithm is implemented with MATALB2015a on the machine equipped with a core 3.5 GHz with 8GB memory without any parallel framework. Dissimilar to most DCF based trackers, we utilize PCA-HOG features [28] to represent images, by using the Matlab Toolbox. In the frame of the translation model, we experimentally set the regularization parameter to 0.01 for both λ and λ_h. The additional regularization factor γ in Eq. (4) is set to 0.5 and the number of boundary patches we sampled is set to 4. The learning rates are set to $\eta = 0.015$ and $\eta_h = 0.04$ respectively. For the joint scale space filters, the regularization parameter is set to $\lambda_{scale} = 0.01$ and the learning rate is set to $\tau = 0.025$. Through experiments on many sequences containing serious scale variation, we find the optimal solution when we interpolate the number of scales from 17 to 33 using the proposed method and use $a = 1.02$ as the scale factor. The weight ω_f and ω_h is set to 0.7 and 0.3 initially.

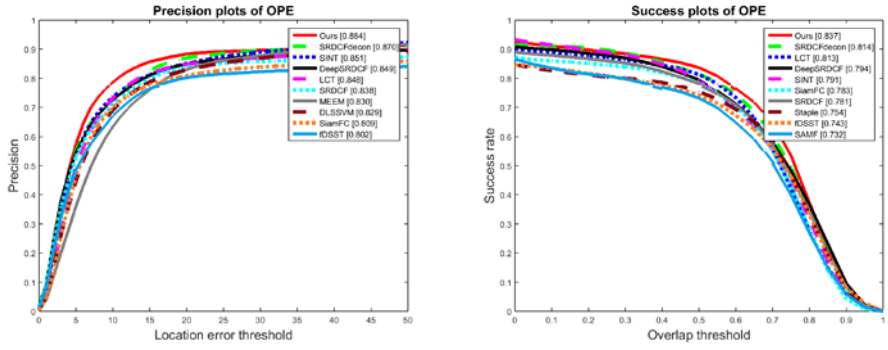
4.2 Evaluation Methodology

OTB2013 [5] and OTB2015 [6] are two popular tracking benchmarks, which contain results of several trackers evaluated on 50 and 100 sequences, respectively. All these sequences are annotated with 11 attributes covering various challenges, including illumination variation (IV), deformation (DEF) , motion blur (MB), occlusion (OCC), scale variation (SV), fast motion (FM), out-of plane rotation (OPR), out-of-view (OV), background clutters (BC), in-plane rotation (IPR) and low resolution (LR).

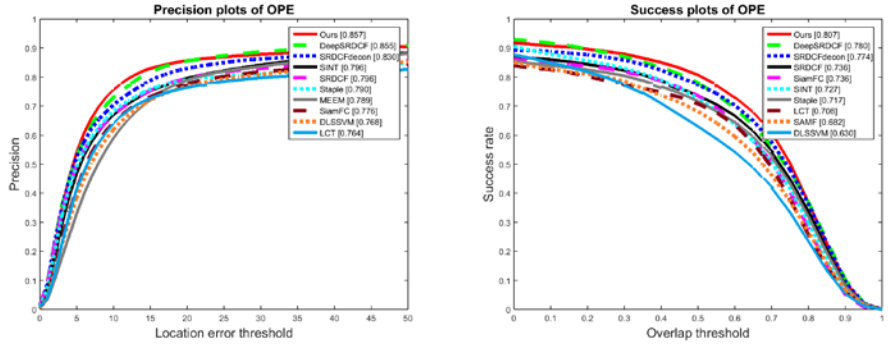
We evaluate our tracker on these benchmarks in comparison with 15 state-of-the-art trackers from three typical categories: (1) correlation filters-based tracking approaches, including DSST [12], LCT [14], SRDCF [17], fDSST [13], KCF [10], SAMF [29], SRDCFdecon [30] and Staple [20]; (2) deep features-based trackers, including SINT [16], SiamFC [15], DLSSVM [31] and DeepSRDCF [32]; and (3) other representative tracking methods, including MEEM [8], TGPR [7] and Struck [9]. All the results of algorithms for comparison mentioned above come from the released source code and results.

Following the protocol in [5, 6], tracking quality is measured by precision rate and success rate. Success rate is defined as the area under curve (AUC) of each success plot, which shows portion of frames with the overlap rates (OR) between predicted and ground

truth bounding box. The OR in each frame is computed from $\frac{\mathcal{R}_c \cap \mathcal{R}_l}{\mathcal{R}_c \cup \mathcal{R}_l}$ with the areas of the predicted regions $\mathcal{R}_c$ and the ground truth regions $\mathcal{R}_l$. Then, for the precision plot, it shows similar statistics on the center location error (CLE). The CLE in each frame can be measured by the Euclidean distance between the centers of the tracking response and the ground truth. For overall performance, we present our results in the one-pass evaluation (OPE) using distance precision rate (DPR) and overlap success rate (OSR) as shown in **Fig. 5**. And DPR is presented at 20 pixels for quantitative comparison, OVR at 0.5. Only the top 10 trackers are displayed to reduce clutter in the graphs.



(a) Comparisons on OTB2013



(b) Comparisons on OTB2015

Fig. 5. Average overall performance on OTB2013 and OTB2015 with distance precision rate (DPR) and overlap success rate (OSR)

Table 1. Comparisons of our tracker with state-of-the-art trackers in terms of distance precision rate (DPR), overlap success rate (OSR) and speed on the OTB2013 and OTB2015 datasets. The best three results are shown in red, green and blue respectively.

	OTB2013			**OTB2015**		
	DPR(%)	**OSR(%)**	**Speed (fps)**	**DPR(%)**	**OSR(%)**	**Speed (fps)**
MEEM	83	69.6	20.8	78.1	62.2	20.8
TGPR	70.5	62.8	1	64.3	53.5	1
Struck	65.6	55.9	10	63.9	51.6	9.8

DSST	74	67	25.9	68.7	60.9	21.9
SAMF	78.5	73.2	18.6	75.8	68.2	16.8
KCF	74	62.3	189.1	70	55.8	183
LCT	84.8	81.3	21.6	76.4	70.8	20.5
SRDCF	83.8	78.1	3.6	79.6	73.6	3.6
Staple	79.3	75.4	44.9	79	71.7	42.9
fDSST	80.2	74.3	61.7	42.7	39.1	58
SRDCFdecon	87	81.4	2.4	83	77.4	2.3
SINT	85.1	79.1	4	79.6	72.7	2.5
SiamFC	80.9	78.3	\	77.6	73.6	\
DLSSVM	82.9	72.4	2	76.8	63	2.1
DeepSRDCF	84.9	79.4	0.2	85.5	78	0.2
Ours	88.4	83.7	46.7	85.7	80.7	45.9

On both the two benchmarks, our tracker performs favorably against all other advanced trackers. Moreover, the speed in **Table 1** shows that our tracker is superior to other advanced trackers in both rates. Among the compared trackers, our tracker provides the best performance with a DPR of 88.4% and an OVR of 83.7% on OTB2013. Compared with SRDCFdecon [30] which ranks 2nd in **Table 1**, our approach performs better compared with its DPR of 87% and OSR of 81.4%. Besides, without the adoption of parallel framework, our tracker (47 fps) is much faster than SRDCFdecon (2 fps). On OTB2015, our tracker achieves a DPR of 85.7% and an OVR of 80.7%, while running at a speed of 46 fps. Though the DeepSRDCF [32] utilizes deep features to represent object appearance, our approach outperforms DeepSRDCF in both DPR and OSR. Besides, our tracker runs in real time while DeepSRDCF does not.

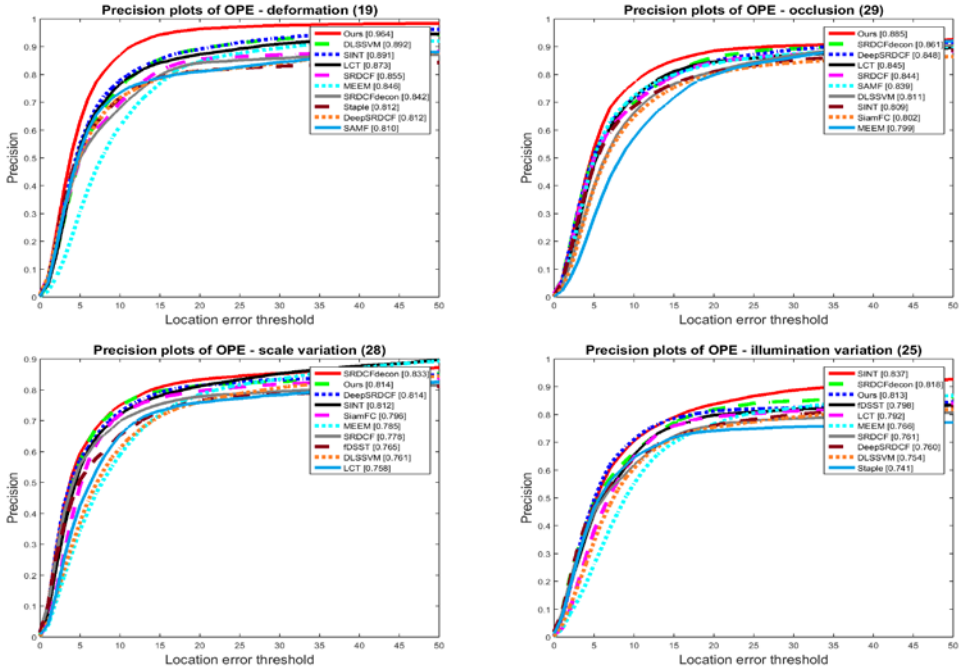
4.3 Precision Evaluation

We further analyze the performance of our tracker under different attributes in OTB2013 [5] and OTB2015 [6]. **Table 2** shows the comparison of our tracker with other top six tracking algorithms for these eleven attributes on OTB2015. And **Fig. 6** illustrates the performance of our tracker with several attributes on both two datasets.

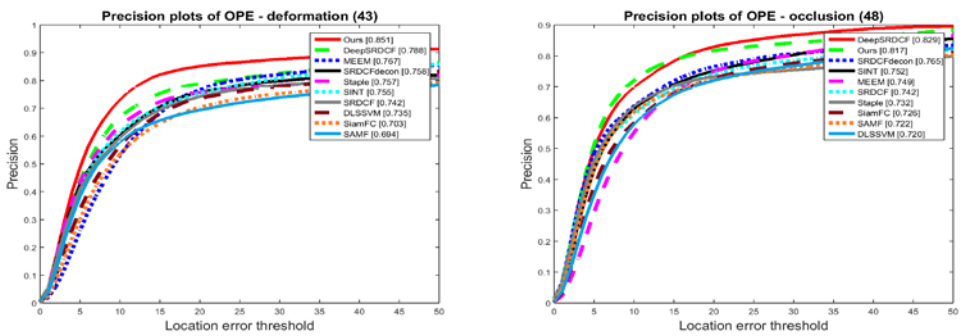
Table 2. Average precision rates (%) of our tracker and other six top trackers on all attributes

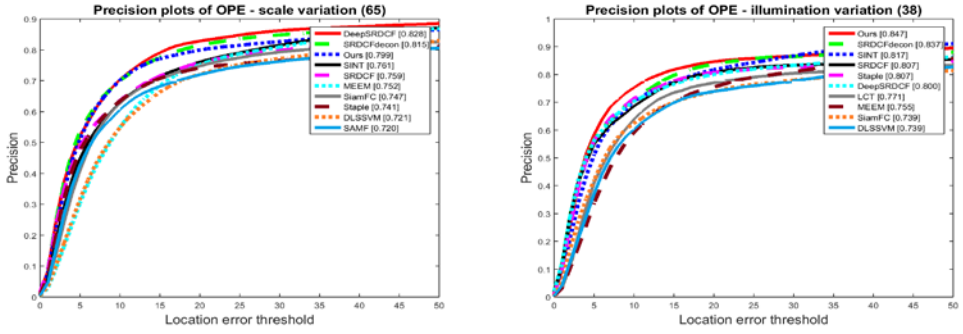
Attribute	Ours	DeepSRDCF	SRDCFdecon	SINT	SRDCF	Staple	MEEM
DEF	85.1	78.8	75.8	75.5	74.2	75.7	76.7
IV	84.7	80.0	83.7	81.7	80.7	80.7	75.5
IPR	83.7	83.0	78.2	84.2	76.3	79.0	81.7
OPR	83.0	84.1	79.8	81.2	75.1	74.7	80.4
OCC	81.7	82.9	76.5	75.2	74.2	73.2	74.9
BC	81.2	85.3	85.3	77.7	79.4	78.5	76.5
SV	79.9	82.8	81.5	76.1	75.9	74.1	75.2
MB	78.3	83.4	82.6	74.7	78.2	72.6	72.2
OV	77.4	79.1	61.9	70.2	61.1	68.0	70.9
FM	77.1	80.1	77.2	73.8	75.4	69.0	72.0
LR	69.5	70.8	64.4	76.8	65.5	63.1	63.1
Overall	85.7	85.5	83.0	79.6	79.6	79.0	78.9

In terms of distance precision rates (DPR), our tracker achieves the top three results on all 11 attributes. As shown in **Fig. 6**, for the attribute of DEF, our approach obtains the most competitive performances and far outshines the second tracker on both two datasets. For the attribute of OCC, our approach achieves the best performance on OTB2013 and gets second on OTB2015. It should be noted that with smaller location error threshold, our method demonstrates its prominent superiority among all the 15 algorithms. In the situations with various challenging attributes, color probability distribution based model plays an important role in improving our tracking performance. Also, the spatial regularization component can further handle with the deformation and occlusion challenge. Compared with other correlation filter-based trackers [14, 17, 20, 29, 30] and MEEM [8], our tracker can better locate the target object in videos and performs more robustly upon most occasions with the help of collaborative translation model.



(a) Comparisons on OTB2013





(b) Comparisons on OTB2015

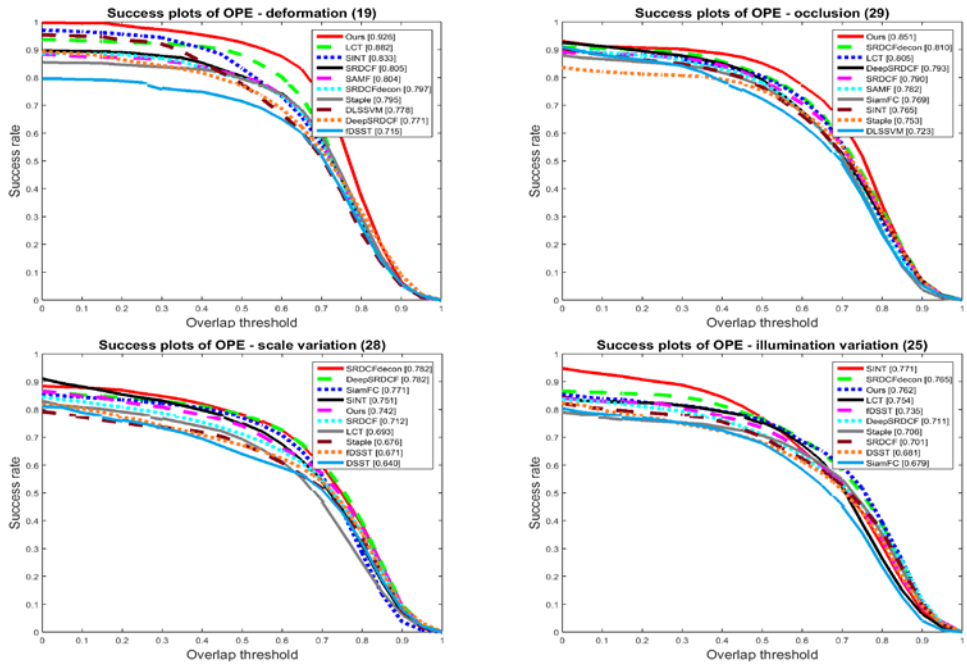
Fig. 6. The precision rates on OTB2013 and OTB2015 with several attributes

4.4 Success Rate

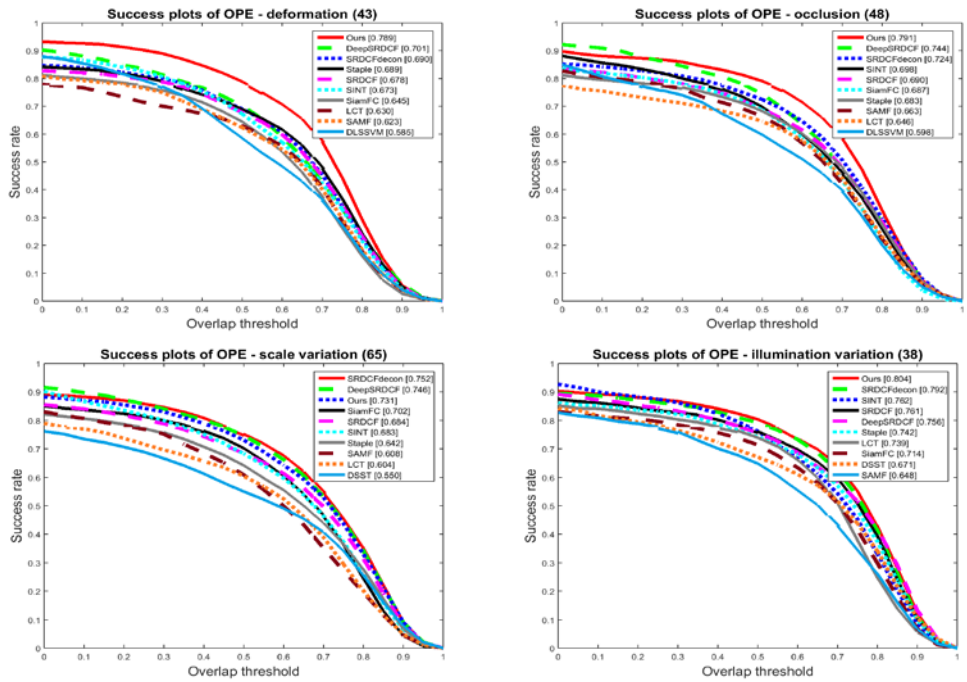
To demonstrate the effectiveness of the proposed multimodal target detection in detail, **Table 3** illustrates the success plots of top seven trackers on OTB2015 dataset under 11 challenging attributes. **Fig. 7** shows that our tracker performs efficiently on most attributes, especially for the variation of deformation and occlusion.

Table 3. Average success rates (%) of our tracker and other six top trackers on all different attributes

Attribute	Ours	DeepSRDCF	SRDCFdecon	SINT	SRDCF	Staple	SiamFC
DEF	78.9	70.1	69.0	67.3	67.8	68.9	64.5
IV	80.4	75.6	79.2	76.2	76.1	74.2	71.4
IPR	75.8	74.0	70.8	73.6	68.3	69.4	71.2
OPR	77.0	74.7	72.0	72.9	67.3	66.3	70.7
OCC	79.1	74.4	72.4	69.8	69.0	68.3	68.7
BC	79.7	76.7	78.5	72.3	71.8	72.8	66.2
SV	73.1	74.6	75.2	68.3	68.4	64.2	70.2
MB	75.9	79.6	81.2	71.4	74.7	67.6	70.3
OV	73.0	67.4	61.6	64.9	56.9	57.2	63.8
FM	73.4	74.2	73.4	68.6	70.7	63.1	69.7
LR	65.7	58.7	61.9	63.7	62.6	49.1	77.7
Overall	80.7	78.0	77.4	72.7	73.6	71.7	73.6



(a) Comparisons on OTB2013



(b) Comparisons on OTB2015

Fig. 7. The success rates on OTB2013 and OTB2015 with several attributes

To demonstrate the effectiveness of our tracker, **Fig. 8** summarizes the qualitative comparisons of our tracker with other six state-of-the-art trackers (DeepSRDCF [32], SRDCFdecon [30], MEEM [8], SRDCF [17], Staple [20] and KCF[10]) on six typical challenging sequences sampled from the benchmark datasets. These six trackers include correlation filters-based trackers, deep features-based trackers and other representative trackers, as well as our baseline.

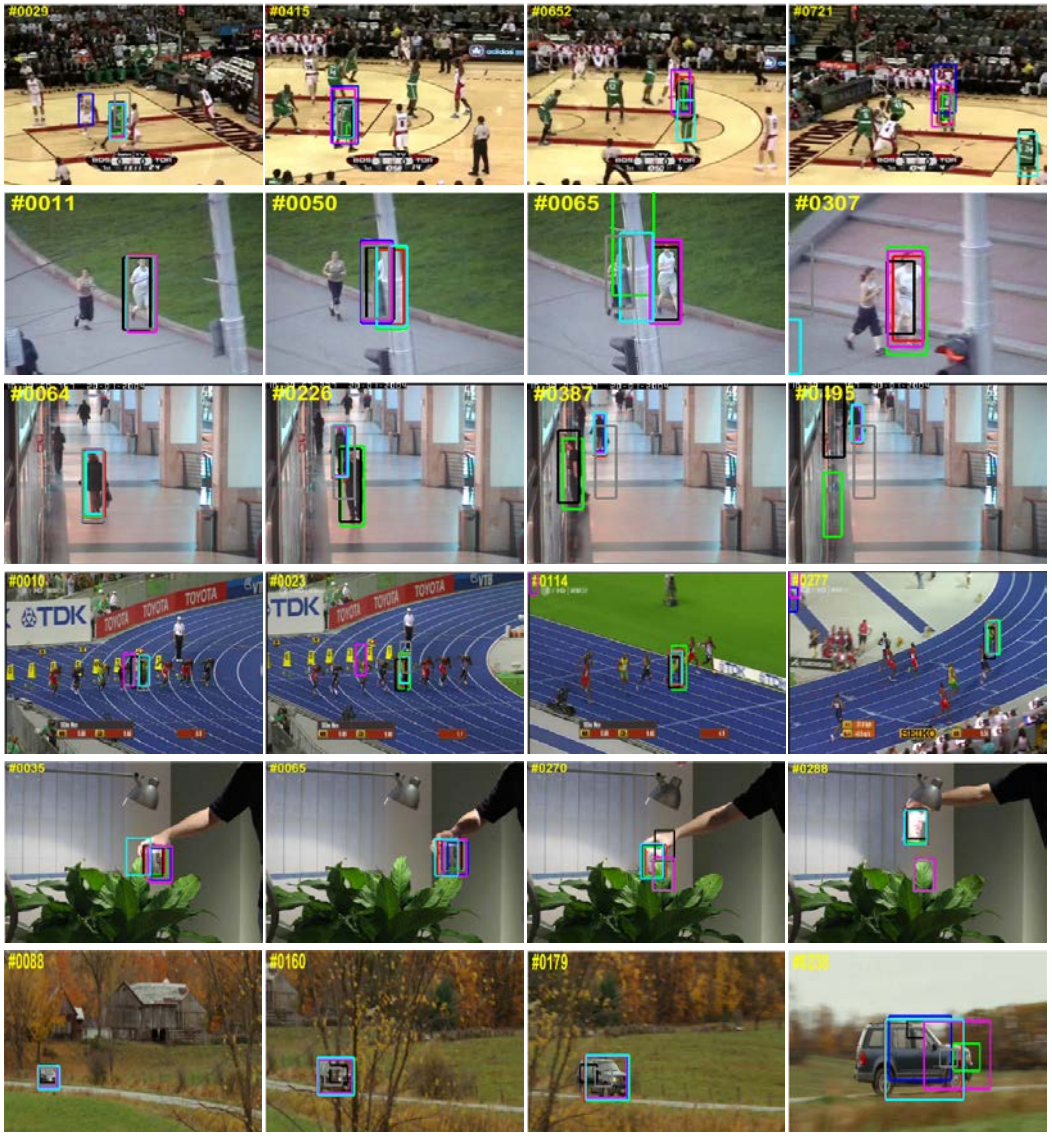


Fig. 8. Qualitative evaluation of the proposed algorithm and other six state-of-the-art trackers on six sequences(from top to bottom: Basketball, Jogging-2, Walking2, Bolt, Coke and CarScale)

The correlation filters-based trackers (KCF [10], SRDCF [17] and SRDCFdecon [30]) perform well in sequences with slight deformation, illumination variation and partial occlusion (Basketball). However, when full occlusion and heavy deformation happens (Coke, Walking2 and Jogging-2), they are limited by the drawback of template models and tend to lose the target. The KCF tracker preforms poor in dealing with the scale variation (CarScale). Staple [20] directly combines the scores of the two complementary models and can solve a portion of the occlusion problems. But, for the ability of re-detect, Staple cannot do well in some complex dynamic environment (Basketball and Jogging-2). DeepSRDCF [32] uses deep features to represent object appearance, and can deal with these cases to some degree. It has the powerful ability of learning and re-detection (Basketball). Nevertheless, it still fails when occlusion happens with other situations such as deformation, rotation and scale variation (Walking2 and CarScale). MEEM [8] utilizes multiple classifiers and chooses the prediction based on the entropy criterion to perform tracking and does well in most cases. However, it may lose the target especially in presence of heavy occlusion and scale variations (Walking2 and CarScale). Our tracker utilizes the high-confidence collaborative strategy which can combine two models efficiently. Compared with the trackers mentioned above, our tracker locates the target object more reliably and can deal robustly with the challenge of full occlusion and heavy deformation notably.

5. Conclusion

In this paper, we propose a novel collaborative object tracking method under the tracking-by-detection framework. The proposed tracker absorbs the strong discriminative ability from collaborative model and speeds up by the compress strategies. To really reduce the background clutter, boundary information punishment mechanism is utilized in the training stage. Otherwise, the high-confidence collaborative strategy can combine the strengths of both DCF based model and color-based model to increase the discriminative power of the learned template model and prevent model drift. Furthermore, the proposed tracking algorithm is equipped with the fast discriminative multi-scale estimate method to cope with the scale variation. Sufficient evaluations on challenging benchmark datasets demonstrate that the proposed tracker can perform well against most advanced methods and maintain the fast speed which can meet the needs of engineering applications.

References

[1] Y. H. Jang, J. K. Suh, K. J. Kim, and Y. J. Choi, "Robust Target Model Update for Mean-shift Tracking with Background Weighted Histogram," *KSII Transactions on Internet & Information Systems*, vol. 10, pp. 181-207, March, 2016. Article (CrossRef Link)

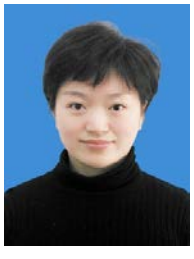
[2] C. Bo and D. Wang, "Robust Online Object Tracking with a Structured Sparse Representation Model," *KSII Transactions on Internet & Information Systems*, vol. 10, no. 5, pp. 2346-2362, May, 2016. Article (CrossRef Link)

[3] X. Zhu, X. Song, X. Chen, Y. Bai, and H. Lu, "Size Aware Correlation Filter Tracking with Adaptive Aspect Ratio Estimation," *KSII Transactions on Internet & Information Systems*, vol. 11, no. 2, pp. 805-825, February, 2017. Article (CrossRef Link)
[4] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2544-2550, June 13-18, 2010. Article (CrossRef Link)
[5] Y. Wu, J. Lim, and M. H. Yang, "Online Object Tracking: A Benchmark," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411-2418, June 23-28, 2013. Article (CrossRef Link)
[6] Y. Wu, J. Lim, and M. H. Yang, "Object Tracking Benchmark," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 9, pp. 1834-1848, September, 2015. Article (CrossRef Link)
[7] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with gaussian processes regression," in *Proc. of European Conference on Computer Vision*, pp. 188-203, September 6-12, 2014. Article (CrossRef Link)
[8] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: robust tracking via multiple experts using entropy minimization," in *Proc. of European Conference on Computer Vision*, pp. 188-203, September 6-12, 2014. Article (CrossRef Link)
[9] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2096-2109, October, 2016. Article (CrossRef Link)
[10] J. F. Henriques, C. Rui, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 3, pp. 583-596, March, 2015. Article (CrossRef Link)
[11] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1090-1097, June 23-28, 2014. Article (CrossRef Link)
[12] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate Scale Estimation for Robust Visual Tracking," in *Proc. of British Machine Vision Conference*, pp. 65.1-65.11, September 1-5, 2014. Article (CrossRef Link)
[13] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative Scale Space Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561-1575, August, 2017. Article (CrossRef Link)
[14] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5388-5396, June 7-13, 2015. Article (CrossRef Link)
[15] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. of European Conference on Computer Vision*, pp. 850-865, October 8-16, 2016. Article (CrossRef Link)
[16] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese instance search for tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1420-1429, June 27-30, 2016. Article (CrossRef Link)
[17] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 4310-4318, December 7-13, 2015. Article (CrossRef Link)

[18] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1030-1037, June 13-18, 2010. Article (CrossRef Link)
[19] J. Choi, H. J. Chang, J. Jeong, Y. Demiris, and Y. C. Jin, "Visual Tracking Using Attention-Modulated Disintegration and Integration," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4321-4330, June 27-30, 2016. Article (CrossRef Link)
[20] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary Learners for Real-Time Tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1401-1409, June 27-30, 2016. Article (CrossRef Link)
[21] T. Zhang, C. Xu, and M. H. Yang, "Learning Multi-task Correlation Particle Filters for Visual Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. Article (CrossRef Link)
[22] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient Convolution Operators for Tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6931-6939, July 21-26, 2017. Article (CrossRef Link)
[23] H. Nam and B. Han, "Learning Multi-domain Convolutional Neural Networks for Visual Tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4293-4302, June 27-30, 2016. Article (CrossRef Link)
[24] J. Gao, T. Zhang, X. Yang, and C. Xu, "Deep Relative Tracking," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1845-1858, April, 2017. Article (CrossRef Link)
[25] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," *European Conference on Computer Vision*, pp. 661-675, May 28-31, 2002. Article (CrossRef Link)
[26] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image and vision computing*, vol. 21, no. 1, pp. 99-110, January, 2003. Article (CrossRef Link)
[27] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2113-2120, June 7-12, 2015. Article (CrossRef Link)
[28] T. Kobayashi, A. Hidaka, and T. Kurita, "Selection of histograms of oriented gradients features for pedestrian detection," in *Proc. of Int. Conference on Neural Information Processing*, pp. 598-607, November 13-16, 2007. Article (CrossRef Link)
[29] Y. Li and J. Zhu, "A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration," in *Proc. of European Conference on Computer Vision*, pp. 254-265, September 6-12, 2014. Article (CrossRef Link)
[30] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1430-1438, June 27-30, 2016. Article (CrossRef Link)
[31] J. Ning, J. Yang, S. Jiang, L. Zhang, and M.-H. Yang, "Object tracking via dual linear structured SVM and explicit feature map," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4266-4274, June 27-30, 2016. Article (CrossRef Link)
[32] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. of the IEEE International Conference on Computer Vision Workshops*, pp. 621-629, December 7-13, 2015. Article (CrossRef Link)

Benxuan Wang is currently pursuing a master degree in the School of Internet of Things Engineering at the Jiangnan University and majored in electrical and communication engineering. Her primary research interests cover computer vision, video information processing and target tracking.

Jun Kong received the M.S. degree in pattern recognition and intelligent system from Institute of Intelligent Machines, Chinese Academy of Sciences, China, in 2003, and PH.D degree in electronic science and technology from Shanghai Institute of Technical Physics, Chinese Academy of Sciences, China, in 2011. He jointed Jiangnan University in 2004, where he is currently an associate professor and the assistant dean of the school of Internet of Things Engineering. His research interests include computer vision, image processing, target tracking and human action recognition.



Min Jiang received her PH.D degree from Institute of Plasma Physics, Chinese Academy of Sciences, China, in 2005. She is currently a professor in the school of Internet of Things Engineering at the Jiangnan University. Her primary research is in the area of machine learning and computer vision with broad applications such as public surveillance, human computer interaction, and biomechanics.

Jianyu Shen is currently a graduate student in the School of Internet of Things Engineering at the Jiangnan University and majored in computer science and technology. His primary research interests cover computer vision, video information processing and target tracking.

Tianshan Liu is currently pursuing a master degree in the School of Internet of Things Engineering at the Jiangnan University and majored in computer science and technology. His primary research interests cover computer vision, human action recognition and target tracking.

Xiaofeng Gu received the Ph.D. degree from the Johns Hopkins University, USA, in 2003. He is currently a professor with Jiangnan University, China. His current research interests include signal processing, video analysis and machine vision.