EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Adaptive response maps fusion of correlation filters with anti-occlusion mechanism for visual object tracking

Jianming Zhang[1,2*], Hehua Liu[1,2], Yaoqi He[1,2], Li-Dan Kuang[1,2] and Xi Chen[1,2]

*Correspondence: jmzhang@csust.edu.cn
[1] School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
Full list of author information is available at the end of the article

Abstract

Despite the impressive performance of correlation filter-based trackers in terms of robustness and accuracy, the trackers have room for improvement. The majority of existing trackers use a single feature or fixed fusion weights, which makes it possible for tracking to fail in the case of deformation or severe occlusion. In this paper, we propose a multi-feature response map adaptive fusion strategy based on the consistency of individual features and fused feature. It is able to improve the robustness and accuracy by building the better object appearance model. Moreover, since the response map has multiple local peaks when the target is occluded, we propose an anti-occlusion mechanism. Specifically, if the nonmaximal local peak is satisfied with our proposed conditions, we generate a new response map which is obtained by moving the center of the region of interest to the nonmaximal local peak position of the response map and re-extracting features. We then select the response map with the largest response value as the final response map. This proposed anti-occlusion mechanism can effectively cope with the problem of tracking failure caused by occlusion. Finally, by adjusting the learning rate in different scenes, we designed a high-confidence model update strategy to deal with the problem of model pollution. Besides, we conducted experiments on OTB2013, OTB2015, TC128 and UAV123 datasets and compared them with the current state-of-the-art algorithms, and the proposed algorithms have impressive advantages in terms of accuracy and robustness.

Keywords: Object tracking, Multiple features, Correlation filter, Anti-occlusion, Response maps fusion

1 Introduction

Object tracking is a challenging topic in computer vision. The main task of object tracking is to predict the position of the object in the subsequent frames by giving the object position information in the first frame. Despite the great advances made in recent years, designing a fast and reliable tracking method is still difficult because of many challenges in the tracking process, such as in-plane rotations, occlusions and fast motions [1, 2]. At present, the robust object tracking can be applied to many promising fields such as human–robotic interaction [3], real-time video surveillance [4] and automatic driving.

Learning an effective appearance model of a target online is crucial for accurate and reliable visual object tracking [5]. Therefore, many powerful models representing the appearance of the target are proposed to build successful visual object trackers, such as subspace learning, correlation filter (CF) [6–16], convolutional neural network (CNN) [16–19] and support vector machine (SVM) [20].

Correlation filter is a valuable tracking solution that allows a robust appearance model for targeted online learning. Moreover, many CF-based trackers obtained impressive performance in both tracking accuracy and speed [2–16]. Bolme et al. [2] first employed CF and proposed minimum output sum of squared errors (MOSSE) filter to track objects and achieved impressive speed and favorable accuracy. Majority of the primitive trackers either relied on brightness information or used a simple color representation to describe the image. However, complex color features combined with brightness are more capable of providing excellent performance in target recognition and detection. Therefore, color name (CN) [21] investigated the contribution of color in the detection tracking framework and showed that color attributes can provide superior visual tracking performance. In addition, Ma et al. introduced the square gradient histogram feature (HOG) [22] in learned kernelized CFs (KCF) [8] and MA et al. in CF2 [23] proposed pre-trained deep convolutional networks to extract target features. The deep convolutional layer outputs semantic information of the encoded target which is robust to drastic appearance changes. However, the spatial resolution of the deep features is coarse and does not accurately localize the target. In contrast, HOG and CN features provide more accurate localization, but are not as robust to scenes when the appearance of the target changes.

Even though there is a widespread use of feature fusion methods [24, 25], it is still promising to improve tracker performance. The weights of most feature fusion methods are either fixed or random. The distinct importance of different features in various complex scenes is ignored. Since most existing CFs used a fixed learning rate, the tracker model is collapsed resulting in tracking failure when the target is obscured. In order to overcome the weaknesses of fixed weights and avoid the problem of pollution of tracking models, the fusion strategy of feature response maps proposed in this paper. It can update the weights dynamically and give full play to the advantages of different parts in different scenarios.

There are many challenges in the tracking task, such as target deformation, illumination variation, background clutter, scale variation, occlusion, etc. With the continuous improvement of the object tracking algorithm, the performance of the tracker in the scene of illumination variation and scale variation has been significantly improved, but there is still much room for enhancement in the robustness of the algorithm in the occlusion scene. Therefore, improving the anti-occlusion ability is still the focus and difficulty in the field of object tracking.

Earlier, researchers have proposed various algorithms to solve the occlusion in tracking process. Wang [26] combines a tracking algorithm with a detection algorithm to solve the problem of re-detecting the target when it is obscured or lost, but the tracking accuracy is low. Yang [27] proposed the object tracking algorithm based on spatio-temporal context information, which utilizes the background information of the target location, and uses it to track when the target is occluded to avoid losing the target, but its tracking robustness is not strong for the case of complete occlusion. Li [28] proposes an approach based on a blocking-based algorithm for object tracking, which can effectively deal with the occlusion problem. Although the above researches have improved the performance of the tracking algorithm to cope with occlusion, there is still plenty of room for improvement. Therefore, we propose the anti-occlusion mechanism. It can effectively overcome these difficulties and achieve favorable results on the benchmark dataset.

In this paper, we first propose a multi-feature response map dynamic fusion strategy. Specifically, we first obtain the tracking results of the target using the individual features separately. And then, we determine the fusion weights for the next frame by comparing the consistency of results between the individual features with the fused feature separately. In this way, adaptive weight factors can be applied for different features in different scenes to build a robust target model. Second, we propose an anti-occlusion mechanism based on response map peak judgment. Specifically, the Gaussian label response corresponds to each tracked target in the frequency domain. When the target is not occluded, the Gaussian response label has only one peak. When the target is in the occlusion case, the Gaussian tag response will have multiple peaks, the location of multiple peaks is the possible target location, and we determine the final location of the target by re-detection of the peak location. In addition, a high-confidence model tracking strategy is proposed based on the above two methods.

The main contributions of this paper are summarized as follows:

1. A multi-feature response map dynamic fusion strategy based on the consistency of individual features and fused features is proposed. It is able to improve the robustness and accuracy of the proposed algorithm by building the better object appearance model.
2. An anti-occlusion mechanism is proposed. It is well known the response map has multiple local peaks when the target is occluded. Furthermore, if the nonmaximal local peak is satisfied with our proposed conditions, we generate a new response map which is obtained by moving the center of the region of interest to the nonmaximal local peak position of the response map and re-extracting features. We then select the response map with the largest response value as the final response map.
3. A high-confidence model update strategy is designed to deal with model contamination. The target area contains a lot of information of the background when the target is occluded. Therefore, we design a model update strategy based on the anti-occlusion mechanism. When the anti-masking mechanism is triggered, the model update is stopped; otherwise, with a standard DCF model update strategy is utilized.

2 Related work

2.1 Correlation filter

Since Bolme et al. [2] first employed correlation filters and proposed MOSSE filter to track objects and achieved impressive speed and favorable accuracy. The correlation filters have been investigated extensively in visual tracking due to its competitive performance and high computational efficiency. Remarkable improvements have been made to this popular tracker to overcome some limitations. For example, Henriques et al. [8] proposed KCF by kernel tricks. A multichannel version of MOSSE was also investigated in [29]. More discriminative features are widely used, such as HOG [22], CN [21], and deep CNN features [26]. In addition, particle filter-based method [6], long-term tracking [30], and continuous convolution [9] have also been to be developed to improve tracking accuracy and robustness.

Correlation filter trackers have also made significant developments in other aspects. Zhang et al. [48] incorporated context information into filter learning. Danelljan et al. proposed SRDCF [14] tracker mitigates boundary effects by penalizing correlation filter factors. SKCF [31] tracker can use an adjustable Gaussian window to extract the target information, which obtains the same results as the SRDCF [14] tracker. At the same time, the boundary effect is well resolved. LCT [30] tracker presents an impressive solution that not only copes with the scale variation excellently, but also mitigates the boundary effect. Dai et al. [15] proposed an adaptive spatial constraint mechanism that can efficiently obtain a spatial weight to adapt target appearance changes, and therefore can obtain more robust target tracking results.

2.2 Feature fusion

Feature fusion is an extremely popular strategy in computer vision and is widely employed in various tasks. By fusing different features, more information of the target can be considered, resulting in improved accuracy and robustness. Along this line, a large number of tracking algorithms combine multiple different features [32]. For example, Danelljan et al. [9, 11] proposed a joint sparse model for multi-feature fusion, and the proposed tracking algorithm can improve tracking performance by efficiently utilizing multiple features and dynamically removing unreliable features. Ma et al. [30] used weighted entropy scheme to fuse multiple visual features to track the target. Li et al. in SAMF [24] proposed to combine HOG and CN features with the CF framework to further improve the tracking performance. Zhang et al. [21] proposed a multi-task correlation filter to track the target object by considering the interdependence among multiple features.

In the tracking scenes, different features model the appearance of the target object from different perspectives [33], which has different importance for the representation of the object. Deep convolutional features can better represent semantic information about the target and are robust to drastic appearance changes [34]. Deep features are also widely employed in object detection [35]. However, the spatial resolution of deep features is coarse and does not accurately localize the target. In contrast, HOG and CN features provide more accurate localization but are less uncertain about appearance changes [36]. The proposed tracker adaptively adjusts the weights of gradient, color and deep features to better build the appearance model of object in challenging tracking scenes.

2.3 Anti-occlusion

In the object tracking process, since the correlation filter will be updated in each frame, when the target is occluded, it may drift, which seriously affects the precision and robustness of the tracking algorithm.

TLD [37] introduces a detection module to solve the problem of target loss when occlusion happens, and performs a global search on each frame to locate the possible location of the target. Li et al. [28] proposed the RPT tracking algorithm, which uses a Monte Carlo framework to estimate the distribution of blocks with high confidence, and locates the global position of the target by finding the high-confidence position of the tracking target. Dong et al. [38] cached previous templates by constructing a classification pool. When heavy occlusion happens, the occlusion problem in it is solved by selecting the best classifier for re-detection. However, the tracker will lose the target in a long occlusion scene.

3 Methods

The procedure of the proposed method is shown in Fig. 1. The green arrows indicate the process of anti-occlusion mechanism and the light blue arrows indicate the steps of multi-feature response map adaptive fusion strategy. h is the learning rate of the filter in Eq. (12).

In this section, we describe the preliminary DCF tracking first, then introduce the multi-feature response map adaptive fusion strategy and the anti-occlusion mechanism. Finally, we describe the high-confidence model update strategy.

3.1 Overview of DCF

The correlation filter detects the location of the target by training a correlation filter w. A typical DCF model is centered on the target, and it is trained by the image area with the size of $M \times N$ is the training sample obtained by the cyclic shift of the image patch. The objective function is as follows:

$$\min_{w} \|Xw - Y\|_2^2 + \lambda \|w\|_2^2, \tag{1}$$

where λ is the regularization parameter to reduce the overfitting of the model, X is the matrix obtained by cyclic shift of our samples. Y is the Gaussian label that represents the ideal output of filtering learning. The closed solution of the objective function can be obtained by Fourier transform. The formula is shown below:

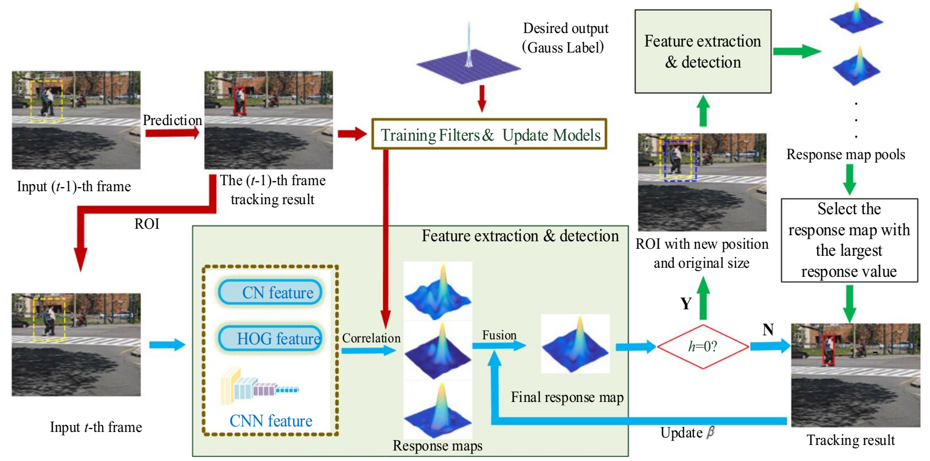

Desired output
(Gauss Label)
Feature extraction & detection
Prediction
Training Filters & Update Models
Input (t-1)-th frame
The (t-1)-th frame tracking result
ROI
ROI with new position and original size
Response map pools
Select the response map with the largest response value
Feature extraction & detection
CN feature
HOG feature
CNN feature
Correlation
Fusion
Response maps
Final response map
Y
N
h=0?
Input t-th frame
Update β
Tracking result

Fig. 1 Procedure of the proposed method

$$\hat{w}_d^{*t} = \frac{\hat{A}_d^t}{\hat{B}_d^t + \lambda}, \tag{2}$$

$$\hat{A}_d^t = \hat{Y} \odot \hat{X}_d^{*t}, \hat{B}_d^t = \sum_{i=1}^{D} \hat{X}_i^{*t} \odot \hat{X}_i^t, \tag{3}$$

where $\odot$ is the element-wise product, the hat symbol denotes the discrete Fourier transform (DFT) of a vector and $\hat{X}^*$ is the complex-conjugate of $\hat{X}$. t and $d(d \in \{1, \ldots, D\})$ denote the index and channel of the current frame, respectively. The response map can be calculated as follows:

$$R = \mathcal{F}^{-1}\left(\sum_{d=1}^{D} \hat{W}_d \odot \hat{X}_d^*\right). \tag{4}$$

The final position of the target can be obtained by finding the maximum value of the response map. Since the training samples by circular shift contain a lot of rich information of the target, it can train a filter with excellent performance, but in the process of generating negative samples from the base samples, the negative samples may have discontinuous edges, which will bring interference information to the filter. By adding the cosine window, the discontinuous regions outside the bottle region can be filtered out, and the tracking regions in the image block can be better highlighted, thus obtain a better training sample.

3.2 Multi-feature response map adaptive fusion strategy

Since HOG, CN, and deep (features extracted with CNN) features have achieved impressive performance in the field of target tracking, the selection and fusion of target features has become an important development direction of target tracking. Therefore, we propose a strategy for dynamic fusion of these three features in the framework of correlation filtering, and the main idea of the proposed tracker is to dynamically fuse multiple features to build the appearance model of the object. Figure 2 illustrates the process of feature fusion and update of fusion weights, the yellow box, black box, and blue box are the prediction results of CN feature, HOG feature and CNN feature response maps, respectively, and the red box is the prediction result obtained by fusing the three feature response maps. The weight of single feature response map is update by Eq. (8). The response maps are linearly added using the feature weights to obtain the final response maps. The individual feature fusion weight is updated by the consistency of the results obtained from the individual feature response map with the final response map.

Specifically, the method in this section uses HOG, CN and deep features to train three independent correlation filters w. In order to build the appearance model of the object in different view. $\mathrm{w}_{\mathrm{hog}}$, w_{cn} and $\mathrm{w}_{\mathrm{cnn}}$ represent the filters obtained by training the HOG, CN and deep features extracted from the target image block, respectively. To fuse multiple features, we let β_{hog}, β_{cn} and β_{cnn} be the weights of $\mathrm{w}_{\mathrm{hog}}$, w_{cn} and $\mathrm{w}_{\mathrm{cnn}}$, respectively. The filters $\mathrm{w}_{\mathrm{hog}}$, w_{cn} and $\mathrm{w}_{\mathrm{cnn}}$ are operated on the HOG, CN and deep feature map, respectively. The response maps F_{hog}, F_{cn} and F_{cnn} are obtained by Eq. (4). Therefore, there are three different kinds of object representations. In order to solve the above three

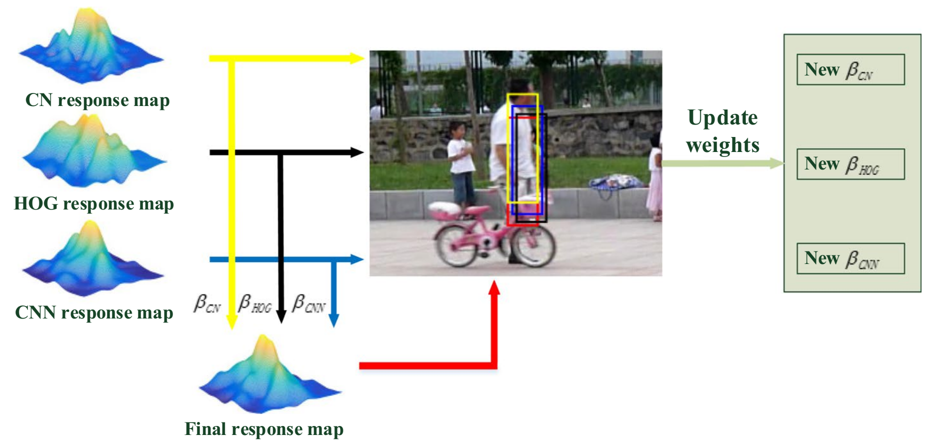


Fig. 2 Feature response map fusion and weight update

different and unattached regression subproblems, we linearly added these response maps according to the following equation:

$$F_{\text{final}} = \beta_{\text{hog}} F_{\text{hog}} + \beta_{\text{cn}} F_{\text{cn}} + \beta_{\text{cnn}} F_{\text{cnn}}. \tag{5}$$

The final result can be obtained by finding the maximum response value position in F_{final}. We adjust the weights of the three features based on the tracking results of the previous frame to take into account the different importance of the three features at different moments or scenes for building the target appearance model. To efficiently adjust the weights and better capture changes in the target appearance, the weight corresponding to a feature is determined by the agreement between the tracking result of that feature alone and the final tracking result F_{final}.

If the tracking result obtained with a feature alone is very closely matched with the final tracking result F_{final}, then the feature can simulate the appearance of the target well and the feature will be given a higher weight. On the other hand, if the tracking result of a feature is very different from the final tracking result F_{final}, then the feature is not suitable for modeling the target appearance in the current tracking scenario, and therefore the feature is assigned by a small weight. We obtain the prediction results P_{hog}, P_{cn}, P_{cnn} and P_{final} by F_{hog}, F_{cn}, F_{cnn} and F_{final} response maps, respectively. The overlap between the individual feature result and the final result indicates the importance of the individual features in the current frame. The agreement between the multiple features and the final result can be calculated as follows:

$$O_j^T = \frac{\text{Area}\left(P_j^T \cap P_{\text{final}}^T\right)}{\text{Area}\left(P_j^T \cup P_{\text{final}}^T\right)}, \tag{6}$$

where P_j^T denotes the results of single feature, P_{final}^T is the final tracking result.

In order to improve the robustness and tracking accuracy of the system, we need to take into account the n most recent consecutive reliable frames and calculate the temporal consistency of individual features, respectively. The specific definition of reliable frames will be given when the model update strategy is introduced. Therefore, considering the previous n reliable frames, we calculate the temporal consistency of single feature as follows (N_r denotes the set of the previous n reliable frames):

$$O_j = (1-\alpha)\left(\sum_{t\in N_r} O_j^t\right) + \alpha O_j^T, \tag{7}$$

where α is learning rate for agreement between the multiple features and the final result. α is a constant that controls the effect of the current frame feature consistency on the overall feature response map fusion weight. Finally, the weight of the single feature can be calculated as follows:

$$\beta_j = \frac{O_j}{\sum_{k\in\{\text{hog,cn,cnn}\}} O_k}, j \in \left\{\text{hog,cn,cnn}\right\}, \tag{8}$$

The larger the β is, the more suitable the current feature is to build the target appearance model in the current scene. In the first frame, we consider that all three features have the same importance, so we set β_{hog}, β_{cn}, β_{cnn} all to 0.33.

3.3 Anti-occlusion mechanism

The target is often occluded by other objects during tracking, resulting in tracking failure, which indicates that how to deal with the occlusion problem becomes a crucial concern for target tracking. After extensive experiments, it is found that there are multiple local peaks in the response map of the tracker when it is partially occluded. However, most current tracking algorithms select only the global peak as the final position of the target and ignore other local peaks. In fact, the features extracted from the target location are also polluted when the target is occluded, which causes the response value of the real location of the target to be low. In addition, the tracker model is updated in each frame and objects that obscure the target are added to the tracking model as tracking targets. However, since the tracking target model is not completely corrupted, the confidence level obtained for the region containing the real target is also higher than that of the general background map, so it means that the local peaks may also be the real position of the target. Therefore, when there are multiple peaks in the response map, we need to evaluate whether the local peaks are the real position of the target.

According to Fig. 3, it can be seen that when the tracking target is partially occluded and the localization is wrong, the response map shows multiple local peaks instead of one independent peak (local peaks do not include global peak). The isolated local peaks may be due to the effects of noise, while the coordinates of the local peaks of the arches may be the real location of the target. According to this characteristic, we design an accurate localization method based on the smoothness of the peaks, the smoothing function of the peaks is defined as follows:

$$SC = \sum_{m=-3}^{3}\sum_{n=-3}^{3}\left(R_{(x+n,y+m)} - R_{(x,y)}\right) \tag{9}$$

where $R_{(\mathrm{x},y)}$ represents the local peak of the response map, $R_{(\mathrm{x}+n,y+m)}$ represents the nearby points of the local peak. After extensive experiments, it has been shown that the

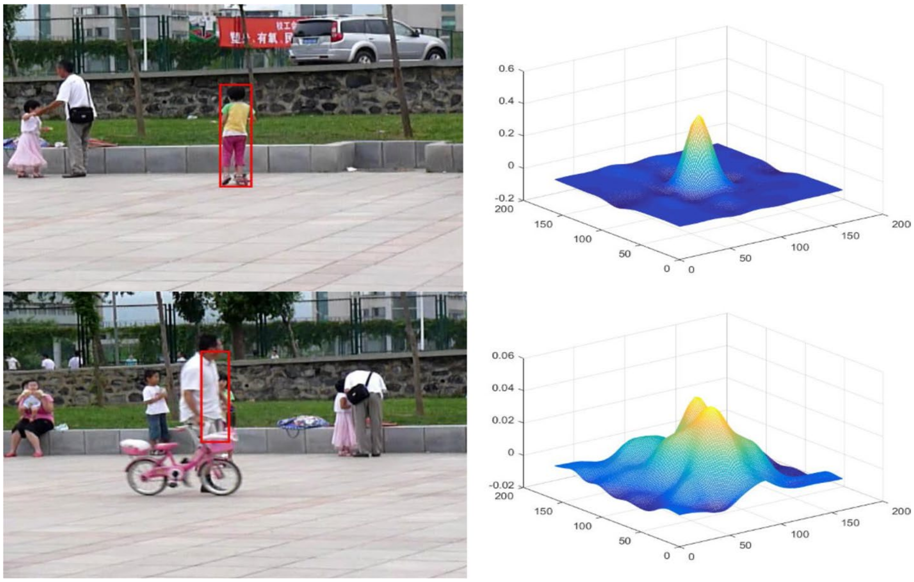


Fig. 3 Results of DCF algorithm for girl2 video sequence and its response map

performance of the algorithm is most reliable when the m, n in the range of [− 3,3], and *SC* represents the smoothing coefficient. It means that the local peaks are likely to be the real tracking target position when *SC* of the local peaks is greater than *SC* of the global peaks. We generate a new response map which is obtained by moving the center of the region of interest to the nonmaximal local peak position of the response map and re-extracting features. We then select the response map with the largest response value as the final response map. Due to the balance between performance and speed, we can select at most the three points of the local peak with the highest smoothness.

3.4 Model update strategy

The features extracted from the target may pollute the tracking model when the target is occluded, so the proposed tracking algorithm has to decide whether to update the tracking model or not. If the tracking results are reliable for the current frame, we update the model to better represent the appearance changes. If the result of the current frame is unreliable, we do not update the model to avoid contaminating it. The online update of the numerator $\hat{A}_d^t$ and denominator $\hat{B}_d^t$ of the filter w is as follows:

$$\hat{A}_d^t = (1-h)\hat{A}_d^{t-1} + h\hat{Y} \odot \hat{X}_d^{*t}, \tag{10}$$

$$\hat{B}_d^t = (1-h)\hat{B}_d^{t-1} + h\sum_i^D \hat{X}_i^{*t} \odot \hat{X}_i^t, \tag{11}$$

where h is the learning rate of the filter w and t is the index of the current frame.

If the local peak smoothing coefficient $\mathrm{SC}_{\mathrm{local}}$ of the response map is greater than the global peak smoothing coefficient $\mathrm{SC}_{\mathrm{global}}$, then the target position of the current frame contains background information and the tracking result of the current frame is unreliable. We reduce the learning rate of the model to avoid the subsequent frame tracking failure caused by model contamination. Otherwise, the tracking result of the current frame is reliable. The model of the proposed algorithm is updated with the normal learning rate. The learning rate h is set by Eq. (12):

$$h = \begin{cases} 0, & if \left(\mathrm{SC}_{\mathrm{local}} > \mathrm{SC}_{\mathrm{global}}\right)\&\&(\mathrm{NLP} \geq 1) \\ \tau, & otherwise \end{cases}, \tag{12}$$

where τ is the learning rate of the standard DCF, $\mathrm{SC}_{\mathrm{local}}$ and $\mathrm{SC}_{\mathrm{global}}$ are the smoothing coefficients of the local peak and global peak in the response map.

4 Experimental setup

The experiments of tracking performance evaluation are conducted using MATLAB R2017a on a PC with an Intel i5-7400 processor (3.0 GHz), 8G RAM and a GeForce GTX1050Ti GPU. The MatConvNet toolbox is used for extracting the deep features from VGG-19. We extract the outputs of the conv4-4 convolutional layers of VGG-19 as the deep feature. We follow the parameters in standard DCF trackers to construct the proposed tracker. The regularization weight λ in Eq. (1) for HOG, CN and deep features are set to 0.01. The size of the collection N_r in Eq. (7) is set to 2. The learning rate τ for the standard DCF in Eq. (12) is 0.025. At the first frame, β_{hog}, β_{cn} and β_{cnn} are all set to 0.33.

The intersection-over-union (IoU) and center location error (CLE) between tracking results and ground truth bounding boxes are used to evaluate a tracker quantitatively [39]. By setting thresholds for IoU and CLE, we get average success rate and precision over all frames, respectively. With a range of thresholds, we finally obtain the success plots of IoU and precision plots for CLE. The area under curve (AUC) of success plots and the precision of CLE at 20 pixels (mean overlap precision) are regarded as the final metrics for each tracker.

5 Results and discussion

In this section, at first, we illustrate the effectiveness of my tracker on the OTB2013 [39], OTB2015 [40], TC128 [42], UAV123 [43] datasets, we compare it with the current dominant advanced object tracker on the benchmark, so as to illustrate the superiority and performance of our tracker. Then, we do the ablation experiments on the OTB2015 [40] dataset and validate the effectiveness of the methods illustrated in Sects. 3.2 and 3.3.

5.1 Evaluation results on the OTB2013 and OTB2015

OTB is widely used for performance evaluation in the visual object tracking. OTB2015 [40] contains 98 sequences with 100 targets. OTB2013 [39] is a subset of OTB2015 [40] and contains 50 sequences with 51 targets. OTB has 11 challenging attributes. The detailed information can be obtained in the reference. We make a comprehensive comparison between our tracker and other 10 trackers.

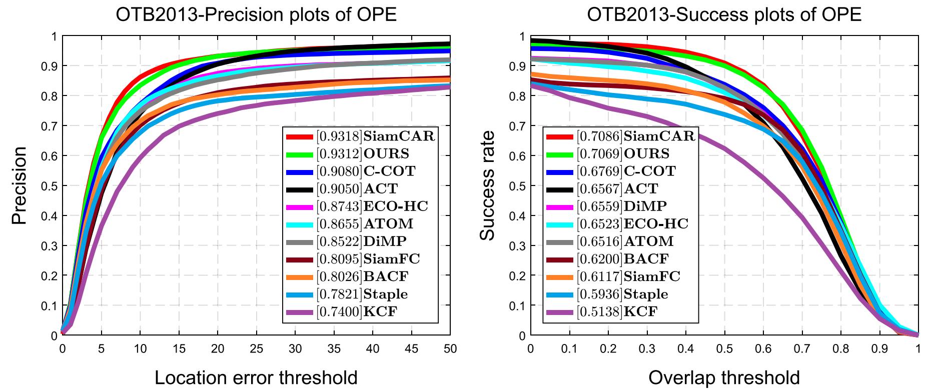


Fig. 4 Performance evaluation results of different algorithms on OTB2013 dataset

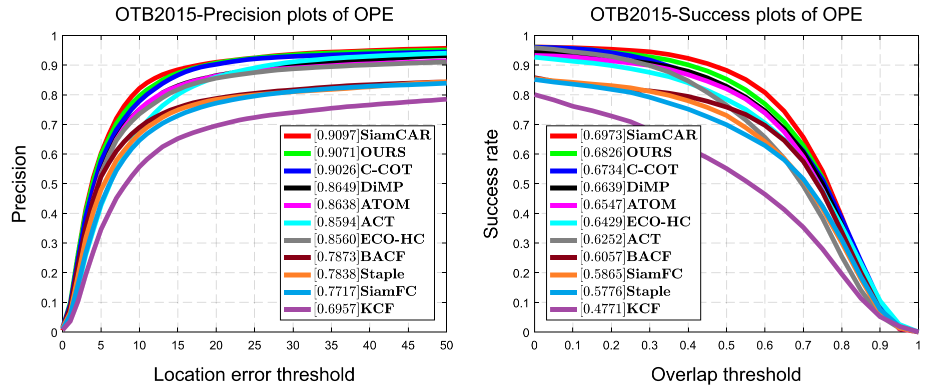


Fig. 5 Performance evaluation results of different algorithms on OTB2015 dataset

5.1.1 Overall performance evaluation

The evaluation results presented in OTB2013 [39] and OTB2015 [40] by comparing the precision and success plots of the trackers are shown in Figs. 4 and 5. The comparative results show that our tracker performs favorably against other state-of-the-art trackers. We compared 10 advanced tracking algorithms including C-COT [9], SiamCAR [44], DiMP [41], ATOM [45], SiamFC [46], BACF [12], ECO-HC [11], Staple [47], ACT [17], KCF [8] in our experiments. KCF [8], C-COT [9], ECO-HC [11] and BACF [12] are correlation filtering based algorithms. SiamFC [46], ATOM [45], DiMP [41] and SiamCAR [44] are methods based on deep learning for object tracking. Our proposed tracker achieves the impressive precision of 93.12% and 90.71% on the OTB2013 [39] and OTB2015 [40] separately, among all the compared trackers. As for the success rate, our tracker achieves the impressive success rate of 70.69% and 68.26% on the relative two benchmarks, respectively.

5.1.2 Attributes-based performance evaluation

For a more detailed analysis of the challenges in the image sequences, the OTB2015 benchmark [40] classifies video sequences into 11 different challenging attributes, including illumination variation (IV), out-of-plane rotation (OPR), scale variation

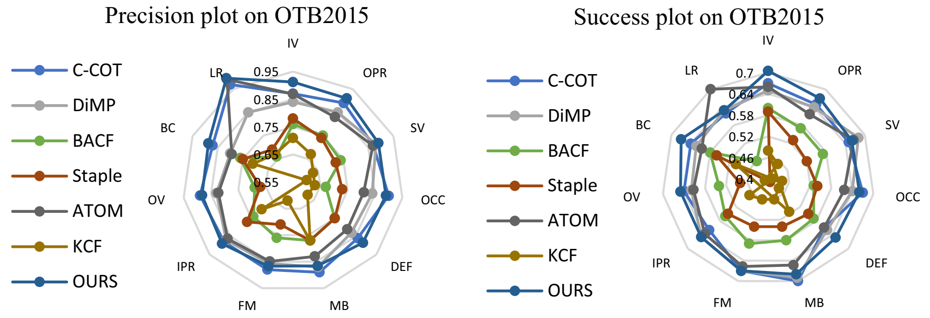


Fig. 6 Radar plot of precision and success based on tracking results on the OTB2015 dataset in 11 challenging attributes

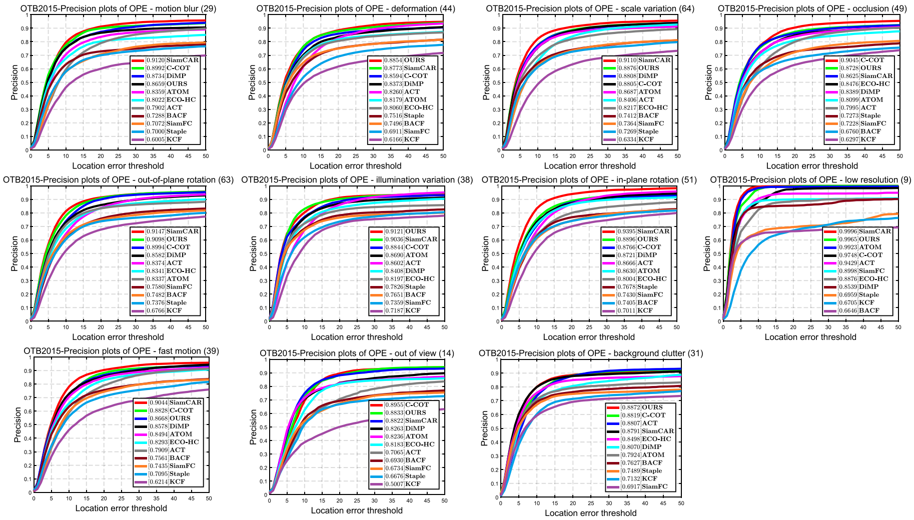


Fig. 7 Precision plots based on tracking results on the OTB2015 dataset in 11 challenging attributes

(SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out of view (OV), background clutter (BC) and low resolution (LR).

Figure 6 is a radar plot of the experimental results of the proposed algorithm with other algorithms such as C-COT [9], DiMP [41], ATOM [45], BACF [12], Staple [47],KCF [8], which clearly illustrates the excellent performance of the proposed algorithm in the face of various challenges.

From Fig. 7, it can be seen that our algorithm achieves the excellent performance in accuracy with other comparable algorithms under DEF, SV, IV, BC, OPR, IPR and LR challenge attributes, with experimental results are 88.54%, 88.76%, 91.21%, 90.98%, 88.96% and 99.65%. Meanwhile, favorable experimental results are obtained under OCC, MB, FM and OV challenge attributes. As can be seen in Fig. 8, our proposed algorithm also achieves an impressive success rate under a variety of challenging attributes. This shows that the proposed tracker can effectively handle occlusion, illumination and distortion. The proposed anti-obscuration mechanism handles tracking failures caused by occlusion. As for deformation and illumination, a multi-feature response map adaptive fusion strategy is adopted construct a better model of capturing object appearance changes.

5.2 Evaluation results on the TC128

The full name of TC128 [42] dataset is Temple Color dataset, which contains 128 annotated video sequences. Similar to the OTB dataset, the TC128 [42] also has 11 different interference factors (readers can refer to [39] for more information). The video sequences of TC128 [42] contain more chromatic information which provides more discriminative information. We compared 10 advanced tracking algorithms including C-COT [9], STRCF [13], SRDCF [14], BACF [12], ECO-HC [10], MCPF [6] in our experiments. Figure 9 shows the evaluation results for the mean overlap precision (OP) from the precision plot and AUC from the success plot of the compared trackers in the TC128.Our proposed tracker achieves the best OP rate of 77.30%, among all the compared trackers. As for the success rate, our tracker also achieves the best AUC of 57.30%.

5.3 Evaluation results on the UAV123

The UAV123 [43] dataset is a collection of 123 high-definition color short sequence video images that were captured during low-altitude aerial photography. As the unmanned aerial vehicle (UAV) moving, the view of the camera mounted on the UAV keeps changing, so the aspect ratio of the target in subsequent video frames in the UAV123 [43] database changes significantly with respect to the target in the initial frame. This dataset contains a large amount of scene, target and motion information.

Figure 10 shows the OP from the precision plot and AUC from the success plot. Our tracker performs very well in the UAV123 [43] dataset which achieves the OP of 73.2% and AUC of 50.2%. Our proposed tracker outperformed ECO-HC [11] by 2.2% and 0.6% in OP and AUC, respectively. These two trackers both performs well in the TC128 [42] benchmark similarly. However, compared with other trackers, for example, our

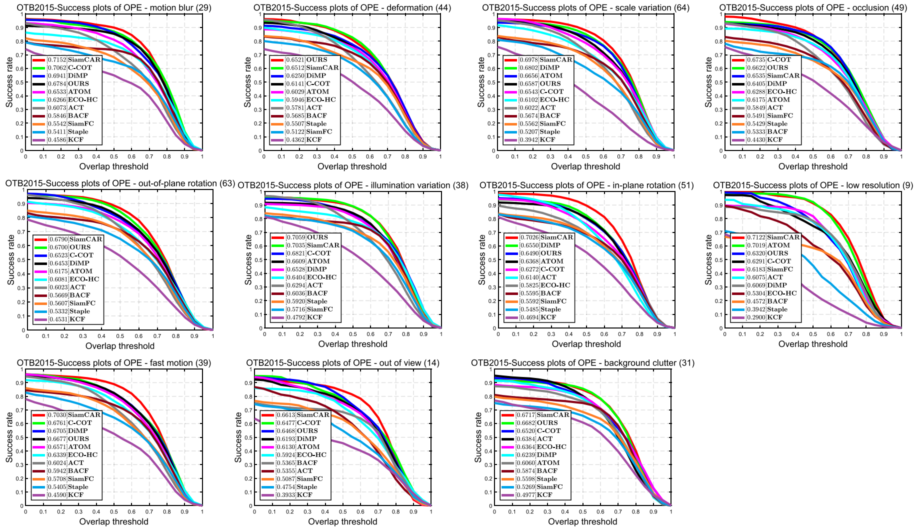
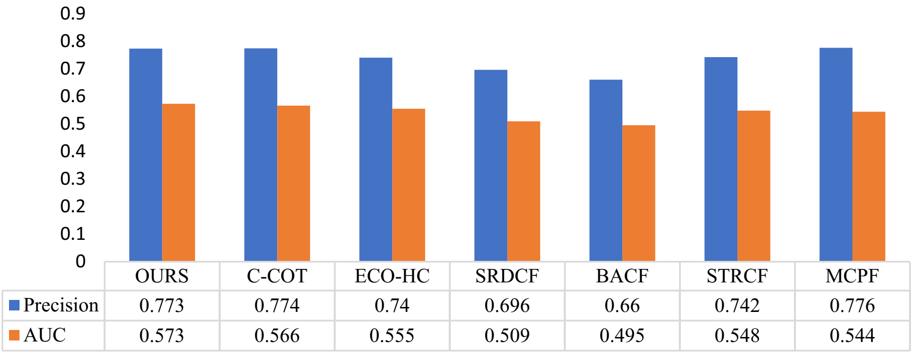
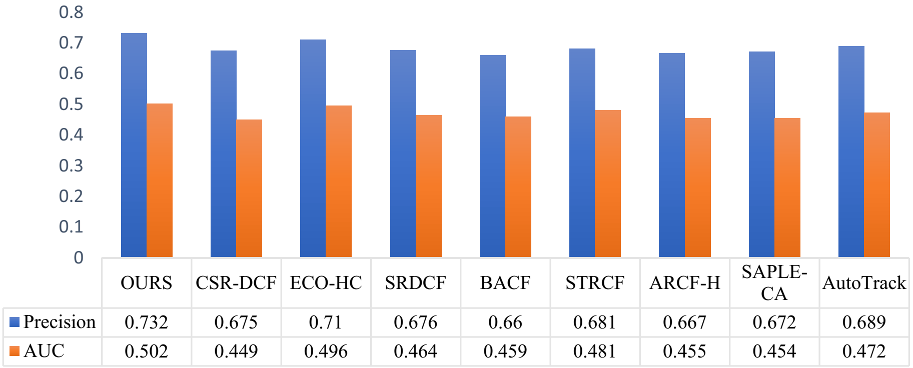


Fig. 8 Success plots based on tracking results on the OTB2015 dataset in 11 challenging attributes



	OURS	C-COT	ECO-HC	SRDCF	BACF	STRCF	MCPF
🟦 Precision	0.773	0.774	0.74	0.696	0.66	0.742	0.776
🟧 AUC	0.573	0.566	0.555	0.509	0.495	0.548	0.544

Fig. 9 Comparison of our tracker and other 6 excellent trackers in terms of precision and success rate on TC128



	OURS	CSR-DCF	ECO-HC	SRDCF	BACF	STRCF	ARCF-H	SAPLE-CA	AutoTrack
🟦 Precision	0.732	0.675	0.71	0.676	0.66	0.681	0.667	0.672	0.689
🟧 AUC	0.502	0.449	0.496	0.464	0.459	0.481	0.455	0.454	0.472

Fig. 10 Comparison of our tracker and other 8 excellent trackers in terms of precision and success rate on UAV123

proposed tracker shows much better performance than SRDCF [14], STRCF [13] and BACF [16]. This indicates our tracking method also performs impressively in the face of video sequences with continuous changes in shooting angle.

5.4 Ablation studies and effectiveness discussion

To verify the effectiveness of the tracker, our tracker is compared itself with different modules enabled. Baseline is the standard DCF algorithm based on multiple features. By only using the multi-feature response map adaptive fusion strategy, we obtain the Baseline–DF without anti-occlusion mechanism (Sect. 3.3) and model update strategy (Sect. 3.4). The Baseline–AO represents the tracker adopts the anti-occlusion mechanism (Sect. 3.3) and model update strategy (Sect. 3.4). Baseline–DF–AO is our final tracker which combines the anti-occlusion mechanism (Sect. 3.3), multi-feature response map adaptive fusion strategy (Sect. 3.2) and model update strategy (Sect. 3.4). The experimental results on the OTB2013 benchmark [39] are shown in Fig. 11. Benefiting from the added anti-occlusion mechanism, Baseline–AO outperforms Baseline in terms of accuracy and success rate by 1.58% and 1.13%, respectively. Besides, Baseline–DF also outperforms Baseline in terms of accuracy and success rate by 2.3% and 1.86%, respectively, which indicates the effectiveness of the multi-feature adaptive fusion

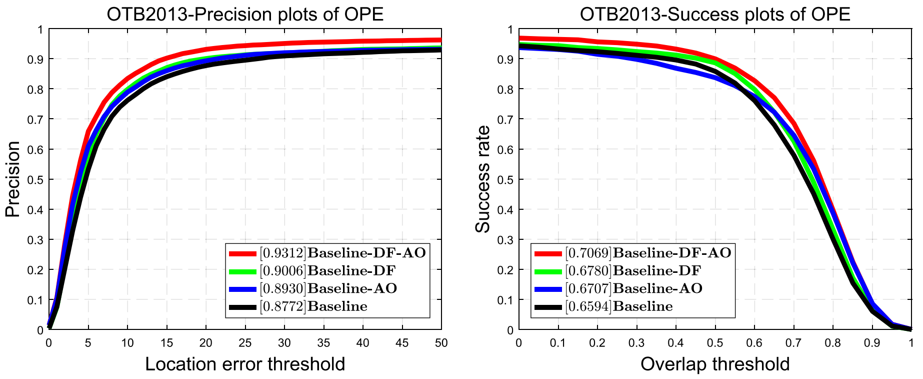


Fig. 11 Performance evaluation results of compared itself with different modules enabled on OTB2013 dataset

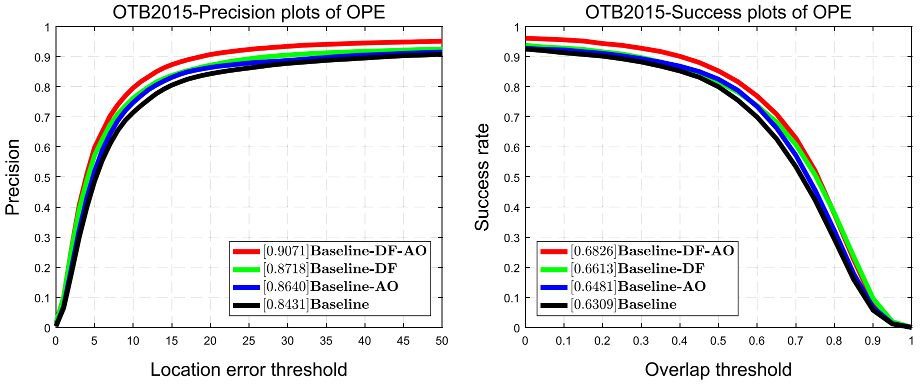


Fig. 12 Performance evaluation results of compared itself with different modules enabled on OTB2015 dataset

strategy. Finally, the final proposed algorithm (Baseline–DF–AO) also has an impressive performance at OTB2013 [39]. As can be seen in Fig. 12, the experimental results on the OTB2015 [40] are similar to those on the OTB2013 [39], which is sufficient to illustrate the effectiveness of the proposed method.

5.5 Qualitative evaluation results

During the tracking of the target, the environmental changes around the target and the deformation of the target may lead to the tracking failure of the tracker. We selected 10 image sequences with challenge attributes on the OTB2015 dataset [40] and compared them with the five trackers SRDCF [14], BACF [12], C-COT [9], KCF [8]. As shown in Fig. 11, it is clear that our tracker has higher robustness compared to other advanced trackers, and our tracker performs better than several other trackers in a variety of challenging scenes. For example, in the shaking sequence, between frames 17 and 104, the environment of the target suffers a drastic lighting change, and some trackers fail to track one after another in this sequence, while our tracker still tracks the target accurately, and is not affected by the lighting change. In the Dragonbaby sequence, the comparative tracker fails to track in this sequence due to the fast motion of the target, while our proposed tracker still tracks the target accurately. All these results indicate that the proposed tracking algorithm has not only impressive accuracy, but also extreme robustness (Fig. 13).

6 Conclusion

In this paper, we have improved the DCF tracker in two major terms, one is the adaptive fusion strategy of target features and the other is the anti-occlusion mechanism. The multi-feature response map adaptive fusion strategy is proposed to better build the appearance model of the target. Different features have different importance for the representation of the outside model of the target when the environment around the tracking target changes. We propose a dynamic feature fusion method that dynamically adjusts the weights of feature fusion by the overlap rate between the result of each feature and the final result. The anti-occlusion mechanism is proposed to cope with the scene where the target is occluded. When the target is occluded, the response map will show multiple local peaks, which are all possible target locations. The final position of the target is further determined by detecting the position of the local peaks of the response map. This method can effectively deal with the case of an obscured target and enhance the robustness of the system. To validate the advancement of our proposed method, we did experiments on several datasets, including OTB2013 [39], OTB2015 [40], UAV123 [43], TC128 [42]. In addition, we also did ablation experiments on OTB2015 [40] as a way to validate the effectiveness of individual methods. From the experimental results, it is easy to know that our algorithm has a large advantage over other advanced algorithms and performs very well on several datasets. Our proposed algorithm achieves good performance in terms of accuracy and success rate, while runs only 5.89 fps. In future work, we intend to introduce context and background from the target into the tracker to better characterize the target's motion, and on the other hand, we intend to extract deep features of different

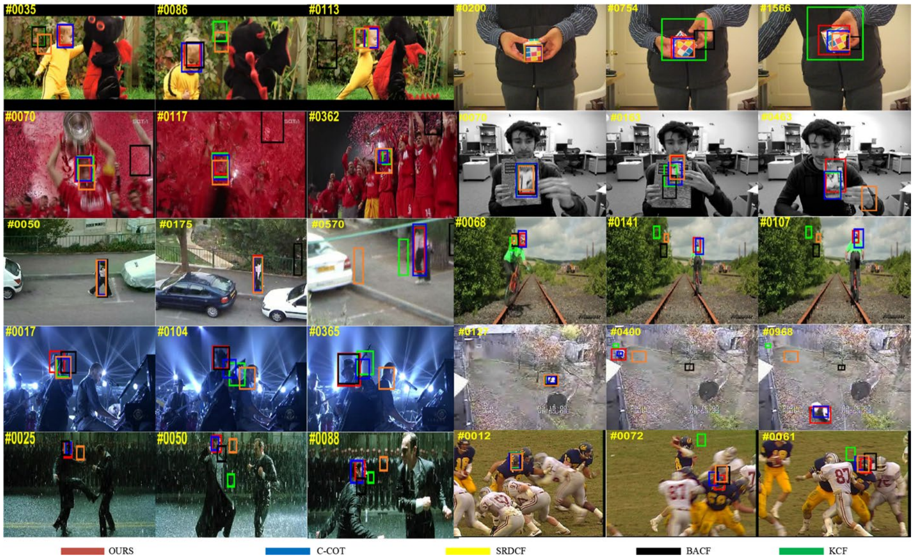


Fig. 13 Qualitative evaluation of the proposed algorithm, C-COT [9], SRDCF [14], BACF [12], KCF [8] methods on ten challenging video sequences

layers and deep features of different network structures to characterize the tracking target. Besides, the improvement of the algorithm efficiency is also an important task for future work.

Abbreviations

AUC: Area under curve; CF: Correlation filter; CNN: Convolutional neural network; FFT: Fast Fourier transformation; DFT: Discrete Fourier transformation; HOG: Histogram of oriented gradient; NLP: Number of local peaks; CN: Color name; IoU: Intersection-over-union; CLE: Center location error.

Acknowledgements

The authors would like to thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Authors' contributions

All authors took part in the discussion of the work described in this paper. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61972056 and 61901061, the Science Fund for Creative Research Groups of Hunan Province under Grant 2020JJ1006, the Postgraduate Training Innovation Base Construction Project of Hunan Province under Grant 2019-248-51, the Basic Research Fund of Zhongye Changtian International Engineering Co., Ltd. under Grant 2020JCYJ07, the Scientific Research Fund of Hunan Provincial Education Department under Grant 19C0028, and the Young Teachers' Growth Plan of Changsha University of Science and Technology under Grant 2019QJCZ011.

Availability of data and materials

We used publicly available dataset in order to illustrate and test our methods. The OTB dataset can be found in http://cvlab.hanyang.ac.kr/trackerbenchmark/datasets.html and the UAV123 dataset can be found in https://ivul.kaust.edu.sa/Pages/Dataset-UAV123.aspx. The TC128 dataset can be found in https://dabi.temple.edu/~hbling/data/TColor-128.html.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China. [2]Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114, China.

Received: 13 July 2021 Accepted: 8 March 2022
Published online: 18 March 2022

References

1. A.W.M. Smeulders, D.M. Chu, R. Cucchiara et al., Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1442–1468 (2014)
2. D. S. Bolme, J.R. Beveridge, B.A. Draper, et al., Visual object tracking using adaptive correlation filters, IEEE Conference on Computer Vision and Pattern Recognition (2010), pp. 2544–2550
3. F. Bonin-Font, A. Ortiz, G. Oliver, Visual navigation for mobile robots: A survey. J. Intell. Robot Syst. **53**, 263 (2008)
4. S. Jung, Y. Kim, E. Hwang, Real-time car tracking system based on surveillance videos. EURASIP J. Image Video Process. **2018**(1), 133 (2018)
5. G. Zhang, J. Yang, W. Wang, Y.H. Hu, J. Liu, Adaptive visual target tracking algorithm based on classified-patch kernel particle filter. EURASIP J. Image Video Process. **2019**(1), 20 (2019)
6. T. Zhang, C. Xu, M.H. Yang, Multi-task correlation particle filter for robust object tracking, IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 4335–4343.
7. X. Li, W. Hu, C. Shen et al., A survey of appearance models in visual object tracking. ACM Trans. Intell. Syst. Technol. (TIST) **4**(4), 1–48 (2013)
8. J.F. Henriques, R. Caseiro, P. Martins et al., High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2015)
9. M. Danelljan, A. Robinson, F. Khan, M. Felsberg, Beyond correlation filters: Learning continuous convolution operators for visual tracking, European Conference on Computer Vision (2016), pp. 472–488.
10. E. Kermani, D. Asemani, A robust adaptive algorithm of moving object detection for video surveillance. EURASIP J. Image Video Process. **2014**(1), 27 (2014)
11. M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, ECO: Efficient Convolution Operators for Tracking, IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 6931–6939
12. H. K. Galoogahi, A. Fagg, S. Lucey, in ICCV. Learning Background-Aware Correlation Filters for Visual Tracking (2017), pp. 1144–1152.

13. F. Li, C. Tian, W. Zuo, et al., Learning spatial-temporal regularized correlation filters for visual tracking, IEEE Conference on Computer Vision and Pattern Recognition (2018), pp. 4904–4913.
14. M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, IEEE International Conference on Computer Vision (2015), pp. 4310–4318.
15. K. Dai, D. Wang, H. Lu, C. Sun, J. Li, Visual Tracking via Adaptive Spatially-Regularized Correlation Filters, IEEE Conference on Computer Vision and Pattern Recognition, (2019), pp. 4665–4674.
16. J. Zhang, J. Sun, J. Wang, X.G. Yue, Visual object tracking based on residual network and cascaded correlation filters. J. Ambient. Intell. Humaniz. Comput. **12**(8), 8427–8440 (2021)
17. B. Chen, D. Wang, P. Li, S. Wang, H. Lu, et al. Real-time 'Actor-Critic' Tracking. Proceedings of the European conference on computer vision (ECCV). 2018: 318–334.
18. H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 4293–4302.
19. Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan and S. Wang, in ICCV. Learning Dynamic Siamese Network for Visual Object Tracking (2017), pp. 1781–1789.
20. S. Avidan, Support vector tracking. IEEE Trans. Pattern Anal. Mach. Intell. **26**(8), 1064–1072 (2004)
21. M. Danelljan, F. S. Khan, M. Felsberg, In CVPR, Adaptive Color Attributes for Real-Time Visual Tracking (2014), pp. 1090–1097.
22. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition (2005), pp. 886–893.
23. C. Ma, J.-B. Huang, X. Yang, M.-H. Yang, in ICCV. Hierarchical convolutional features for visual tracking (2015), pp. 3074–3082.
24. Y. Li, J. Zhu, A scale adaptive kernel correlation filter tracker with feature integration. In European conference on computer vision (2014), pp. 254–265.
25. O. Khalid, J.C. SanMiguel, A. Cavallaro et al., Multi-tracker partition fusion. IEEE Trans. Circuits Syst. Video Technol. **27**(7), 1527–1539 (2016)
26. J. Wang, H. Yang, N. Xu et al., Long-term target tracking combined with re-detection. EURASIP J. Adv. Signal Process. **2021**(1), 2 (2021)
27. X. Yang, S. Zhu, D. Zhou et al., An improved target tracking algorithm based on spatio-temporal context under occlusions. Multidim. Syst. Sign. Process. **31**, 329–344 (2020)
28. Y. Li, J. Zhu, Ho s C H. Reliable patch trackers: robust visual tracking by exploiting reliable patches. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. (Boston, MA, USA 2015), pp. 353–361.
29. H. K. Galoogahi, T. Sim and S. Lucey, Multi-channel Correlation Filters. IEEE International Conference on Computer Vision (2013), pp. 3072–3079.
30. C. Ma, X. Yang, C. Zhang, et al., Long-term correlation tracking, IEEE conference on computer vision and pattern recognition (2015), pp. 5388–5396
31. A. S. Montero, J. Lang, R. Laganière, Scalable Kernel Correlation Filter with Sparse Feature Integration. IEEE International Conference on Computer Vision Workshop. 2015: 587–594.
32. J. Zhang, Y. Liu, H. Liu, J. Wang, Y. Zhang, Distractor-aware visual tracking using hierarchical correlation filters adaptive selection. Appl. Intell. (2021). https://doi.org/10.1007/s10489-021-02694-8
33. S. He, Z. Li, Y. Tang, Z. Liao, F. Li, S.J. Lim, Parameters compressing in deep learning. CMC: Comput. Mater. Continua **62**(1), 321–336 (2020)
34. J. Zhang, X. Jin, J. Sun, Spatial and semantic convolutional features for robust visual object tracking. Multimed. Tools Appl. **79**(21–22), 15095–15115 (2020)
35. J. Zhang, W. Wang, C. Lu, J. Wang, Lightweight deep network for traffic sign classification. Ann. Telecommun. **75**(7–8), 369–379 (2020)
36. J. Zhang, X. Jin, J. Sun, J. Wang, K. Li, Dual model learning combined with multiple feature selection for accurate visual tracking. IEEE Access **7**, 43956–43969 (2019)
37. Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409–1422 (2011)
38. X. Dong, J. Shen, D. Yu et al., Occlusion-aware real-time object tracking. IEEE Trans. Multimedia **19**(4), 763–771 (2016)
39. Y. Wu, J. Lim, M.-H. Yang, Online object tracking: a benchmark, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2013), pp. 2411–2418.
40. Y. Wu, J. Lim, M. H. Yang, Object Tracking Benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015)
41. G. Bhat, M. Danelljan, L. V. Gool, et al. Learning discriminative model prediction for tracking[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (2019), pp. 6182–6191.
42. P. Liang, E. Blasch, H. Ling, Encoding color information for visual tracking: algorithms and benchmark. IEEE Trans. Image Process. **24**(12), 5630–5644 (2015)
43. M. Mueller, N. Smith, B. Ghanem. A benchmark and simulator for uav tracking. European conference on computer vision (Springer, Cham, 2016), pp. 445–461.
44. D. Guo, J. Wang, Y. Cui, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020), pp. 6269–6277.
45. M. Danelljan, G. Bhat, F. S. Khan, et al. Atom: Accurate tracking by overlap maximization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 4660–4669.
46. L. Bertinetto L, J. Valmadre, JF. Henriques, et al. Fully-convolutional siamese networks for object tracking. European conference on computer vision (Springer, Cham, 2016), pp. 850–865.
47. L. Bertinetto, J. Valmadre, S. Golodetz, et al., Staple: Complementary learners for real-time tracking, IEEE conference on computer vision and pattern recognition (2016), pp. 1401–1409.
48. Zhang, K., Zhang, L., Liu, Q., Zhang, D., & Yang, M. H. (2014). Fast visual tracking via dense spatio-temporal context learning. *European conference on computer vision*, pp 127–141.

Publisher's Note

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