Journal of Ambient Intelligence and Humanized Computing

Visual Object Tracking Based on Residual Network and Cascaded Correlation Filters

--Manuscript Draft--

Manuscript Number:	
Full Title:	Visual Object Tracking Based on Residual Network and Cascaded Correlation Filters
Article Type:	Original Research
Keywords:	Object Tracking, Deep Learning, Residual Network, Resnet Features, Cascaded Correlation Filters
Corresponding Author:	Jianming Zhang Changsha University of Science and Technology CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Changsha University of Science and Technology
Corresponding Author's Secondary Institution:	
First Author:	Jianming Zhang
First Author Secondary Information:	
Order of Authors:	Jianming Zhang
	Juan Sun
	Jin Wang
	Xiao-Guang Yue
Order of Authors Secondary Information:	
Funding Information:	National Natural Science Foundation of China (61972056) — Prof. Jianming Zhang National Natural Science Foundation of China (61772454) — Prof. Jianming Zhang National Natural Science Foundation of China (61811540410) — Prof. Jianming Zhang
Abstract:	Great progress is made in the field of object tracking recently. Especially, trackers based on deep learning and correlation filters both have achieved good performance. However, object tracking still faces some challenging problems such as deformation and illumination. In such kinds of situations, the accuracy and precision of tracking algorithms plunge as a result. It is imminent to find a solution to this situation. In this paper, we propose a tracking algorithm based on features extracted by residual network called Resnet features and cascaded correlation filters to improve the precision and accuracy. Firstly, features extracted by deep residual network trained on other image processing datasets, are robust enough and retain higher resolution, therefore, we exploit Resnet-101 pretrained offline to obtain features extracted by middle and high layers for target appearance model representation. Resnet-101 is deeper compared with other deep neural networks which means it contains more semantic information. Then, the method we propose to combine our correlation filters are superior. We propose cascaded correlation filters generated by handcraft, middle-level and high-level features from residual network to gain better competence. Handcraft features localize target precisely because they contain more spatial details while Resnet features are robust to the target appearance change because they retain more semantic information. Finally, we conduct extensive experiments on OTB2013 and OTB2015 benchmark. The experimental results show that our tracker achieves high performance under all kinds of challenges and performs favorably against other

	state-of-the-art trackers.

Journal of Ambient Intelligence and Humanized Computing manuscript No.
(will be inserted by the editor)

Visual Object Tracking Based on Residual Network and Cascaded Correlation Filters

Jianming Zhang · Juan Sun · Jin Wang · Xiao-Guang Yue

Received: date / Accepted: date

Abstract Great progress is made in the field of object tracking recently. Especially, trackers based on deep learning and correlation filters both have achieved good performance. However, object tracking still faces some challenging problems such as deformation and illumination. In such kinds of situations, the accuracy and precision of tracking algorithms plunge as a result. It is imminent to find a solution to this situation. In this paper, we propose a tracking algorithm based on features extracted by residual network called Resnet features and cascaded correlation filters to improve the precision and accuracy. Firstly, features extracted by deep residual network trained on other image processing datasets, are robust enough and retain higher resolution, therefore, we exploit Resnet-101 pretrained offline to obtain features extracted by middle and high layers for target appearance model representation. Resnet-101 is deeper compared with other deep neural networks which means it contains more semantic information. Then, the method we propose to combine our correlation filters are superior. We propose cascaded correlation filters generated by handcraft, middle-level and high-level features from residual network to gain better competence. Hand-

J. Zhang
Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
E-mail: jmzhang@csust.edu.cn

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

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

X.G. Yue
Rattanakosin International College of Creative Entrepreneurship, Rajamangala University of Technology Rattanakosin, Nakhon Pathom 73170, Thailand

craft features localize target precisely because they contain more spatial details while Resnet features are robust to the target appearance change because they retain more semantic information. Finally, we conduct extensive experiments on OTB2013 and OTB2015 benchmark. The experimental results show that our tracker achieves high performance under all kinds of challenges and performs favorably against other state-of-the-art trackers.

Keywords Object Tracking · Deep Learning · Residual Network · Resnet Features · Cascaded Correlation Filters

1 Introduction

Given the status (specified by a bounding box with coordinate, width and height) of an unknown target in the first frame, object tracking is a task of predicting the status of the target in the following frames. Object tracking is a basic but significant problem in the field of computer vision [32, 5, 19, 20]. And it has a wide range of practical applications such as video surveillance system, video-based human-computer interaction, flight control system in modern militarization and aerospace industry. Magnificent progress are made in object tracking in the past decade. Some celebrated tracking research groups propose widely-used benchmarks and trackers with high accuracy and good competence. Nevertheless, object tracking is still a hard nut to crack as a result of many challenges such as illumination changes and scale variation in the continuous video frames [35]. With the fast development of object tracking, more dedicated tracking platforms such as OTB2013 [30], LaSOT [12], UAV123 [22], Temple-Color 128 [18], which have large scale dataset with higher quality and more challenges, are proposed. Therefore, it ratchets up the difficulty in object tracking and requires higher performance with respect to trackers.

The selection of target appearance models is of great importance in the entire tracking process. On account of stochastic targets in the tracking frames, the generalizing ability and adaptability of tracking algorithms are considered to be one of the most important part. Both of them determine the accuracy and precision directly [39]. There are rich representations of different target appearance models, for instance, handcraft features put forward in the early days and deep learning features extracted from pretrained network models like Convolutional Neural Network (CNN) offline. Much attempt has been made to exploit traditional handcraft features like HOG (Histogram Oriented Gradient) [6], CN (Color Names) [10], color histogram [1] and Haar-like features [24]. Handcraft features require less computational ability but represent clearly [33]. Those features mentioned above, especially HOG, are capable of coping with target variations. Neural network and some other machine learning method have been applied to many different fields and has good performance as well [25, 27, 40, 26]. Compared with handcraft features, CNN features gains better performance in image processing [36, 38], object tracking included. CNN features are classified into three types: shallow features, middle features and high features, which address different kinds of subproblems in object tracking. Shallow and middle CNN features retain more details while high features has more semantic information. The previous features are used for precise localization and the later for preventing model drift. However, when it comes to obvious target deformation, neither of handcraft or CNN features can gain good performance. They can't handle cases where video frames are of large target appearance variation.

These days, Correlation Filters (CFs) are prevailingly used in visual tracking because of their effective computational ability. There is a large volume of published studies exploiting CFs [8,17,9]. Tracking algorithms utilizing CFs are discriminative methods other than generative methods. Discriminative trackers are considered as a classification problem. In the video frames, CFs calculate the correlation between the target model obtained in the last frame and the uncertain target samples in the present frame. CFs are used to distinguish the foreground from the background actually. In spite of advantages of correlation filters, they are not capable of dealing with the fast motion with a single one correlation filters. At the same time, some multiple correlation filter models [34,37,4] are combined by a linear way which also have limitations that the tracking algorithm are not robust to some of the challenges such as target deformation and fast motion. Moreover, many end-to-end deep learning-based trackers are also proposed but with the limit of complex computation[29].

To address the problems mentioned above, we propose an object tracking algorithm based on residual network and cascaded correlation filters in this paper. The main three contributions of our paper are described as follows:

1. We extract deep Resnet features from a pretrained residual network model, and choose features extracted by middle and high network layers to represent our target appearance model. Different features can express different information of the target, which are combined to represent the overall target appearance that contributes to the improved tracking.

2. We propose a cascaded way to combine these correlation filters. Correlation filters generated by color histogram, HOG, middle and high Resnet features respectively, are cascaded in the outline of our algorithm (see in Fig. 1). Features and CFs are fused properly and the tracking performance is improved.

3. We conduct extensive experiments on OTB2013 and OTB2015 benchmarks [30,31]. The experimental results show that our proposed tracker achieves good performance improvement compared with other state-of-the-art trackers.

2 Related Work

2.1 Tracking by CFs

With the successful application in object detection and recognition, correlation filters are applied to object tracking properly as well with the help of Fast Fourier Transform (FFT). Surprisingly, such kinds of trackers achieve remarkable performance due to efficient computational ability. The first formal discussion and analysis of CFs-based tracking, proposed by Bolme et al. emerged in 2010 [3]. It trained correlation filter by using Minimum Output Sum of Square Error (MOSSE) to obtain response map, then found the max response. Numerous studies have attempted to apply CFs later, including Circulant Structure of Tracking-by-Detection with Kernels (CSK) [15], Kernelized Correlation Filters (KCF) [14]. In 2015, Martin Danelljan et al. proposed Spatially Regularized Discriminative Correlation Filters (SRDCF) [9] in order to reduce the boundary effect caused by circulant shifts. It introduced spatial regularization term and used iterative Gauss-Seidel method to optimize. In 2018, Feng li et al. put forward Spatial-Temporal Regularized Correlation Filters (STRCF) [16]. It added temporal regularization term extensively based on SDRCF to gain time pertinence between frames with Alternating Direction Method of Multipliers (ADMM) optimal method. Nevertheless, these tracking methods use one single correlation filter to track target, which limits the precision and robustness during tracking process. In this paper, we utilize cascaded correlation filter to improve tracking accuracy.

2.2 Tracking by Handcraft Features

Handcraft features achieve really good performance in the early years of tracking and they attract more researchers to work on[13]. It has been conclusively shown that handcraft features can localize the precise target position. In 2010, Bolme et al. used original image content to gain the target model called MOSSE tracking method [3]. In 2012, Henriques et al. utilized gray feature to represent the target samples and tracking at a relatively high speed [15]. In 2014, Martin Danelljan et al. proposed a tracker based on Color Names (CN) space feature [10]. The tracker divided RGB color mode into eleven categories including black, blue, brown, gray, green, orange, pink, purple, white and yellow. Considered complex computation caused by eleven color channels, CN reduced the number of channels by Principle Component Analysis (PCA). In 2015, Henriques et al. exploited Histogram Oriented Gradient feature with multiple channels to represent the target samples and gained higher tracking speed [14]. Moreover, color histogram [23] is robust dealing with deformation while HOG feature is remarkable when it comes to motion blur and illumination changes. Bertinetto et al. fused color histogram and HOG effectively with training and testing method [1] in 2016. In this paper, we exploit deep features extracted from networks pretrained by image processing datasets.

2.3 Tracking by CNN Features

CNN features shows promising results for visual object tracking. In 2013, Naiyan Wang et al. combined deep learning and object tracking for the first time [28]. Even though the precision is low, it provided feasible method using offline training and online fine-tuning for tracking algorithm followed. In 2015, Chao Ma et al. proposed Hierarchical Convolutional Features (HCF) tracking method. It discovered the fact that features extracted by CNN layers differed from each other [21]. HCF used features from Conv3-4, Conv4-4 and Conv5-4 in VGG-19 to generate three types of correlation filters. Then each response map obtained from correlation filter was weighted and summed to gain the final response map. In 2016, Martin Danelljan et al. proposed Continuous Convolution Operator Tracker (CCOT) achieved the fusion of handcraft features and deep features [11] to complement each other and achieved high scores. In 2017, Martin Danelljan et al. speeded up CCOT by proposing Efficient Convolution Operators (ECO) [7]. ECO reduced the dimension of features and simplified the training dataset in order to reduce the time and space complexities and prevent the model drift. In this paper, we utilize cascaded correlation filters generated by handcraft and Resnet middle and high features to gain better competence. In 2016, SiamFC [2] is proposed by using a Siamese network to find the correlation between the target and samples. It is effective to exploit such kind of network model proved by the experimental results.

3 Our approach

Fig. 1 illustrates the framework of our whole tracking algorithm. When the target in the present frame is being tracked, the region of interest (ROI) is cropped from the input video frame. The size of the ROI is often twice or three times the size of the target bounding box predicted in the last tracking frame. Then, for one thing, the ROI image patch is ejected into a pretrained residual network model named Resnet-101[] to extract middle and high layers' features. For another, the ROI image patch is processed with the color histogram and HOG and then extract handcraft color histogram and HOG features. Handcraft features localize target precisely because they contain more spatial details while Resnet features are robust to the target appearance change because they retain more semantic information. For the purpose of generating response maps, features obtained are convolved by correlation filters. Then, the response maps are combined with each other by a cascaded method. This kind of method to combine the response maps is motivated by the cascaded networks in pose estimation. Cascaded network combines two similar network models. These two similar models have differences as well and cope with different situations. Therefore, we propose a cascaded correlation filters to combine different response maps which deal with various challenges respectively. Finally, from the combined response map result, we can find the maximum response value to obtain the final target position. The final response map will be mapped to the size of the original image patch as well as the position of the maximum response value. As a result, the mapped position of the maximum response value represents the final target position. The scale of the target will also be predicted at the same time. The process of the target scale estimation are not illustrated in the Fig. 1 for clarity.

In this section, we give a comprehensive review of our tracking algorithm in details. Our proposed tracking algorithm is divided into three parts: (i) Extract Resnet features to represent the target appearance model; (ii) Use discriminative correlation filter to find the correlation between target model and image patch samples; (iii) Exploit cascaded correlation filters to obtain final response map.

3.1 Resnet Feature

As the present convolutional neural network is limited by exploding or vanishing gradients, the residual network is trained to a deeper structure with the building residual blocks. Residual network is deeper and achieves high performance as well. Compared to ordinary convolutional network, Resnet-101 contains five convolutional layer stages including 101 layers which is deeper with less redundancy.

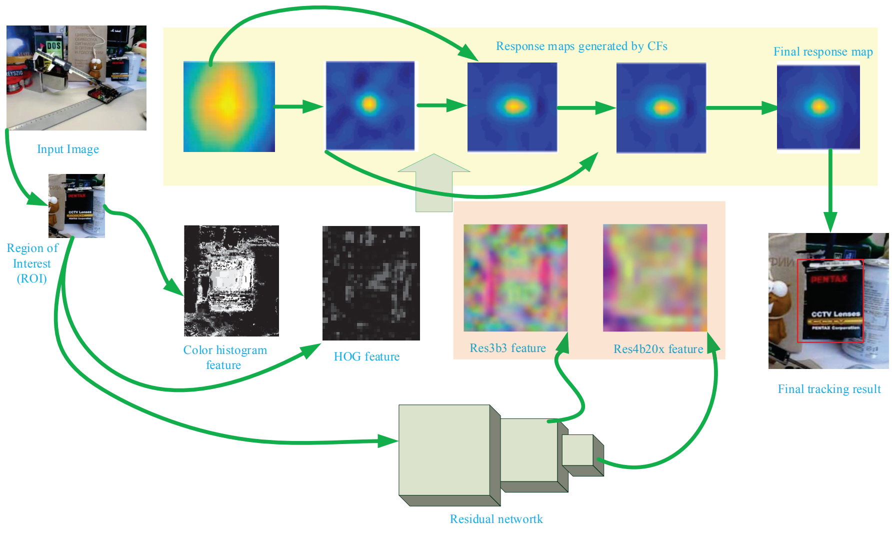

Fig. 1: Comprehensive framework of our tracking algorithm.

Fig. 2 represents a building residual block in the residual network. x denotes the input to the first layer of these layers. $F(x)$ denotes a network map before the summary. $H(x)$ denotes the desired output to these layers after summary. By introducing $H(x)$, it is easier to fit the residual value by equation in Fig. 2.

Resnet features is defined as feature extracted from residual network. Considering the features extracted by adjacent layers has less difference as well, we select the best feature output in each stage to reduce the redundancy and the interference caused by similar features in the tracking process. More specifically, we use res3b3 of size 28×28 in the middle stage and rest4b20x of size 14×14 in the high stage empirically. The layer res3b3 is in the third layer of Res101 and it is a branch layer of the third layer. The layer res4b20x is in the fourth layer of Res101 and it is also a branch layer of the fourth layer. The features extracted from res3b3 and res4b20x are rich in semantic information. The raw images are of size 224×224.

Fig. 3 illustrates the visualization of the middle Resnet features and high Resnet features from unsampled images (the 1st column in Fig. 3) in different categories. In Fig. 3, our tracking target is in the green bounding box. We find that features extracted from middle Resnet layers (the 2nd column in Fig. 3) has richer detailed information while features extracted from high Resnet (the 3rd column in Fig. 3) retain more semantic information. What's more, these features are more able to distinguish targets from background.

3.2 Discriminative Correlation Filters

As is known to us all, CFs-based trackers have raised much concern in recent years. They have been widely used in the field of object tracking as a result of cyclic samples

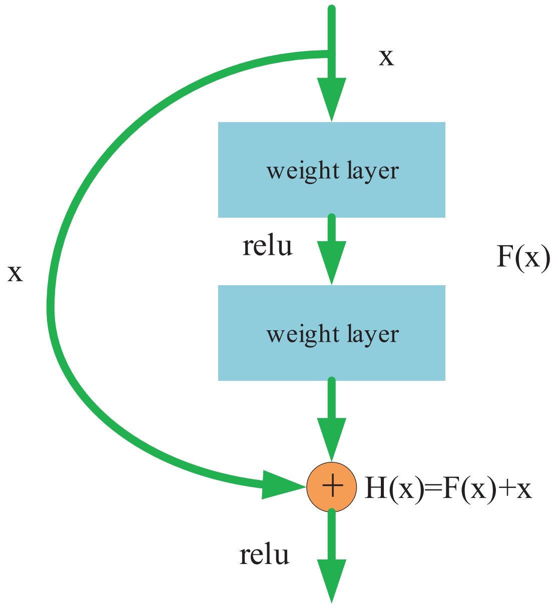

Fig. 2: The basic unit of deep residual network.

and FFT method, which are used to accelerate the tracking process collaboratively. Tracking methods based on discriminative correlation filters, trains a correlation filter by densely generated samples in the previous frame. In the next frame, the trained correlation filter is utilized to calculate the target position response maps. The maximum response represents the optimal target center position. Finally, the correlation filter is updated by the new predicted target position.

Firstly, we denote x as a feature of a single channel in a feature vector of size $M \times N \times D$ extracted from the original image patch. Moreover, M refers to the width, N refers to the height and D refers to the channel of the feature vector. $x_{m,n}$ denotes our image cyclic shifted sample, $m \in \{0, 1, ..., M-1\}$, $n \in \{0, 1, ..., N-1\}$. Gaussian function label is defined as a 2D Gaussian distributions:

$$y(m,n) = e^{\frac{(m-M/2)^2+(n-N/2)^2}{2\sigma^2}} \tag{1}$$

where σ denotes the kernel width.

Each correlation filter trained by shifted samples follows the following minimum cost function:

$$\arg\min_{w} \sum_{m,n} \|w \cdot x_{m,n} - y(m,n)\|_2^2 + \lambda \|w\|_2^2 \tag{2}$$

In equation (1), w is the correlation filter that needs to be learned and λ denotes a regularization parameter. λ belongs to $[0, +\infty]$ To simplify the calculating process in practice, the FFT is introduced to transform the calculation from time domain to

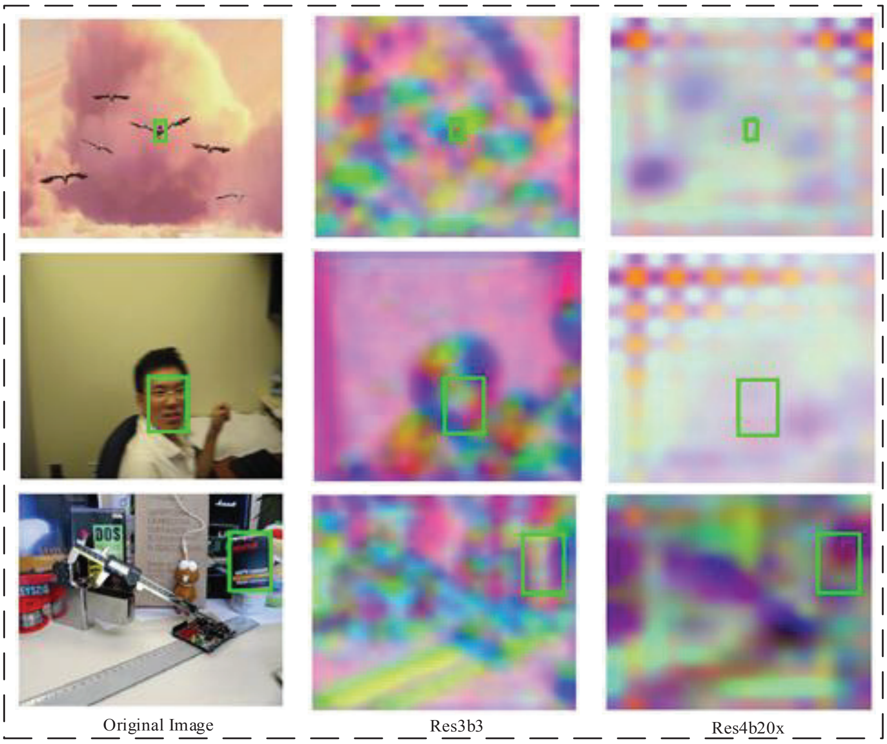

Fig. 3: Visualization of the middle Resnet features and high Resnet features from difference sequences' frame. Video sequence in the first row is *bird2*, the second row is *blurface*, and the third row is *box*.

frequency domain. Thus, the equation (2) can also be written on each channel as follows:

$$W^d = \frac{Y \odot \bar{X}^d}{\sum_{i=1}^{D} X^i \odot \bar{X}^i + \lambda} \tag{3}$$

where X, Y, W are the Fourier Transformation form of $x_{m,n}$, $y(m,n)$ and w respectively. In equation (3), the bar means complex conjunction. $\odot$ means the element-wise product. d means the d-th channels. D denotes the number of channels, $d \in \{1,..,D\}$.

After the correlation filter w is trained in the $t-1$ frame, we localize the position of the target in the t frame by w. Given the image patch in the t frame, we extract features for each category mentioned above. Each feature vector is denoted by p of size $M \times N \times D$. Then the response map R of size $M \times N$ generated by correlation filter for each feature is computed as:

$$R = \mathscr{F}^{-1}\left(\sum_{d=1}^{D} W^d \odot \bar{P}^d\right) \tag{4}$$

where $\mathscr{F}^{-1}$is an inverse FFT operator. The maximum value of R is our final tracking position.

The correlation filter models are updated per frame to avoid model drift because the appearance variation of target as the video frame goes forward. In the process of object tracking, a well-optimized correlation filter is trained with lots of training samples over all tracking results by the minimum output error. Nevertheless, it is found that the computation of the optimal correlation filter is rather time-consuming because each deep feature from layers of residual network in each position estimated, has D channels (e.g., Res3b3 and Res4b20x in the Resnet-101 has 512 and 1024 channels, that is $D = 512$ and $D = 1024$ respectively), which are too large for computation. To achieve the robustness of our optimal correlation filters estimation, we need to update the W^d in equation (4) of the$l$layer of the t-th frame. The numerator A^d and the denominator $B^d$are updated separately by employing a moving average as follows:

$$A_t^d = (1-\eta)A_{t-1}^d + \eta Y \odot \bar{X}_t^d \tag{5a}$$

$$B_t^d = (1-\eta)B_{t-1}^d + \eta \sum_{i=1}^{D} X_t^i \odot \bar{X}_t^i \tag{5b}$$

$$W_t^d(l) = \frac{A_t^d}{B_t^d + \lambda} \tag{5c}$$

where l denotes the layer of the residual network, t denotes the frame index of the video sequences, η is the learning rate. What's more, we update the correlation filters per frame to achieve the robust approximation.

3.3 Cascaded Correlation filters

There are some combinations of multiple correlation filters like linear combination. These methods can't achieve high performance because the correlation filters are not fully used by linear combination only. Therefore, we propose cascaded correlation filters in this paper to combine the correlation filters more precisely to improve the tracking performance. Cascaded discriminative correlation filters are clearly shown in Fig. 4. Each correlation filter generates one kind of response map. Thus, we represent the correlation filters by response maps generated. Based on the color histogram, HOG, middle and high layers in the residual network, we combine these features extracted by cascading them. Different from linear combination, the cascaded correlation filters are able to fully utilize the information obtained from different features.

There are four kinds of response maps and a final response map, but the former response maps are combined with all following response maps by a cascaded method. From top to down, first response map is generated by color histogram, second is generated by HOG, third is generated by Res3b and forth is generated by Res4b20x. Green arrow represents the combination of different response maps. The first is combined with both second and third response map. The second combines with third and the forth. The final response map is our final result to search for the target position.

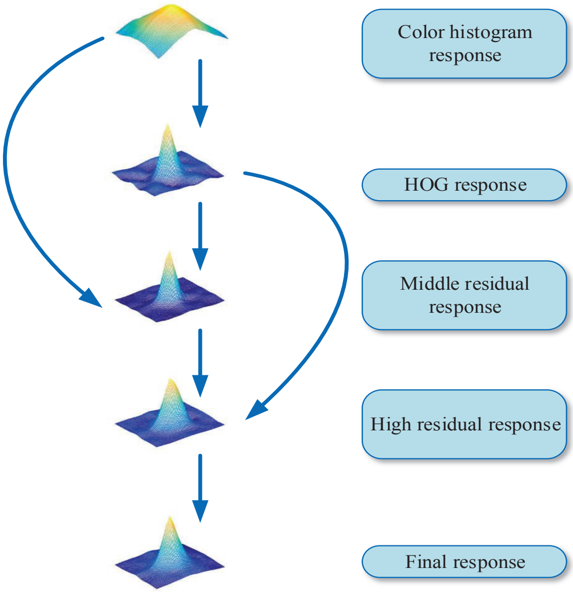

Fig. 4: Cascaded discriminative correlation filter represented by four different kinds of response maps. The final response map is obtained by combining the previous response maps in a cascaded way.

According to Fig. 4, it can be found that each response map has different information. As a result, we empirically combines these complementary information to diminish the model drift of our tracking algorithm. With the help of cascaded correlation filters, our proposed tracker has achieved favorable tracking performance.

4 Experiment

In this section, we state the detailed experiments conducted to evaluate our tracker with Resnet features and cascaded correlation filters. We use the combined handcraft features and Resnet features to localize the target collaboratively. Firstly, we discuss the details of our experiments. Then, we verify the effectiveness of our each contribution. We also make comparison between our tracker and some other state-of-the-art trackers on OTB2013 and OTB2015 datasets comprehensively. We compare our tracker with others in 11 different attributes annotated on OTB benchmark and these tracker are also compared in quality and quantity respectively. The experimental results prove that our tracker is of high precision and success rate. As a result, our tracker shows great performance in the field of object tracking.

4.1 Implementing details

The image patch passes through Res3b3 and Res4b20x in Resnet-101 for target appearance with network forward propagation and residual operation. We use the output of Res3b3 and Res4b20x for target representation. Besides Res3b3, Res4b20x features, we select two kinds of handcraft features, specifically, color histogram, HOG, to represent our target. HOG has 31 channels, Res3b3 is of size 28×28 and Res4b20x is of size 14×14. Handcraft and Resnet features have different specialties which cope with different problems. Handcraft features localize the target precisely while Resnet features obtains richer semantic information. These complementary features generate response maps through cascaded correlation filters. Combined features predict the status of the target quite well.

Our tracker is implemented on a computer with an Intel I7-6700 K 4.0 GHz CPU, 16GB RAM, and a GeForce GTX980Ti GPU model. And we use one single GPU and Matconvnet Toolbox to compute Resnet Features only. Our proposed tracker is implemented in MATLAB R2018b. Some parameters in the equations should be set to a fixed value in advance. Specifically, σ in equation (1) is set to 1/16, the regularization parameter λ in equation (2) is set to 0.001, the learning rate η in equation (5) is set to 0.01.

4.2 Evaluation metrics

OTB is consisted of two datasets: OTB2013 proposed in 2013 and OTB2015 proposed in 2015, which is an extension on the basis of OTB2013. OTB2013 contains 50 sequences of different categories which has 51 various targets, while OTB2015 contains 98 video sequences which has 100 targets. OTB has 11 challenging attributes, namely, illumination variation (IV), motion blur (MB), deformation (DEF), fast motion (FM), scale variation (SV), occlusion (OCC), background clutter (BC), in-plane rotation (IPR), out-of-plane rotation (OPR), low resolution (LR), out-of-view (OV).

There are two metrics in OTB benchmark, which are precision plots and success rate plots, corresponding to Center Location Error (CLE) and overlap respectively. CLE is defined as the Euclidean distance between the predicted central coordinate position and the labelled groundtruth central coordinate position. Furthermore, overlap refers to the Intersection over Union (IoU) between predicted target bounding box and labelled groundtruth target bounding box. IoU is greater than zero and less than one according to mathematical theory. We often use the ratio of the number of frames whose tracking IoU is greater than IoU threshold to the number of all frames as the overlap success rate. Meanwhile, we use the ratio of the number of frames of which Euclidean Distance less than CLE threshold to the number of all frames as the distance precision rate. Success plot uses overlap success rate at 0.5 as the specific threshold to evaluate trackers. Precision plot uses distance precision rate at 20 pixels to evaluate trackers. OPE is mentioned below which means One-pass Evaluation proposed in OTB2013.

4.3 Ablation studies

Effectiveness of Resnet Features: Ordinary convolutional network uses the top-down feature pyramid network units to extract features. CF2 captures the target with the fusion of hierarchical features. However, the convolutional network that CF2 uses is not deep enough. Therefore, we introduce the residual network into our tracking framework. In order to analysis the effectiveness of different layers of the deep residual network, we do the experiment to approve that. The performance of different layers of Resnet-101 in several situations provided in OTB2013 based on the framework in CF2 is illustrated in Fig. 5 and Fig. 6. From the tracking results in the figures, features extracted from middle and high level layers in the residual network handle some specific challenges. The information from middle layers (like Res3b3) can deal with the challenge of deformation properly. At the same time, the information from high layers (like Res4b20x) contains more semantic information which is able to cope with the challenge of occlusion. Comprehensively, the performance of using Res3b3 and Res4b20x to extract features separately are both high as well.

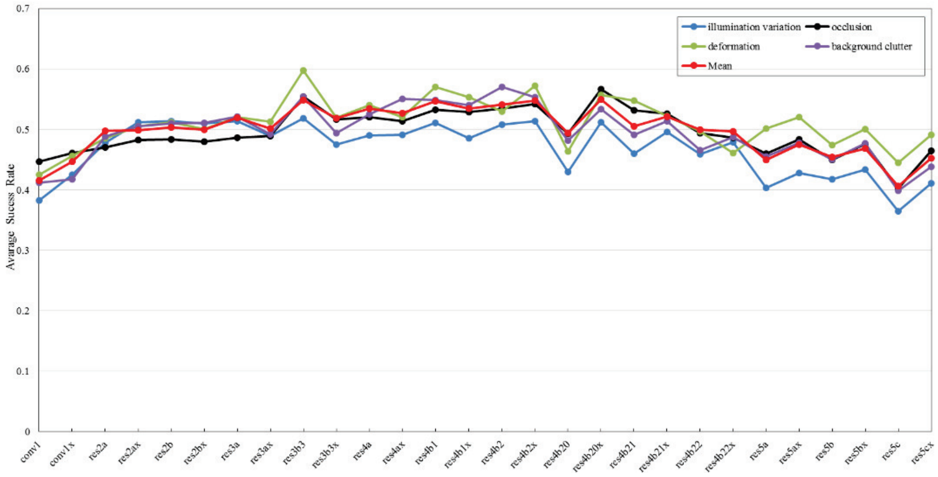

Fig. 5: The average success rate of different layers in four challenging situations: illumination variation, deformation, occlusion and background clutter. The red line indicts the overall performance over 50 videos with 51 targets in OTB2013. The x-axis represents the layers of Resnet-101 and the y-axis represents the average success rate to evaluate the performance.

We make comparison between our baseline and baseline with features extract from Res3b3, Res4b20x on the benchmark OTB2013, which contains 50 challenging tracking videos. We use the correlation filter-based tracking algorithm named Staple [1] as our baseline. Fig. 7 shows that the *baseline* combined with features extracted from Res3b named *baseline+Res3b* increases by 1.38% and 1.6% over the precision and success plots. The baseline with features extracted from Res4b20x named *base-*

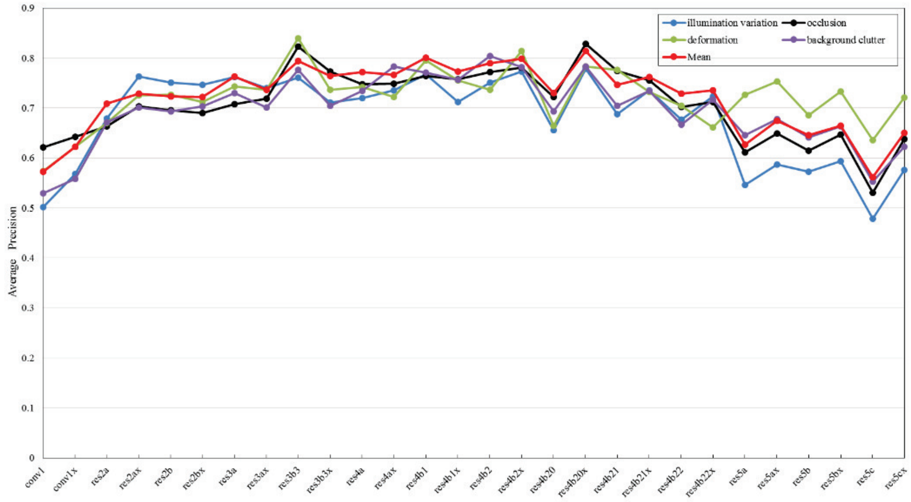

Fig. 6: The average precision of different layers in four challenging situations: illumination variation, deformation, occlusion and background clutter. The red line indicts the overall performance over 51 videos in OTB2013. The x-axis represents the layers of Resnet-101 and the y-axis represents the average precision to evaluate the performance.

line+Res4b20x increases by 3.4% and 1.84% over the precision and success plots. From the tracking results, we can draw the conclusion that features extracted from Res3b3 and Res4b20x shows a promising performance in the object tracking. The effectiveness of our Resnet features are demonstrated at the same time.

Effectiveness of Cascaded Correlation filters: Based on the features extracted from the residual network, we proposed a cascaded correlation filters to obtain the final correlation filter as well get the final response map. To demonstrate the effectiveness of the proposed cascaded correlation filter further, we conduct the experiments on the basis of Baseline on OTB2013 and OTB2015. At first, we use linear method to combine the correlation filters. It means that we simply add the response maps. From the experimental results shown in the Table 1, we can see that after incorporating the linear combination entitled *Baseline + Linear Combination* and cascaded combination entitled *Baseline + Cascaded Combination* of correlation filters into our tracker, the tracking performance of baseline with cascaded combination are much better than the baseline and the baseline with linear combination both in precision and success rate. As a result, the performance of our tracker only based on the features extracted from residual network is still boosted further by cascaded correlation filters.

4.4 Overall performance

In this section, we state the experiments conducted to evaluate our proposed tracker over benchmark OTB2013 and OTB2015. The proposed tracking algorithm is com-

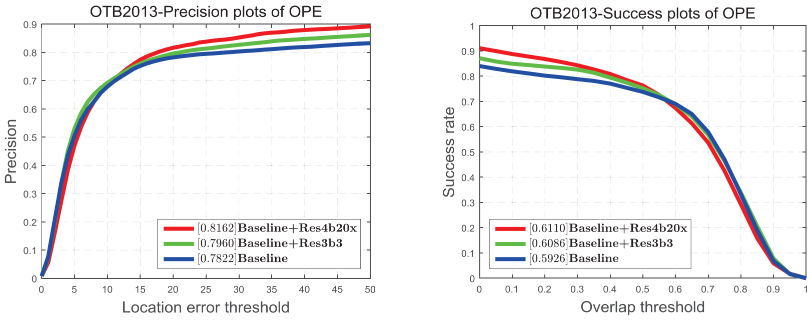

Fig. 7: The effectiveness study of our proposed Resnet features named Res3b3 and Res4b20x on OTB2013. The comparison between correlation filter-based tracking algorithm entitled *Baseline*, Baseline combined with Res3b3 entitled *Baseline+Res3b3* and Baseline combined Res4b20x entitled *Baseline+Res4b20x*. The validation of the residual network is proved by incorporating Res3b3 and Res4b20x into basic correlation filter.

Table 1: Comparison between different methods of combining the correlation filters. *Baseline* refers to the Staple, *linear combination* refers to add the response maps simply and *cascaded combination* refers to combining the correlation filters in a cascaded way. We conduct the experiments over OTB2013 and OTB2015 by precision and success rate. The highest results are marked in bold. The excellent performance of cascaded correlation filters are validated by comparative experiments.

Combination of correlation filters	OTB-2013		OTB-2015	
	Precision	Success rate	Precision	Success rate
Baseline	0.7822	0.5926	0.7840	0.5789
Baseline+Linear Combination	0.7373	0.5628	0.7307	0.5408
Baseline+Cascaded Combination	**0.8354**	**0.6179**	**0.8121**	**0.5938**

pared with 7 other state-of-the-art trackers, namely SiamFC [2], KCF [14], DSST [8], SAMF [17], Staple [1], MEEM [37]. SiamFC is an end-to-end tracking algorithm based on deep learning. KCF, DSST, SAMF and Staple are tracking algorithms using handcraft features based on the correlation filters. What's more, MEEM is a tracking method with multiple correlation filter models combined for object tracking.

Quantitative Evaluation: The OPE results of our proposed tracker and other 6 trackers are compared to demonstrate the robustness of our tracker on the OTB2013 and OTB2015 respectively. OTB2015 is more challenging than OTB2013 because OTB2015 adds some large and difficult videos. Fig. 8 and Fig. 9 presents that our tracker performs favorably against other 6 state-of-the-art trackers according to the metrics presented in the OTB benchmark. In particular, our tracker performs excellently with the precision of 83.54% and success rate of 61.79% on OTB2013. In addition, our tracker is of precision at 81.21% and success rate at 59.38%, which is the highest results over all compared tracking algorithms.

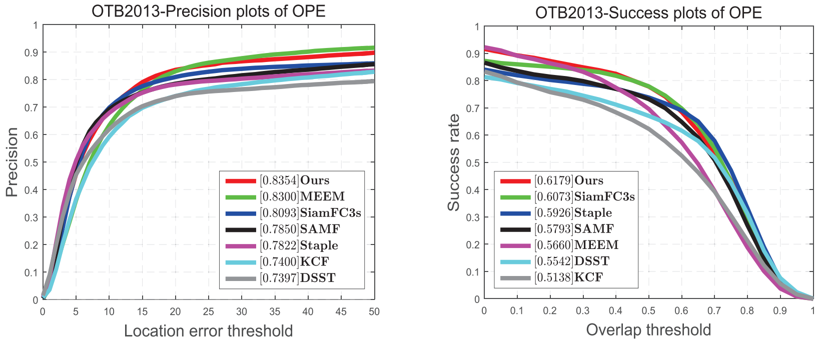

Fig. 8: Precision plot and success plot of OPE on OTB2013. The precision plots computed by center location error are on the left. The success plots computed by overlap are on the right. Our proposed tracker performs favorably against other state-of-the-art trackers.

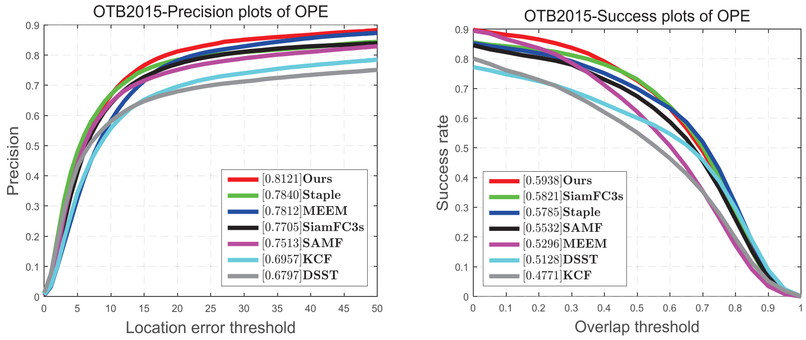

Fig. 9: Precision plots and success plots of OPE on OTB2015. The precision plots computed by center location error are on the left. The success plots computed by overlap are on the right. Our proposed tracker performs favorably against other state-of-the-art trackers.

Attribute-based Evaluation: Fig. 10 and Fig. 11 shows the precision plots and success plots of 11 attributes-based comparison tracking results on OTB2013 dataset. Our proposed tracker performs better in most of these 11 attributes. For example, when it comes to deformation, our tracking result in precision plot is up to 90.9% while in success plot is about 59.5%. It proves that our tracker can address the situation where the tracking target is under the challenge of deformation. Moreover, our proposed tracker also performs favorably when facing the background clutter, illumination variation, rotation and low resolution.

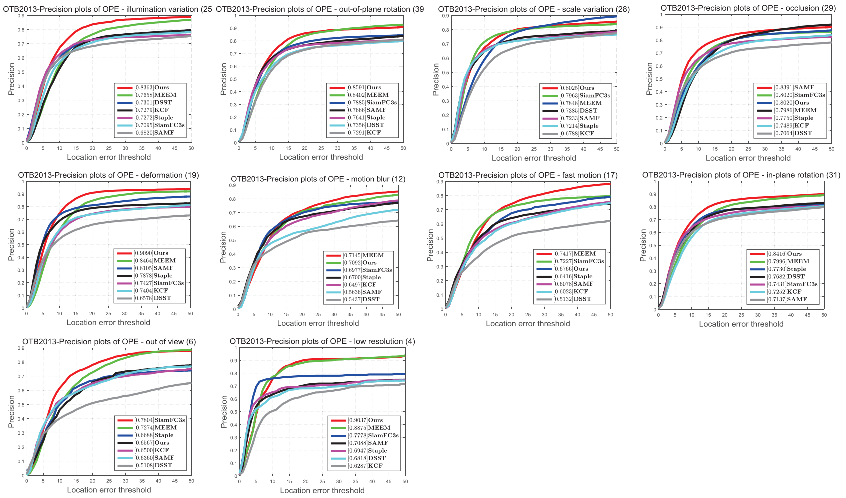

Fig. 10: Precision Plot of 11 Attribute-based comparison of our tracker and other existing trackers on OTB2013.

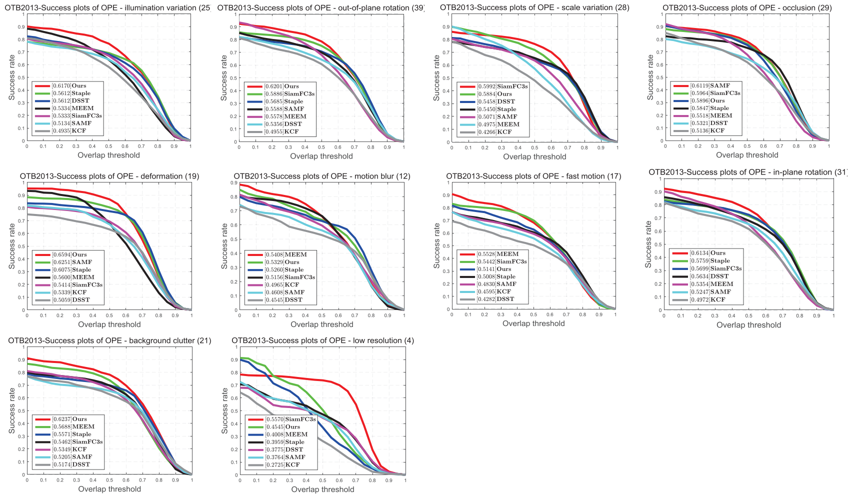

Fig. 11: Success plots of 11 Attribute-based comparison of our tracker and other existing trackers on OTB2013.

5 Conclusion

In this paper, we propose a tracking method based on the Resnet features and cascaded correlation filters to address the problem that the target appearance model is weak and the tracking methods have bad performance under the challenge of obvious target deformation. Our tracking algorithm can solve the problems mentioned above. The target model represented by Resnet features is reinforced apparently. What's more, we exploit cascaded correlation filters trained by specific features to track the target. Complementary response maps are obtained from these correlation filters. And various response maps are fused by cascading method to gain the final response map, where the maximum value of the response map is our final tracking result. Extensive experiments are carried out in a large-scale benchmark named OTB2013 and OTB2015. The experimental results show that our proposed tracker performs favorably against other state-of-the-art tracking algorithm. In the future, we would improve our method and see if Resnet feature is compatible to other existing trackers. Moreover, it is also feasible to train a network model similar to residual network and extract stronger features.

Acknowledgements This research was funded in part by the National Natural Science Foundation of China under Grant Nos. 61972056, 61772454 and 61811540410, the ”Double First-class” International Cooperation and Development Scientific Research Project of Changsha University of Science and Technology under Grant No. 2019IC34, the Postgraduate Scientific Research Innovation Fund of Hunan Province under Grant Nos. CX20190696 and CX20190695.

References

1. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S.: Staple: Complementary learners for real-time tracking. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 1401–1409 (2016). DOI 10.1109/CVPR.2016.156. URL https://doi.org/10.1109/CVPR.2016.156
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking. In: Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II, pp. 850–865 (2016). DOI 10.1007/978-3-319-48881-3_56. URL https://doi.org/10.1007/978-3-319-48881-3_56
3. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pp. 2544–2550 (2010). DOI 10.1109/CVPR.2010.5539960. URL https://doi.org/10.1109/CVPR.2010.5539960
4. Chen, Y., Wang, J., Liu, S., Chen, X., Xiong, J., Xie, J., Yang, K.: Multiscale fast correlation filtering tracking algorithm based on a feature fusion model. Concurrency and Computation: Practice and Experience p. e5533 (2019). DOI 10.1002/cpe.5533. URL https://doi.org/10.1002/cpe.5533
5. Chen, Y., Wang, J., Xia, R., Zhang, Q., Cao, Z., Yang, K.: The visual object tracking algorithm research based on adaptive combination kernel. J. Ambient Intelligence and Humanized Computing **10**(12), 4855–4867 (2019). DOI 10.1007/s12652-018-01171-4. URL https://doi.org/10.1007/s12652-018-01171-4
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, pp. 886–893 (2005). DOI 10.1109/CVPR.2005.177. URL https://doi.org/10.1109/CVPR.2005.177
7. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ECO: efficient convolution operators for tracking. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 6931–6939 (2017). DOI 10.1109/CVPR.2017.733. URL https://doi.org/10.1109/CVPR.2017.733
8. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014 (2014). URL http://www.bmva.org/bmvc/2014/papers/paper038/index.html
9. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015,

Santiago, Chile, December 7-13, 2015, pp. 4310–4318 (2015). DOI 10.1109/ICCV.2015.490. URL https://doi.org/10.1109/ICCV.2015.490

10. Danelljan, M., Khan, F.S., Felsberg, M., van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pp. 1090–1097 (2014). DOI 10.1109/CVPR.2014.143. URL https://doi.org/10.1109/CVPR.2014.143
11. Danelljan, M., Robinson, A., Khan, F.S., Felsberg, M.: Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V, pp. 472–488 (2016). DOI 10.1007/978-3-319-46454-1_29. URL https://doi.org/10.1007/978-3-319-46454-1_29
12. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 5374–5383 (2019). DOI 10.1109/CVPR.2019.00552. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Fan_LaSOT_A_High-Quality_Benchmark_for_Large-Scale_Single_Object_Tracking_CVPR_2019_paper.html
13. Gao, Z., Xia, S., Zhang, Y., Yao, R., Zhao, J., Niu, Q., Jiang, H.: Real-time visual tracking with compact shape and color feature. Comput. Mater. Continua **55**(3), 509–521 (2018)
14. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2015). DOI 10.1109/TPAMI.2014.2345390. URL https://doi.org/10.1109/TPAMI.2014.2345390
15. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.P.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV, pp. 702–715 (2012). DOI 10.1007/978-3-642-33765-9_50. URL https://doi.org/10.1007/978-3-642-33765-9_50
16. Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.: Learning spatial-temporal regularized correlation filters for visual tracking. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 4904–4913 (2018). DOI 10.1109/CVPR.2018.00515. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Li_Learning_Spatial-Temporal_Regularized_CVPR_2018_paper.html
17. Li, Y., Zhu, J.: A scale adaptive kernel correlation filter tracker with feature integration. In: Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II, pp. 254–265 (2014). DOI 10.1007/978-3-319-16181-5_18. URL https://doi.org/10.1007/978-3-319-16181-5_18
18. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE Trans. Image Processing **24**(12), 5630–5644 (2015). DOI 10.1109/TIP.2015.2482905. URL https://doi.org/10.1109/TIP.2015.2482905
19. Liu, F., Guo, Y., Cai, Z., Xiao, N., Zhao, Z.: Edge-enabled disaster rescue: A case study of searching for missing people. ACM TIST **10**(6), 63:1–63:21 (2019). DOI 10.1145/3331146. URL https://doi.org/10.1145/3331146
20. Liu, W., Liu, Z., Wang, L., Li, B., Jing, N.: Human movement detection and gait periodicity analysis via channel state information. Computer Systems Science and Engineering **33**(2) (2018)
21. Ma, C., Huang, J., Yang, X., Yang, M.: Hierarchical convolutional features for visual tracking. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 3074–3082 (2015). DOI 10.1109/ICCV.2015.352. URL https://doi.org/10.1109/ICCV.2015.352
22. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, pp. 445–461 (2016). DOI 10.1007/978-3-319-46448-0_27. URL https://doi.org/10.1007/978-3-319-46448-0_27
23. Possegger, H., Mauthner, T., Bischof, H.: In defense of color-based model-free tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 2113–2120 (2015). DOI 10.1109/CVPR.2015.7298823. URL https://doi.org/10.1109/CVPR.2015.7298823
24. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA, pp. 511–518 (2001). DOI 10.1109/CVPR.2001.990517. URL https://doi.org/10.1109/CVPR.2001.990517

25. Wang, J., Gao, Y., Liu, W., Sangaiah, A.K., Kim, H.: An intelligent data gathering schema with data fusion supported for mobile sink in wireless sensor networks. IJDSN **15**(3) (2019). DOI 10.1177/1550147719839581. URL `https://doi.org/10.1177/1550147719839581`
26. Wang, J., Gao, Y., Liu, W., Wu, W., Lim, S.J.: An asynchronous clustering and mobile data gathering schema based on timer mechanism in wireless sensor networks. Comput. Mater. Contin **58**, 711–725 (2019). URL `https://doi.org/10.32604/cmc.2019.05450`
27. Wang, J., Ju, C., Gao, Y., Sangaiah, A.K., Kim, G.j.: A pso based energy efficient coverage control algorithm for wireless sensor networks. Comput. Mater. Contin **56**(3), 433–446 (2018)
28. Wang, N., Yeung, D.: Learning a deep compact image representation for visual tracking. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pp. 809–817 (2013). URL `http://papers.nips.cc/paper/5192-learning-a-deep-compact-image-representation-for-visual-tracking`
29. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.S.: Fast online object tracking and segmentation: A unifying approach. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 1328–1338 (2019). DOI 10.1109/CVPR.2019.00142. URL `http://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Fast_Online_Object_Tracking_and_Segmentation_A_Unifying_Approach_CVPR_2019_paper.html`
30. Wu, Y., Lim, J., Yang, M.: Online object tracking: A benchmark. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, pp. 2411–2418 (2013). DOI 10.1109/CVPR.2013.312. URL `https://doi.org/10.1109/CVPR.2013.312`
31. Wu, Y., Lim, J., Yang, M.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015). DOI 10.1109/TPAMI.2014.2388226. URL `https://doi.org/10.1109/TPAMI.2014.2388226`
32. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38**(4), 13 (2006). DOI 10.1145/1177352.1177355. URL `https://doi.org/10.1145/1177352.1177355`
33. Zhang, D., Yang, G., Li, F., Wang, J., Sangaiah, A.K.: Detecting seam carved images using uniform local binary patterns. Multimedia Tools and Applications pp. 1–16 (2018). DOI 10.1007/s11042-018-6470-y. URL `https://doi.org/10.1007/s11042-018-6470-y`
34. Zhang, J., Jin, X., Sun, J., Wang, J., Li, K.: Dual model learning combined with multiple feature selection for accurate visual tracking. IEEE Access **7**, 43956–43969 (2019). DOI 10.1109/ACCESS.2019.2908668. URL `https://doi.org/10.1109/ACCESS.2019.2908668`
35. Zhang, J., Jin, X., Sun, J., Wang, J., Sangaiah, A.K.: Spatial and semantic convolutional features for robust visual object tracking. Multimedia Tools and Applications pp. 1–21 (2018). DOI 10.1007/s11042-018-6562-8. URL `https://doi.org/10.1007/s11042-018-6562-8`
36. Zhang, J., Lu, C., Li, X., Kim, H.J., Wang, J.: A full convolutional network based on densenet for remote sensing scene classification. Math. Biosci. Eng **16**(5), 3345–3367 (2019)
37. Zhang, J., Ma, S., Sclaroff, S.: MEEM: robust tracking via multiple experts using entropy minimization. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI, pp. 188–203 (2014). DOI 10.1007/978-3-319-10599-4_13. URL `https://doi.org/10.1007/978-3-319-10599-4_13`
38. Zhang, J., Wang, W., Lu, C., Wang, J., Sangaiah, A.K.: Lightweight deep network for traffic sign classification. Annals of Telecommunications pp. 1–11 (2019). DOI 10.1007/s12243-019-00731-9. URL `https://doi.org/10.1007/s12243-019-00731-9`
39. Zhang, J., Wu, Y., Feng, W., Wang, J.: Spatially attentive visual tracking using multi-model adaptive response fusion. IEEE Access **7**, 83873–83887 (2019). DOI 10.1109/ACCESS.2019.2924944. URL `https://doi.org/10.1109/ACCESS.2019.2924944`
40. Zhou, Z., Qin, J., Xiang, X., Tan, Y., Liu, Q., Xiong, N.N.: News text topic clustering optimized method based on tf-idf algorithm on spark. Computers, Materials & Continua **62**(1), 217–231 (2020)