EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Background perception for correlation filter tracker

Yushan Zhang[1], Jianan Li[1], Fan Wu[1], Lingyue Wu[1] and Tingfa Xu[1,2*]

Abstract

Visual object tracking is one of the most fundamental tasks in the field of computer vision, and it has numerous applications in many realms such as public surveillance, human-computer interaction, robotics, etc. Recently, discriminative correlation filter (DCF)-based trackers have achieved promising results in short-term tracking problems. Most of them focus on extracting reliable features from the foreground of input images to construct a robust and informative description of the target. However, it is often ignored that the image background which contains the surrounding context of the target is often similar across consecutive frames and thus can be beneficial to locating the target. In this paper, we propose a background perception regulation term to additionally exploit useful background information of the target. Specifically, invalid description of the target can be avoided when either background or foreground information becomes unreliable by assigning similar importance to both of them. Moreover, a novel model update strategy is further proposed. Instead of updating the model by frame, we introduce an output evaluation score, which serves to supervise the tracking process and select high-confidence results for model update, thus paving a new way to avoid model corruption. Extensive experiments on OTB-100 dataset well demonstrate the effectiveness of the proposed method BPCF, which gets an AUC score of 0.689 and outperforms most of the state-of-the-art.

Keywords: Correlation filter, Background perception, Model update, Visual tracking

1 Introduction

Discriminative correlation filter (DCF) trackers [1–17] have shown remarkable progress in recent years. The first CF-based method is Minimum Output Sum of Squared Error (MOSSE) [10], which receives a speed of more than 600 frames per second (FPS). After that, many improvements have been made to escalate its performance. The circulant structure of sequences is exploited to augment training samples [1]. Kernelized correlation filter is proposed to get a multi-channel extension of linear correlation filters [9]. To integrate multi-resolution feature maps, continuous convolution operators for visual tracking are also proposed [4] and utilized by many state-of-the-art trackers, such as ECO [3] and CFWCR [12], among which CFWCR exploits the great power of deep convolutional neural networks (CNN) features without using any hand-crafted features such as HOG [18] or color names [19], and achieves great performance in both accuracy and robustness. Afterwards, there are also trackers focusing on foreground feature selection [14] and reliability learning [20].

However, most of these methods only focus on foreground information, while they do not take good advantage of the background information which is also beneficial for tracking. Moreover, most trackers update the model after each frame, or after every N frames by using a sparse update scheme to avoid the tracker being dominated by recent samples. Nevertheless, such trackers still suffer from model corruption since they update the model indiscriminately regardless of whether the tracking result is accurate or not.

In this paper, aiming at the above issues, we propose a novel tracker, background perception correlation filter tracker (BPCF), based on an improved version of ECO [3], Correlation Filters with Weighted Convolution Responses (CFWCR) [12], which achieves remarkable results on VOT challenges [21]. In order to better exploit and make full use of the background information of the input, we propose a background perception regulation term. Particularly, we first divide the search area of the input images into several small pieces. By introducing a regulation

* Correspondence: ciom_xtf@bit.edu.cn
[1] School of Optics and Photonics, Image Engineering & Video Technology Lab, Beijing Institute of Technology, Beijing 100081, China
[2] Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing 100081, China

term that minimizes the sum of the L2 distances between the convolutional outputs of all possible pairs of the pieces, we assign similar importance to all the different small pieces regardless of whether the region belongs to the background or the foreground part of the input samples. In addition, as for the problem of indiscriminate model update, we introduce a novel model update strategy by computing a confidence score for the tracking result after every N frames, and only update the model when the confidence score is higher than a preset threshold, i.e., a particular proportion of the average of all the previous confidence scores. Figure 1 shows some qualitative results of our proposed method BPCF compared to some state-of-the-art on sample sequences of OTB-2015, from which we can see that our method outperforms all the other trackers. Moreover, quantitative results on OTB-2015 dataset show that our tracker BPCF has achieved state-of-the-art performance.

To sum up, our work makes the following contributions:

We propose a new DCF-based tracking model which integrates a background perception regulation term to stress the equal contribution of the foreground and background information and a novel model update strategy to supervise the tracking results into a unified tracking framework.

A background perception regulation term is introduced to the existing CFWCR tracker to exploit the background information of the input samples and emphasize equal contributions of the background and foreground information to avoid the tracker being dominated by unreliable parts.

A novel model update strategy is proposed to avoid model corruption. Instead of updating the model by frame, we compute a confidence score for the tracking result after every N frames, and select high-confidence results for model update.

2 Related work

Correlation filter-based methods for visual object tracking have shown dominant results in recent years. The first CF-based tracker, MOSSE [10], which only uses grayscale image and single-channel feature, could produce stable tracking results when initialized by a single frame and achieve the speed of 669 frames per second. After that, CF-based trackers become increasingly popular and have received impressive results on OTB-2015 object tracking benchmarks [22]. Due to the short of training samples when initializing the tracker, Henriques et al. [1] introduced the circulant structure of input images to augment training samples. Later, Henriques et al. [9] used kernel regression, which has exactly the same complexity as its linear counterpart, to combine different features and received better results. Nevertheless, using the circulant structure of image sequences could beget boundary effect. To solve this problem, Danelljan et al. [8] introduced a spatial regularization term. The proposed regularization weights penalize the correlation filter coefficient by assigning higher value at the edge of the filter and lower value at the central part. Spatial-temporal regularization term is also introduced by Danelljan et al. [10]. The new filter could be solved efficiently via alternating direction method of multipliers (ADMM) and provide a 5× speedup. Later, Danelljan et al. [4] proposed continuous convolution

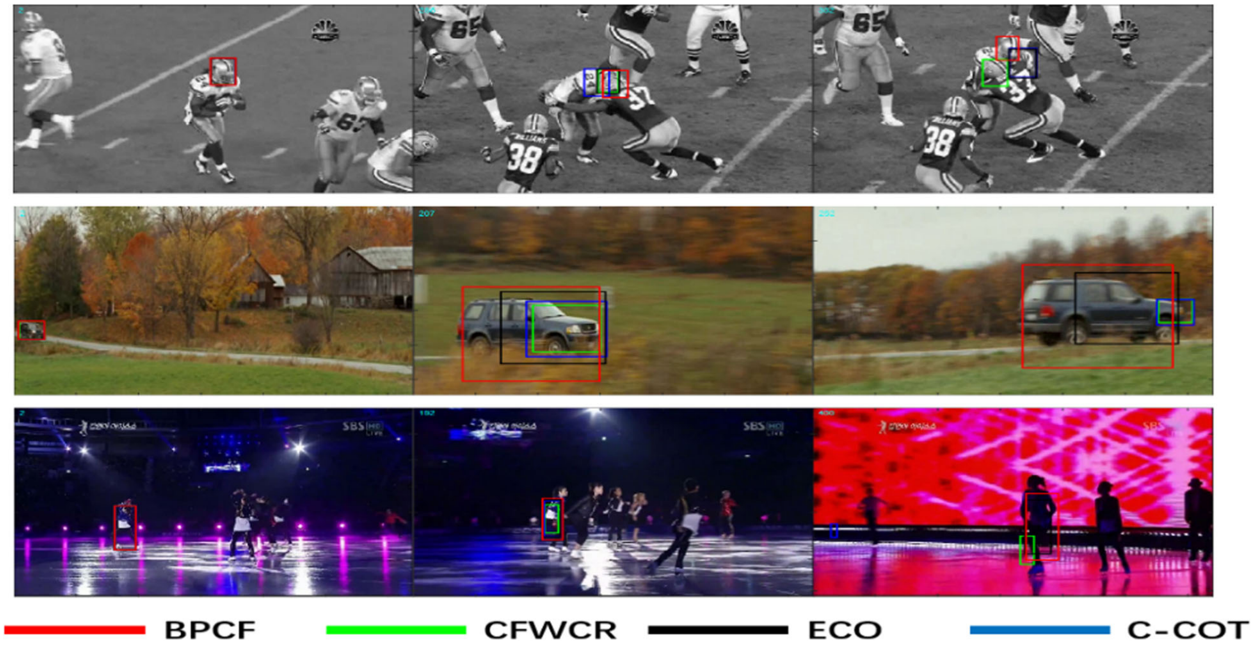


Fig. 1 Results on OTB2015. The comparison of BPCF and some state-of-the-art trackers on sample sequences

operators to enable the integration of multi-resolution deep feature maps. Many subsequent trackers based on this received good results. Danelljan et al. [3] revised C-COT [4] by introducing a factorized convolution operator, a compact generative model, which significantly reduced the computational complexity. He et al. [12] exploited the great power of deep CNN features without using any hand-crafted features and got great performance both in accuracy and robustness. Gundogdu et al. [14] put forward the importance of feature selection. Sun et al. [20] introduced a joint discrimination and reliability learning method, which highlighted the importance of the foreground and its different reliability.

Most of the previous methods only focus on the foreground information and rely on hand-crafted features such as HOG and CN. Differently, we only take deep features as our input and propose a background perception regulation term to ensure that the background and foreground of the input samples have similar contributions during tracking. By adding this regulation term, the tracker is unlikely to be dominated by unreliable parts of the target object, which could solve the problem of overfitting. Moreover, a novel model update strategy is proposed. Instead of updating the model regardless of whether the tracking result is precise or not, we update only when the tracking result is reliable and stop updating when the tracking result is incorrect or the tracking target is undergoing severe occlusion

3 Methods

The approach is actually twofold: firstly, we found that the background information in a given image sequence is always similar in consecutive frames, which could help to recognize a given target efficiently. Thus, we introduce a background perception regulation term, which could help us additionally exploit the background information and learn a more robust correlation filter. Secondly, most existing trackers update the tracker indiscriminately regardless of whether the tracking result is precise or not. The problem of such update strategy is that when the target is experiencing severe occlusion or the tracking result is imprecise, it will cause the tracker to corrupt. To solve this problem, a self-adaptive model update strategy is proposed. We introduce an output evaluation score, where the score is lower while the object is being occluded or the tracking result is incorrect. We then can select those reliable samples to update the model.

3.1 Base framework

Our framework, like many other DCF trackers, is based on C-COT, a theoretical framework for learning continuous convolution operators. This kind of tracker adopted an implicit interpolation model for the training samples. Assuming each sample x_j contains D feature channels$x_j^1, x_j^2, \dots, x_j^D$, N_d as the number of spatial samples in x_j^d, where $d \in \{0, 1, 2, \dots, D\}$. Unlike the traditional DCF trackers, where each feature channel $x_j^d \in \mathbb{R}^{N_d}$ consists of N_d discrete spatial variables, it transfers the feature maps into continuous spatial domain by introducing an interpolation kernel b_d to get an interpolation operatorJ_d:

$$J_d\left\{x^d\right\}(t) = \sum_{n=0}^{N_d-1} x^d[n] b_d\left(t - \frac{T}{N_d} n\right), \tag{1}$$

where the interpolation kernel b_d has a period of T ($T > 0$). Thus, the result $N_d\{x^d\}$ is a continuous feature map with a period of T to be used for further computation. CFWCR framework negates hand-crafted features such as HOG [18] and CN [19], and adopts CNN features to extract feature maps. Specifically, they use VGG-M [23] network pre-trained on ILSVRC [24] dataset to extract multi-resolution continuous feature maps, and employ the first and the fifth convolutional layer as two deep feature channels. The filter f is trained by minimizing the following function:

$$E(f) = \sum_{j=1}^{m} \alpha_j \left\| W_1 \sum_{a=1}^{D_{\text{conv1}}} f^a * J_a\left\{x_j^a\right\} + W_2 \sum_{b=1}^{D_{\text{conv5}}} f^b * J_b\left\{x_j^b\right\} - y_j \right\|^2 + \sum_{d=1}^{D} \left\|\omega f^d\right\|^2, \tag{2}$$

where α_j represents the importance of each training sample, and ω is the spatial regulation term to avoid boundary effect.

The convolution responses of the two channels are weighted summed to get a final confidence response:

$$S_f(x) = W_1 \sum_{a=1}^{D_{\text{conv1}}} f^a * J_a\{x^a\} + W_2 \sum_{b=1}^{D_{\text{conv5}}} f^b * J_b\{x^b\}, \tag{3}$$

where the feature maps extracted from the first and fifth convolutional layers are first interpolated using Eq.1 and then convoluted with filter f^a and f^b trained by Eq.2. The assigned weights W_1 and W_2 denote the significance of each layer.

3.2 Background perception

We propose a background perception regulation term$R(h, X)$, by which we regulate the filter to assign larger importance to the region where the extracted feature map X has a smaller value. By this means, we can assign similar importance to different regions of the training samples and avoid the tracker being dominated by unreliable parts. The regulation term can be formulated as:

$$R(h, X) = \sum_{k=1}^{K} \sum_{m,n}^{M} \left(\sum_{d=1}^{D} \left(P_d^m x_{k,d}\right)^T h_d - \sum_{d=1}^{D} \left(P_d^n x_{k,d}\right)^T h_d \right)^2, \tag{4}$$

where $x_{k,\ d} \in \mathbb{R}^{K \times 1}$ is the kth cyclically shift of the input vector $x_d \in \mathbb{R}^{K \times 1}$ for the dth channel. $P_d^m = \operatorname{diag}(p_d^m(1), \dots, p_d^m(K)) \in \mathbb{R}^{K \times K}$ is the mth binary mask (Fig. 2) which crops the samples to the mth sub-region. $h_d \in \mathbb{R}^{K\times 1}$ is the target filter of dth channel. To simplify Eq.4, the formula is rewritten as follows:

$$R(h,X) = \sum_{m,n}^{M} \left\| \sum_{d=1}^{D} X_d^T P_d^m h_d - \sum_{d=1}^{D} X_d^T P_d^n h_d \right\|_2^2 = \sum_{m,n}^{M} \left\| X^T P^m h - X^T P^n h \right\|_2^2, \tag{5}$$

where $P^m = diag(P_1^m, \ldots, P_D^m) \in \mathbb{R}^{DK\times DK}$ is a block diagonal matrix where P_d^m is the dth diagonal block. $X_d = [x_{1,\ d}, x_{2,\ d}, \ldots, x_{K,\ d}] \in \mathbb{R}^{K\times K}$ denotes all the cyclical shift of the input vector x_d. $X = [X_1^T, X_2^T, \ldots, X_D^T]^T \in \mathbb{R}^{DK\times K}$ is a matrix that fits all the circulant matrices of different channels together. $h \in \mathbb{R}^{DK\times 1}$ is the ultimate filter.

By introducing this background perception term, the filters are learned by minimizing the following objective:

$$E(h) = \left\| y - \sum_{d=1}^{D} X_d^T h_d \right\|^2 + \eta \sum_{m,n}^{M} \left\| \sum_{d=1}^{D} X_d^T P_d^m h_d - \sum_{d=1}^{D} X_d^T P_d^n h_d \right\|_2^2 + \sum_{d=1}^{D} \left\| \omega h_d \right\|^2, \tag{6}$$

$$E(h) = \left\| y - X^T h \right\|^2 + \eta \sum_{m,n}^{M} \left\| X^T P^m h - X^T P^n h \right\|_2^2 + \left\| Wh \right\|^2, \tag{7}$$

where y is the predefined Gaussian window objective function. The binary mask $P^m = \mathrm{diag}(P_1^m, \ldots, P_D^m) \in \mathbb{R}^{DK\times DK}$ is a block diagonal matrix where P_d^m is the dth diagonal block. Input sample $X = [X_1^T, X_2^T, \ldots, X_D^T]^T \in \mathbb{R}^{DK\times K}$ is a matrix that fits all the circulant matrices of different channels together. And $h \in \mathbb{R}^{DK\times 1}$ is the ultimate filter that we get.

To get the optimal filter h, we can solve the minimization problem by using conjugate gradient descent method. We first compute the derivative of Eq.7, and then set it to zero to get the following equation:

$$Ah = Xy, \tag{8}$$

where A is defined as:

$$A = XX^T + 2\eta \sum_{m=1}^{M} M (P^m)^T XX^T (P^m) - 2\eta \left[\sum_{m=1}^{M} P^m \right]^T XX^T \left[\sum_{m=1}^{M} P^m \right] + W^T W, \tag{9}$$

To solve the normal equation by conjugate gradient descent method, we employ the following iterative procedure:

$$\begin{aligned} a_i &= \frac{r_i^T r}{P_i^T A P_i}, \\ h_{i+1} &= h_i + a_i P_i, \\ r_{i+1} &= r_i + a_i A P_i, \\ P_{i+1} &= r_{i+1} + \left(\frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} \right) P_i, \end{aligned} \tag{10}$$

where we set $h_0 = 0$, $r_0 = Xy$, $P_0 = r_0$. And the number of conjugate gradient descent iterations is set to 5.

3.3 Target localization

In the target localization step in Nth frame, we first extract the feature map x of the search region. x^ddenotes the feature map of the dth channel. The convolutional outputs of the two channels are computed by convolution operation and then weighted summed up to get a final confidence response $S_h(x)$. The convolution process is computed in Fourier domain to reduce the computing burden.

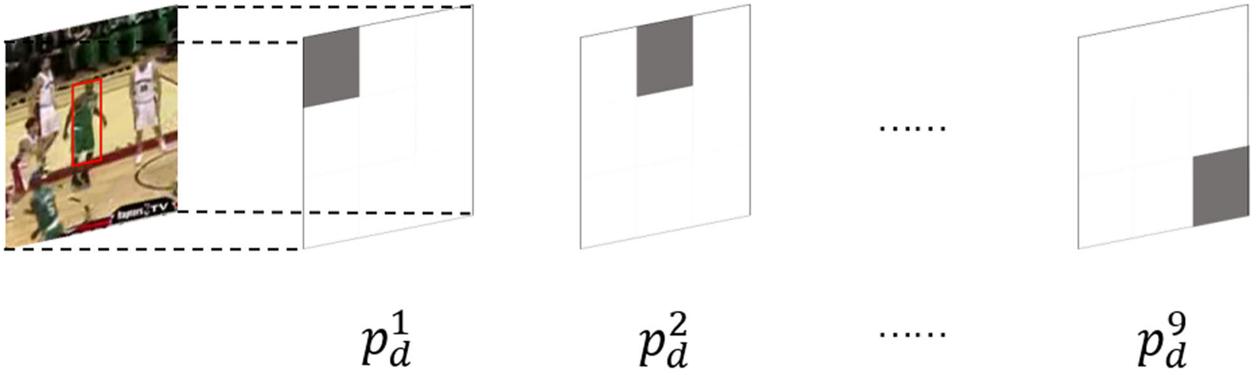


Fig. 2 Binary masks. When considering two-dimensional case and suppose$m = 9$, the binary masks from $p_d^1, p_d^2, \ldots, p_d^9$ are shown in the figure where the gray pixels are assigned the value of 1 and the white pixels are assigned the value of 0

$$S_h(x) = \sum_{d=1}^{D} \alpha_d \mathbb{F}^{-1}\left(\mathbb{F}\left(h^d\right)\odot\mathbb{F}\left(x^d\right)\right), \tag{11}$$

where $d = 2$ and the two feature channels are the first and the fifth convolutional layer of VGG-16 net pre-trained on ILSVRC dataset. And α_d is the weight of each layer which denotes the importance of each feature channel.

3.4 Model update strategy

Most of the existing state-of-the-art DCF-based trackers update the tracking model after each frame [8, 10]. These methods suffer from model corruption since they update the model indiscriminately no matter the tracking result is accurate or not. Moreover, the model is easy to be dominated by recent frames, in which case, if the tracking result of the recent frames is imprecise, the tracking mission is prone to be failed. Thus, when the target object is experiencing severe background cluster, deformation, and occlusion, the model will become highly unreliable. Other trackers such as ECO [3] and CFWCR [12] adopt the sparser updating scheme, where the filter is only updated by starting the optimization process at every N frames. Specifically, the traditional method of model update usually sets $N = 1$, and the sparser updating method mostly sets $N = 5$. By applying this scheme, the tracker could avoid being dominated by recent samples. Nevertheless, the problem of indiscriminate model updating still exists.

Our proposed model update strategy also adopts the sparse updating scheme, but unlike the traditional methods where the input samples extracted from each frame are used to update the model, we compute a confidence score for the tracking result and discard those with the confidence less than a threshold, i.e., a certain proportion of the average of all the previous scores. The confidence score for the tracking result is defined as:

$$\text{Score} = \frac{F_{\max} - F_{\min}}{\left(\sum_{x,y}\left(F_{x,y} - F_{\text{ave}}\right)^2\right)^{1/2}}, \tag{12}$$

where $F_{\max}$ and $F_{\min}$ denote the maximum and minimum value of the confidence response $S_h(x)$. $F_{x,\ y}$ denotes the xth row and the yth column of $S_h(x)$. In this equation, we can see that the confidence score is in direct proportion to the peak value of the response and in inverse proportion to the standard deviation of the response.

Figure 3 illustrates the significance of the proposed method, from which we can see that when the target object is experiencing severe occlusion, the confidence score will reduce drastically due to the reduction of the peak value of the confidence response $S_h(x)$ and the increase of the standard deviation of the response. When the output localization of the object is inaccurate, the fluctuation of the confidence response is far more intense and the confidence score reduces significantly. Moreover, the confidence response of different sequences can vary a lot while all of their tracking results are good. Thus, we cannot choose an invariable threshold to decide whether the result is reliable or not. We use the historical average values of the computed score as a testing criterion. If the computed score of the current frame is greater than its respective historical average values with a certain ratio, which is set to 0.6 in our proposed method, the output is considered reliable. In sum, by applying our novel model update strategy, we could discard those unreliable training samples during the process of tracking and only preserve the reliable ones for model update.

3.5 Tracking with background perception and self-adaptive template update

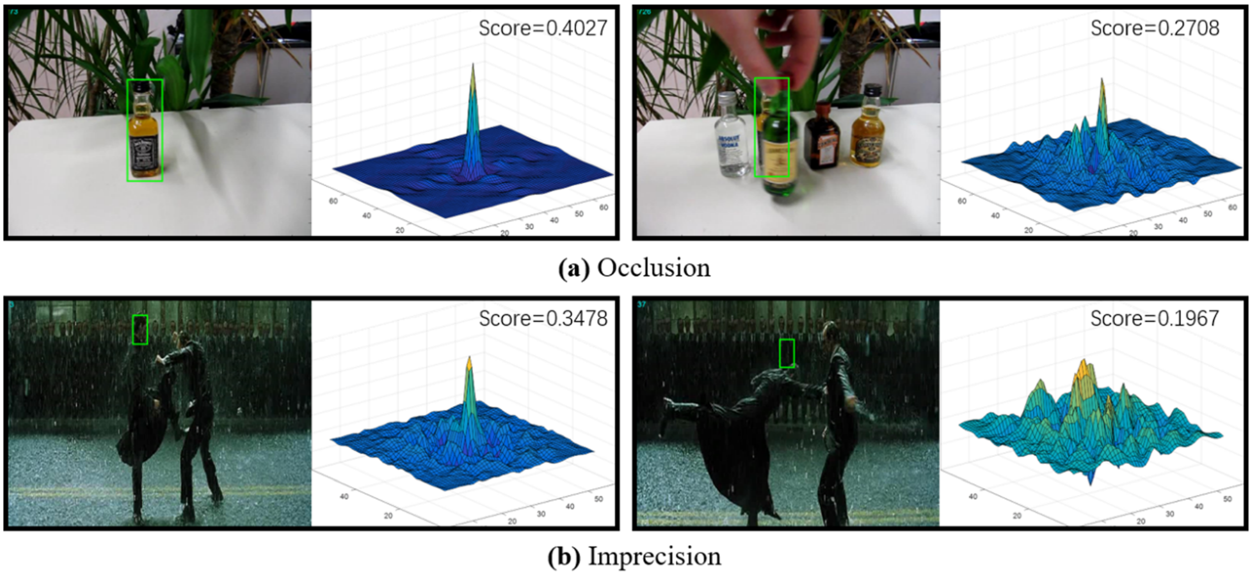
The algorithm flow of our proposed method is described in Table 1. For the target localization step, we compute the confidence response $S_h(x)$ with Eq. (11) and then optimize $S_h(x)$ by the standard Newton's method. The optimization process will be completed in 5 iterations. As for the update step, we first calculate the confidence score S of the tracking result for the current frame and stop model update while the result has low confidence. The filter h is optimized with Eq. (8) by conjugate gradient descent method.

4 Results and discussion

In this section, we conduct comprehensive experiments to evaluate our proposed method BPCF and compare it with state-of-the-art DCF-based trackers. We first present the implementation details in Section 4.1. Then we compare our tracker with its baseline CFWCR [12] tracker and some other state-of-the-art trackers on OTB-2015 dataset [22] both quantitatively and qualitatively in Section 4.2 and Section 4.3.

4.1 Implementation details

Our proposed method is implemented in MATLAB using Matconvnet tools, and the basic settings are the same as CFWCR. The features are extracted from the conv1 layer and the conv5 layer. The relative weight $\sigma = W_2/W_1$ is set to 2. The search image is 4 times the size of the target object. The maximum number of stored training samples is set to 50, and the learning rate is 0.012 with a raining gap 5. The parameter of the proposed regulation term is set to 1.5.

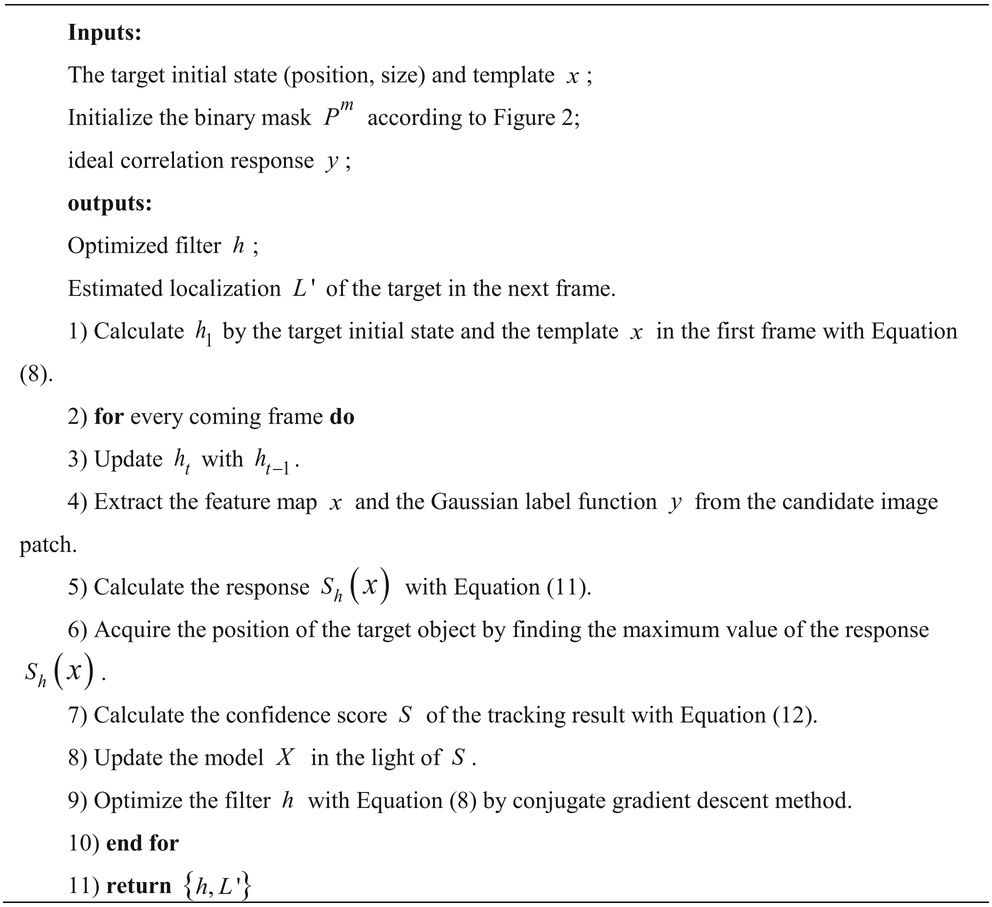


(a) Occlusion

(b) Imprecision

Fig. 3 Confidences scores. **a** shows the change of the confidence score when the target object is experiencing severe occlusion. **b** shows the change of the confidence score when the localization is imprecise

Table 1 The algorithm flow



```
Inputs:
    The target initial state (position, size) and template $x$;
    Initialize the binary mask $P^m$ according to Figure 2;
    ideal correlation response $y$;
outputs:
    Optimized filter $h$;
    Estimated localization $L'$ of the target in the next frame.
1) Calculate $h_1$ by the target initial state and the template $x$ in the first frame with Equation (8).
2) for every coming frame do
3) Update $h_t$ with $h_{t-1}$.
4) Extract the feature map $x$ and the Gaussian label function $y$ from the candidate image patch.
5) Calculate the response $S_h(x)$ with Equation (11).
6) Acquire the position of the target object by finding the maximum value of the response $S_h(x)$.
7) Calculate the confidence score $S$ of the tracking result with Equation (12).
8) Update the model $X$ in the light of $S$.
9) Optimize the filter $h$ with Equation (8) by conjugate gradient descent method.
10) end for
11) return $\{h, L'\}$
```

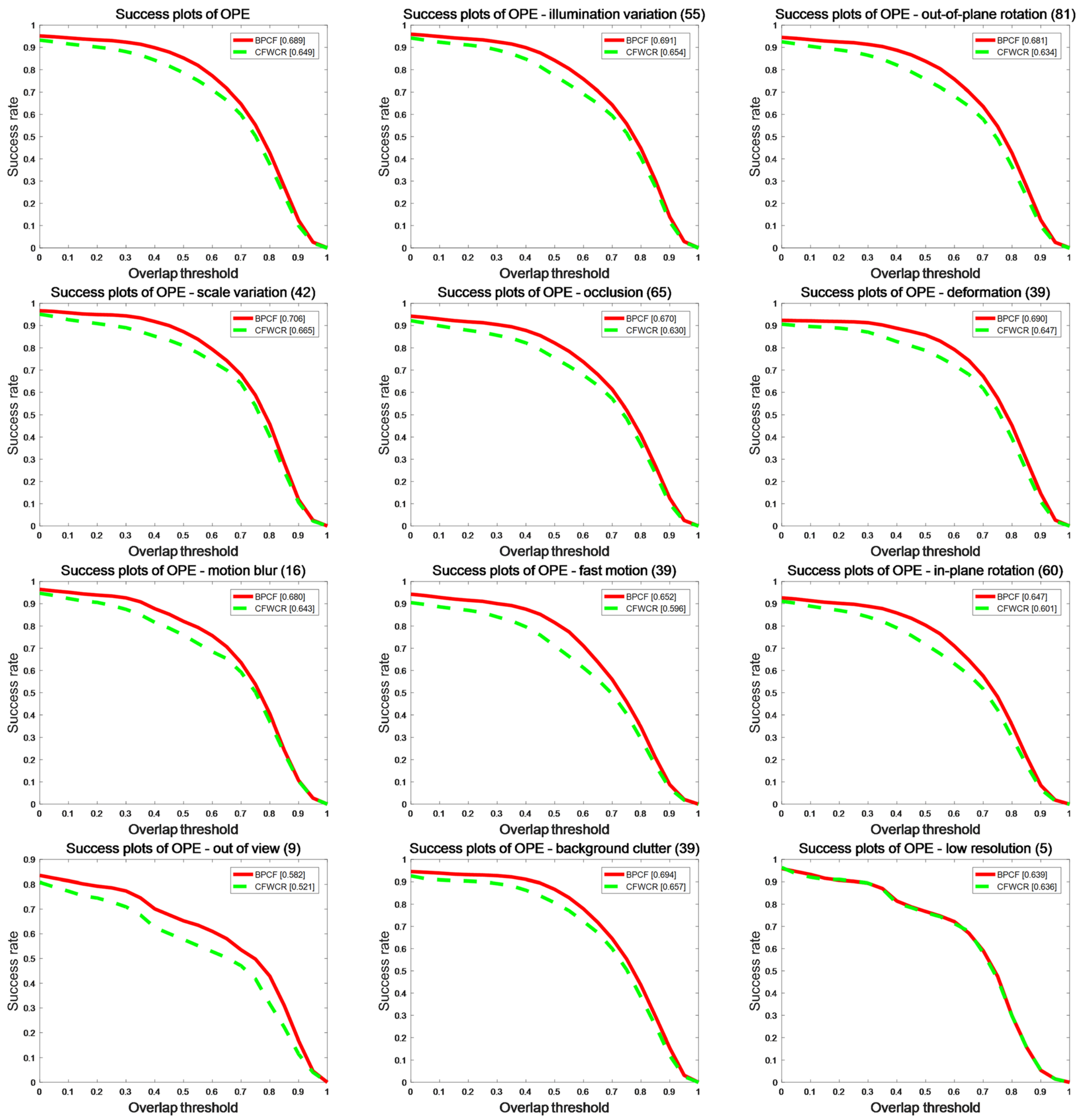


Fig. 4 Baseline comparison. Success plot of OPE on OTB-2015 benchmark. The red line denotes BPCF with background perception regulation term and the self-adaptive model update strategy jointly. The green line denotes CFWCR

Table 2 Baseline comparisons when using background perception (BPCF-BP) and model update (BPCF-MU) separately and jointly (BPCF). We report the AUC score on the OTB-2015 dataset

Trackers	CFWCR	BPCF-BP	BPCF-MP	BPCF
AUC	0.649	0.685	0.654	0.689

4.2 Quantitative evaluation

In this section, we mainly evaluated our algorithm on OTB-2015 [22] benchmark. We tested our tracker on all the 100 sequences with tracking difficulties including illuminate variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out of view, background clutter, low resolution, etc. One commonly used evaluation metric is the overlap score, which is defined as $R_t \cap R_g / R_t \cup R_g$, where R_t denotes the bounding box of the tracking result and R_g denotes the ground-truth bounding box. Given a threshold between 0 and 1, we can get an average success rate by comparing the overlap score and the certain threshold. The main criteria of OTB-2015 to determine whether the result is good or not is the area under the curve (AUC) of each success plot, which is the average of the success rates corresponding to different overlap thresholds. In our experiment, we use the area under the curve (AUC) to generate the success plot of OPE.

4.2.1 Baseline comparison

We first compare BPCF with the based work CFWCR. Both are tested on OTB-2015 benchmark. We can see in Fig. 4 that the performance of BPCF is escalated in all the situations including illuminate variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out of view, background clutter, low resolution, etc.

As shown in Fig. 4, our tracker has a significantly better performance compared with CFWCR. By implementing the background perception regulation term and the

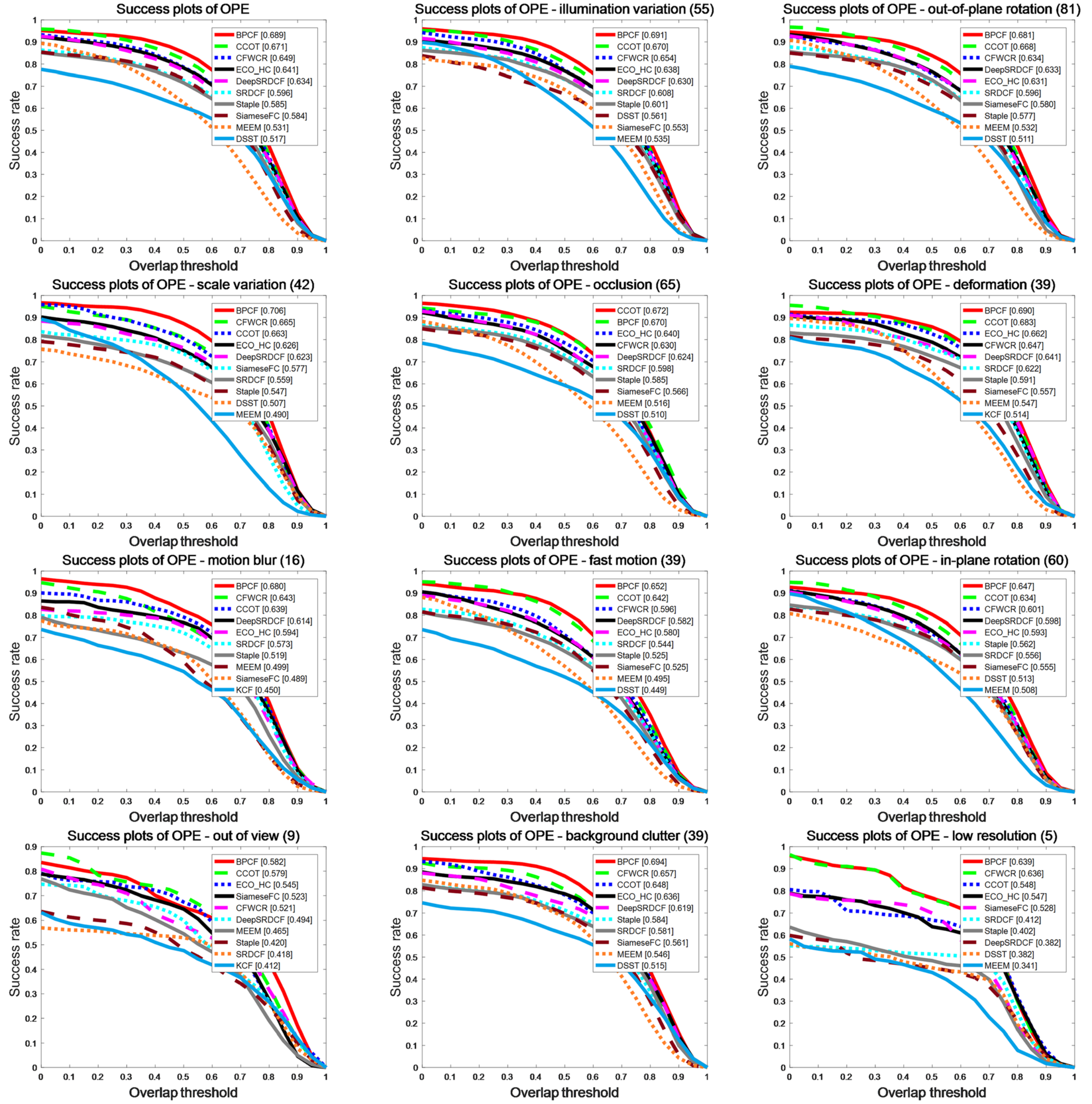


Fig. 5 State-of-the-art comparison. Success plot of OPE for some of the state-of-the-art trackers and BPCF on OTB-2015 benchmark

self-adaptive model update strategy jointly, the performance of our tracker improves by about 4%.

To get a better evaluation, we extend our experiment on OTB-2015 benchmark by first separately implement our proposed background perception regulation term and model update strategy and then integrate them up at the same time. Table 2 shows the analysis of our contributions, from which we can see that our tracker BPCF has a significantly better performance compared with the based tracker. By implementing the background perception regulation term, our tracker improves by about 3.6%, and by adding the self-adaptive model update strategy furthermore, it improves by about 4%.

4.2.2 State-of-the-art comparison

Here, we use both success plots and the precision plots [22] over all 100 videos on OTB-2015 dataset to analyze our approach BPCF. In the evaluation of success plots, the area under the curve (AUC) of success plots is used to rank the trackers. The precision plot reports the

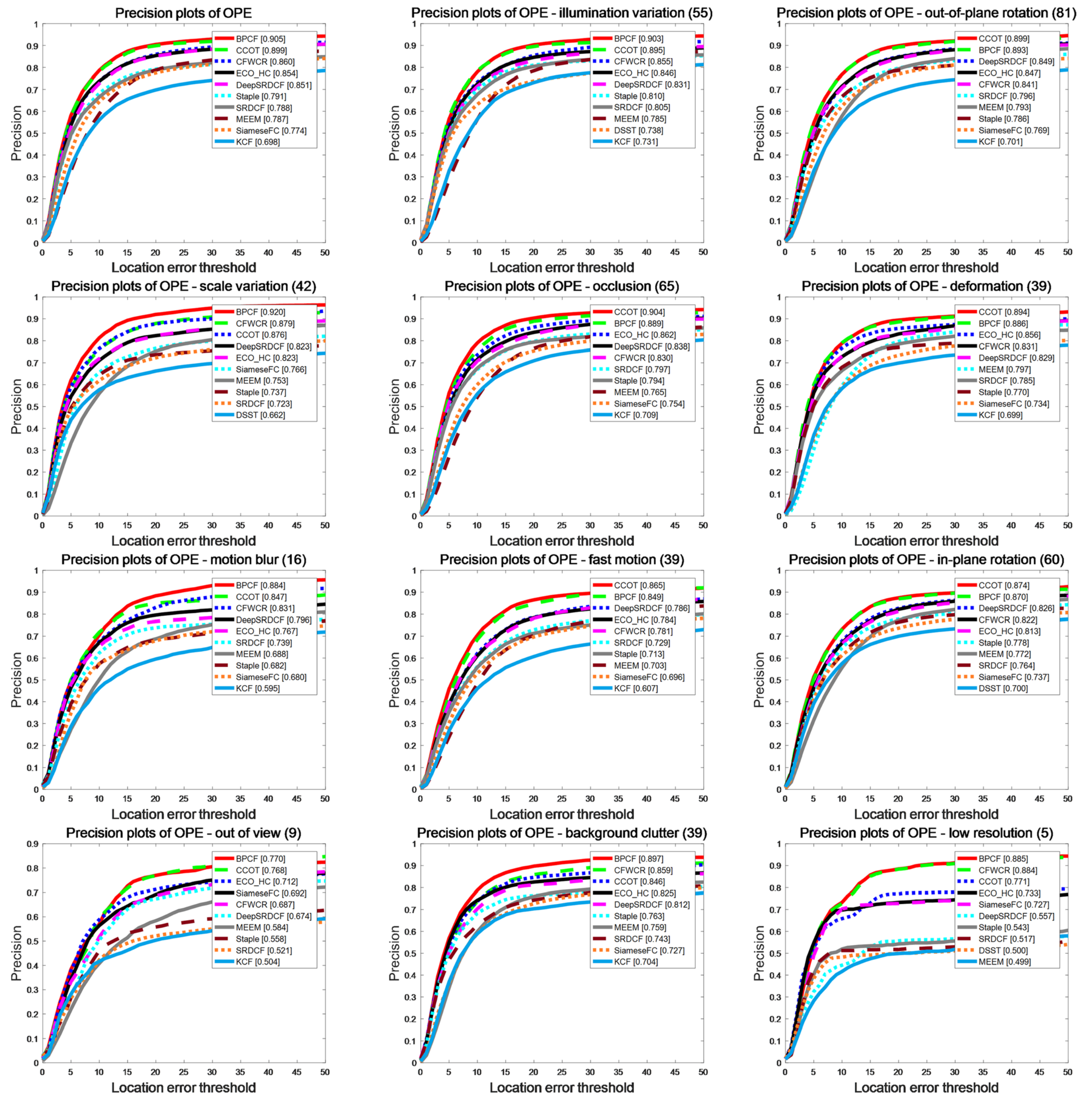


Fig. 6 State-of-the-art comparison. Precision plot of OPE for some of the state-of-the-art trackers and BPCF on OTB-2015 benchmark

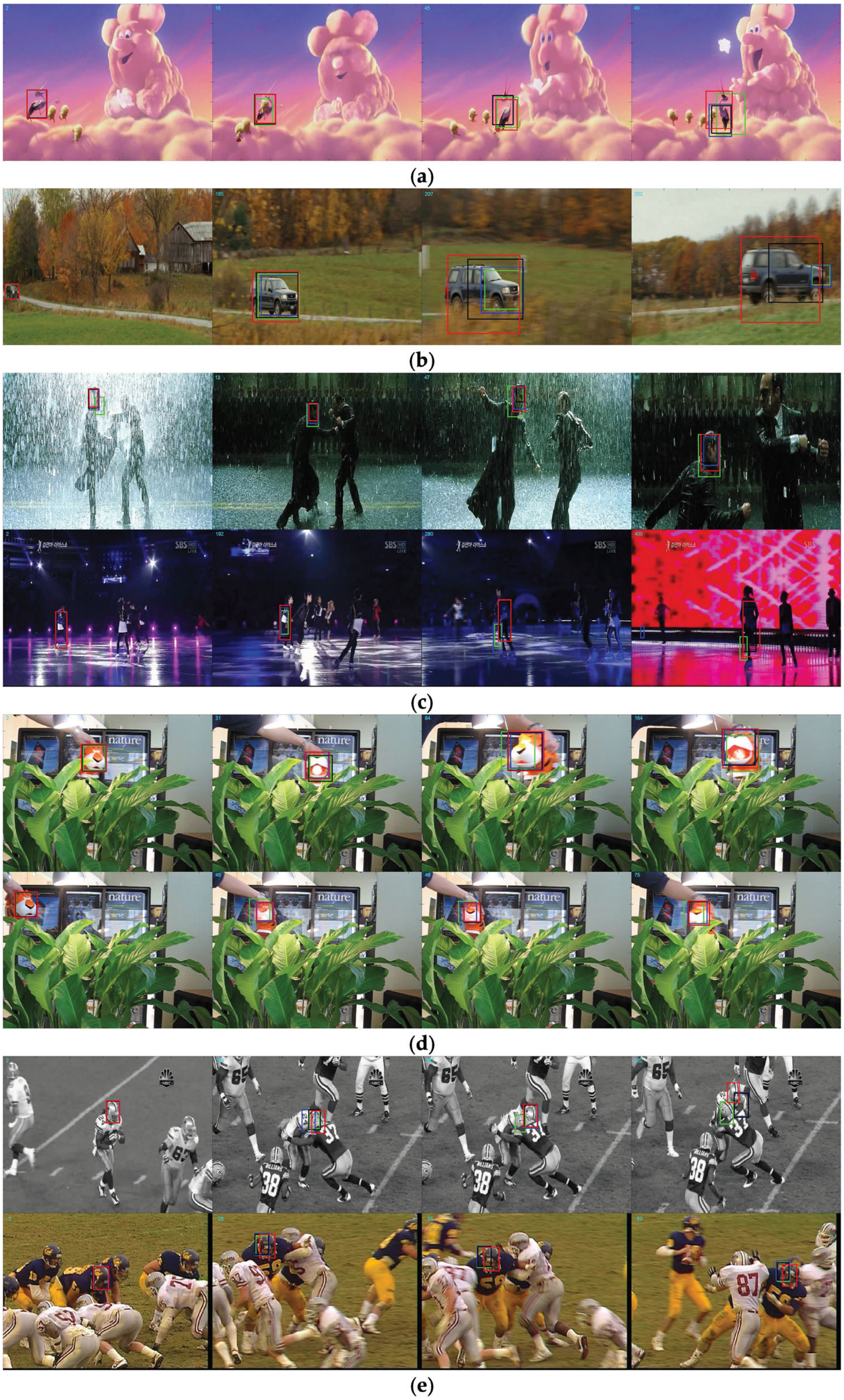

(a)
(b)
(c)
(d)
(e)

Fig. 7 Qualitative comparison. **a** bird2; **b** carscale; **c** matrix, skating1; **d** tiger1, tiger2; **e** football, football1. The tracking result of BPCF, ECO, C-COT, and CFWCR are marked in red, black, blue, and green

average distance precision score at 20 pixels for each method. We conduct extensive experiment and compare our tracker with 10 state-of-the-art methods: MEEM [25], C-COT [4], CFWCR [12], ECO [3], DeepSRDCF [26], SRDCF [8], Staple [27], SiameseFC [28], DSST [29], and KCF [9].

The evaluation results of our proposed tracker BPCF and the 10 competitive trackers are demonstrated in Figs. 5 and 6, from which we can see that among all the existing trackers including some DCF-based trackers and some CNN based trackers, C-COT [4] and our based method CFWCR [12] provide the best results on both success plots and precision plots. The result of our proposed method BPCF outperforms both of them and provides the best result of 0.689 AUC score and 0.905 precision score. Besides, ECO [3] with hand-crafted features gets an AUC score of 0.641 and a precision score of 0.854, and SRDCF [26] with deep features gets an AUC score of 0.634 and a precision score of 0.851. Other trackers have an AUC score less than 0.6 and a precision score less than 0.8.

4.3 Qualitative evaluation

To evaluate the performance of our tracker alone with some other state-of-the-art trackers qualitatively, the tracking results of BPCF along with ECO, C-COT, CFWCR, etc. are presented in Fig. 7.

In Fig. 7a, the results of the sequence bird2 is shown, in which the tracking target is undergoing out-of-plane rotation and occlusion. We can intuitively see from the result that, among all the trackers, BPCF tracker works best and remains the most robust. All the other trackers drift more or less during tracking.

Figure 7b shows the results of the sequence carscale, where the tracking target is experiencing scale variation and occlusion. We can see from the figure that BPCF tracker stays stable while the target object experiences severe occlusion. The other trackers experience overfitting while the object's scale varies.

In Fig 7c, the results of the sequences matrix and skating1 are presented, where target objects are undergoing illuminate variation. We can see from the result of the sequence matrix that due to the illuminate variation, other trackers experience some degree of imprecision, while BPCF remains the most robust. In the sequence skating1, CFWCR, and C-COT have lost the target halfway, while BPCF and ECO remain robust to the end of the sequence.

In Fig. 7d, we present the tracking result of sequences tiger1 and tiger2, in which motion blur and occlusion occur frequently. It is shown in the figure that all of the trackers managed to track the target successfully, although CFWCR has some occasional drift.

In Fig. 7e, the tracking results of sequences football and football1 are given. Background clutter can be seen in these two video sequences. We can see from Fig. 6e that when the target is occluded by other similar objects, all the other trackers except BPCF have a certain degree of drift. In the sequence football, one of the trackers even fail to locate the correct object after the occlusion.

5 Conclusions

In this paper, we discard the traditional use of the hand-craft features and adopt deep features only for visual tracking. Moreover, we propose a background perception regulation term to alleviate the overmuch highlighting of the foreground of the input training samples. By making full use of the background information, our tracker works more robustly. The self-adaptive model update strategy is also implemented to avoid model corruption by selecting high confidence tracking results as training samples. We evaluate our method on OTB-2015 benchmark and experimental results show that our tracker achieves the state-of-the-art performance.

Abbreviations

CN: Color names; CNN: Convolutional neural network; DCF: Discriminative correlation filter; HOG: Histogram of gradient; VGG: Visual geometry group

Acknowledgements

The authors would like to thank Tingfa Xu for the support.

Authors' contributions

JL and TX conceived of the tracking method. YZ was responsible for the programming. FW and LW verified the analytical methods. YZ wrote the manuscript, and all authors revised the final manuscript. In addition, TX and JL are the corresponding authors. All authors read and approved the final manuscript.

Funding

This work was supported by the Major Science Instrument Program of the National Natural Science Foundation of China under Grant 61527802.
This work was supported by the General Program of National Nature Science Foundation of China under Grants 61371132 and 61471043.

Availability of data and materials

All data are fully available without restriction.

Competing interests

The authors declare that they have no competing interests.

Received: 14 September 2019 Accepted: 18 December 2019
Published online: 16 January 2020

References

1. J Henriques, R Caseiro, P Martins, J Batista, in European Conference on Computer Vision. Exploiting the circulant structure of tracking-by-detection with kernels (2012), pp. 702-715.
2. M Danelljan, F. S. Khan, M Felsberg, J. V. D. Weijer, in IEEE Conference on Computer Vision and Pattern Recognition. Adaptive color attributes for real-time visual tracking (2014), pp. 1090–1097.
3. M Danelljan, G Bhat, F. S. Khan, M Felsberg, in IEEE Conference on Computer Vision and Pattern Recognition. ECO: Efficient Convolution Operators for Tracking (2017), pp. 21–26.
4. M Danelljan, A Robinson, F. S. Khan, M Felsberg, in European Conference on Computer Vision. Beyond correlation filters: learning continuous convolution operators for visual tracking (2016), pp. 472-488.
5. R. Yao, S. Xia, F. Shen, Y. Zhou, Q. Niu, Exploiting spatial structure from parts for adaptive kernelized correlation filter tracker. IEEE Signal Process Lett **23**, 658–662 (2016)

6. M. Danelljan, G. Häger, F.S. Khan, M. Felsberg, Discriminative scale space tracking. IEEE Trans Pattern Anal Machine Intell **39**, 1561–1575 (2016)
7. M Wang, Y Liu, Z Huang, Large margin object tracking with circulant feature maps. IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 4800-4808.
8. M Danelljan, G Häger, F. S. Khan, M Felsberg, in IEEE International Conference on Computer Vision. Learning spatially regularized correlation filters for visual tracking (2015), pp. 4310–4318.
9. J.F. Henriques, C Rui, P Martins, J Batista, High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis & Machine Intelligence 37, 583-596 (2014).
10. D.S. Bolme, J.R. Beveridge, B.A. Draper, Y.M. Lui, in IEEE Conference on Computer Vision and Pattern Recognition. Visual object tracking using adaptive correlation filters (2010), pp. 2544–2550.
11. F Li, C Tian, W Zuo, L Zhang, M. H. Yang, in IEEE Conference on Computer Vision and Pattern Recognition. Learning spatial-temporal regularized correlation filters for visual tracking (2018), pp. 4904-4913.
12. Z He, Y Fan, J Zhuang, Y Dong, H. L. Bai, In IEEE International Conference on Computer Vision Workshop. Correlation filters with weighted convolution responses (2017), pp. 1992-2000.
13. C Sun, D Wang, H Lu, M.H. Yang, in IEEE Conference on Computer Vision and Pattern Recognition. Learning spatial-aware regressions for visual tracking (2018), pp. 8962-8970
14. E. Gundogdu, A.A. Alatan, Good features to correlate for visual tracking. IEEE Transactions on Image Process. **27**(2526-2540) (2018)
15. T. Xu, Z.H. Feng, X.J. Wu, J. Kittler, Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE Trans Image Process **28**, 5596–5609 (2019)
16. B Huang, T Xu, B Liu, B Yuan, Context constraint and pattern memory for long-term correlation tracking. Neurocomputing. In Press, https://doi.org/10.1016/j.neucom.2019.10.021
17. B. Huang, T. Xu, S. Jiang, Y. Bai, Y. Chen, SVTN: Siamese Visual Tracking Networks with Spatially Constrained Correlation Filter and Saliency Prior Context Model. IEEE Access. **7**, 144339–144353 (2019)
18. N Dalal, B Triggs, in IEEE Conference on Computer Vision and Pattern Recognition. Histograms of oriented gradients for human detection (2005), pp. 886-893.
19. M Danelljan, G Häger, F.S. Khan, M Felsberg, in Scandinavian Conference on Image Analysis. Coloring channel representations for visual tracking (2015), pp. 117–129.
20. C Sun, D Wang, H Lu, M. H. Yang, in IEEE Conference on Computer Vision and Pattern Recognition. Correlation tracking via joint discrimination and reliability learning (2018), pp. 489-497.
21. M Kristan, A Leonardis, J Matas, M Felsberg, Z He, et al, in IEEE International Conference on Computer Vision Workshop. The visual object tracking vot2017 challenge results (2017), pp. 1949-1972.
22. Y. Wu, J. Lim, M.H. Yang, Object tracking benchmark. IEEE Trans Pattern Anal Machine Intell **37**, 1834–1848 (2015)
23. K Simonyan, A Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
24. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge. Int J Comput Vision **115**, 211–252 (2015)
25. J Zhang, S Ma, S Sclaroff, in European Conference on Computer Vision. MEEM: robust tracking via multiple experts using entropy minimization (2014), pp. 188-203.
26. M Danelljan, G Häger, F. S. Khan, M Felsberg, in IEEE International Conference on Computer Vision (ICCV) Workshops. Convolutional features for correlation filter based visual tracking (2015), pp. 59-66.
27. L Bertinetto, J Valmadre, S Golodetz, O Miksik, P. H. S. Torr, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Staple: complementary learners for real-time tracking (2016), pp. 1401-1409.
28. L Bertinetto, J Valmadre, J. F. Henriques, A Vedaldi, P. H. Torr, in European Conference on Computer Vision (ECCV) Workshops. Fully-Convolutional Siamese Networks for Object Tracking (2016), pp. 850-865.
29. M Danelljan, G Häger, F. S. Khan, M Felsberg, Accurate scale estimation for robust visual tracking. In BMVC (2014)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.