IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 6, 2019, accepted March 25, 2019, date of publication April 1, 2019, date of current version April 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908668

Dual Model Learning Combined With Multiple Feature Selection for Accurate Visual Tracking

JIANMING ZHANG[1,2], (Member, IEEE), XIAOKANG JIN[1,2], JUAN SUN[1,2], JIN WANG[1,2], (Senior Member, IEEE), AND KEQIN LI[3], (Fellow, IEEE)

[1]Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114, China
[2]School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
[3]Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

Corresponding author: Jin Wang (jinwang@csust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772454, Grant 61811540410, and Grant 61811530332, in part by the Scientific Research Fund of Hunan Provincial Education Department under Grant 16A008, in part by the Postgraduate Scientific Research Innovation Fund of Hunan Province under Grant CX2018B565, in part by the Undergraduate Inquiry Learning and Innovative Experimental Fund of the Changsha University of Science and Technology (CSUST) under Grant 2018-6-119, and in part by the Postgraduate Training Innovation Base Construction Project of Hunan Province under Grant 2017-451-30.

ABSTRACT Over these years, object tracking algorithms combined with correlation filters and convolutional features have achieved excellent performance in accuracy and real-time speed. However, tracking failures in some challenging sequences are caused by the insensitivity of deeper convolutional features to target appearance changes and the unreasonable updating of correlation filters. In this paper, we propose dual model learning combined with multiple feature selection for accurate visual tracking. First, we fuse the handcrafted features with the multi-layer features extracted from the convolutional neural network to construct a correlation filter learning model, which can precisely localize the target. Second, we propose an index named hierarchical peak to sidelobe ratio (HPSR). The fluctuation of HPSR determines the activation of an online classifier learning model to redetect the target. Finally, the target locations predicted by the dual learning models mentioned above are combined to obtain the final target position. With the help of dual learning models, the accuracy and performance of tracking have been greatly improved. The results on the OTB-2013 and OTB-2015 datasets show that the proposed algorithm achieves the highest success rate and precision compared with the 12 state-of-the-art tracking algorithms. The proposed method is better adaptive to various challenges in visual object tracking.

INDEX TERMS Convolutional neural network, correlation filter, learning models, multiple feature selection, object tracking.

I. INTRODUCTION

A. BACKGROUND

Visual tracking is an important and fundamental problem in the field of computer vision. Basically, it involves constructing a model by information of video and predicting the posture and trajectory of object according to the correlation between temporal and spatial context. Despite wide use in video surveillance, human-computer interaction, driverless vehicle and so on, object tracking is still challenging task due to factors such as illumination variation, scale variation, partial and full occlusion, fast motion, background clutter, in-plane and out-of-plane rotation.

With significant progress in object tracking recently, many competitive algorithms have been proposed, which are mainly divided into generative [1]–[3] and discriminative [4]–[6] methods according to different appearance models. The generative algorithms model the foreground, search the candidate regions by the minimum reconstruction error to find the best matching position in the current frame, then update the target model by online learning mechanism. Discriminative algorithms transform object tracking into a binary classification problem. By collecting a set of positive and negative samples in each frame, a discriminative classifier is trained to maximize the differences between the target

The associate editor coordinating the review of this manuscript and approving it for publication was Huazhu Fu.

and background. The performances of discriminative tracking algorithms primarily depend on feature extraction methods, designed classifiers and online updating mechanism of the classifiers. In order to overcome the problem that target appearance changes as time goes on, it is essential to use appropriate feature descriptors such as color histogram, Haar-like, SURF, HOG, subspace representation, super-pixel and even multiple features integration to represent the target. This work aims to improve the tracking accuracy by multiple level features selection and robust combination mechanism of dual learning models.

Deep convolutional neural network (CNN) have shown impressive performance in many tasks such as pattern classification [7], object detection [8] and region of interest detection [9] due to their success on automatic feature extraction via multi-layer nonlinear transformations. A CNN consists of several convolutional layers, pooling operation and softmax. Deep features extracted from convolutional layers have strong discriminative ability and preserve both spatial and structured information. There is an increasing amount of work on combination of deep convolutional features and correlation filters to predict the target location in object tracking. Convolutional features from deeper layers contain richer semantic information while features from earlier layers provide spatial resolution and edge details which play a vital role in localizing target accurately.

B. MOTIVATION

Among discriminative tracking algorithms, discriminative correlation filters (DCFs) based approaches [10]–[12] have gradually become popular and achieved outstanding results in object tracking benchmarks (OTB) [13], [14] as well. In these algorithms, the filters are trained to obtain target classification score and discrete Fourier transformation (DFT) is used for all cyclically shifted samples to do efficient calculation. DFT guarantees the real-time tracking performance. Therefore, many algorithms based on the combination of CNN and DCF are proposed in recent years [15], [16]. These algorithms rely on the powerful representation of convolutional features to obtain excellent tracking results. At the same time, they do not update the deep model online, which greatly improves the real-time performance of algorithms.

However, there are still some limitations to integrate CNN with DCF framework: 1) In the multi-layer features fusion, it is reasonable that deeper convolutional features are given larger weight because they have richer semantic information than earlier layers. However, the tracking process is easily disturbed by various factors, which misguides semantic information representation and enlarges the errors by online updating, resulting in transient drift. Therefore, the methods of multi-layer features fusion do not fully explore the effective relationship between features. Moreover, the multi-layer filters use weighted features to fuse do not effectively utilize response map information, which causes waste of information. 2) Correlation filters are difficult to adapt to severe occlusion or fast motion in the target movement, so the tracker will bring interference information into the continuous updating of correlation filters, resulting in the accumulation of errors, tracking drift or failure. In this case, a single learning model cannot effectively track target and a new intervention mechanism need to be involved.

C. MAIN CONTRIBUTIONS

In view of the above two points, we propose a dual learning model and multiple feature selection to implement accurate visual tracking. The main contributions are summarized as follows:

(1) A correlation filter learning model based on fusion of multi-level features is proposed. We exploit low, middle, high features to construct six-layer correlation filter banks and obtain target position in a recursive layer-by-layer method. The low and high features are selected as feature descriptors in the end.

(2) The index of Hierarchical Peak to Sidelobe Ratio (HPSR) is presented. Response maps are obtained via different correlation filters. We calculate the PSR of each response map and weigh it to obtain a confidence index for the current position.

(3) A classifier learning model for online redetection is proposed. The discriminative classifier is trained by collecting positive and negative samples around the target. It can detect the target again online and get a new position. The final target position is obtained by dual models.

We compare the proposed algorithm with the state-of-the-art trackers on large benchmark datasets OTB-2013 [13] and OTB-2015 [14]. Experiments show that the proposed method can effectively improve the tracking accuracy and success rate and can better adapt to challenges such as fast motion, occlusion and illumination.

II. RELATED WORKS

In this section, we mainly introduce three categories of trackers closely related to our algorithm: tracking by deep learning, tracking by correlation filters and tracking by multiple models respectively.

A. TRACKING BY DEEP LEARNING

Although deep learning is limited to large training datasets and high computational complexity, the CNN-based tracker become popular in the field of object tracking because of their robust feature representation capabilities. In 2013, Wang and Yeung [17] proposed deep learning tracking (DLT) method, which applies the deep neural network to single object tracking task for the first time. Later, trackers based on CNNs are frequently proposed. For example, hierarchical convolutional features(HCF) tracking [15], generic object tracking using regression networks (GOTURN) [18] with high processing speed, end-to-end learning tracker based on multi-domain networks (MDNet) [19], adaptive tracking with deep feature cascades (EAST) [20], and convolutional residual learning tracking (CREST) [21] with reformulating the discriminative correlation filter as one-layer convolutional network.

In addition, trackers based on Siamese networks are also popular, including the fully- convolutional Siamese network (SiamFC) [22], correlation filter network tracking by interpreting the correlation filter as a differentiable CNN layer (CFNet) [23], dynamic Siamese network tracking with online update (DSiam) [24]. The algorithms based on Siamese network provide competitive performance. In this work, we exploit different properties of features. The hand-craft feature is simple but not effective. Moreover, the deeper convolutional features contain rich semantic information while the earlier features carry abundant spatial edge details. We make full use of these feature properties to improve the tracking accuracy.

B. TRACKING BY CORRELATION FILTERS

CFs-based trackers transform correlation into element-wise multiplication in Fourier domain by fast Fourier transform. The maximum value on the response map is regarded as the tracking result. In 2010, Bolme *et al.* [10] utilize the minimum output sum of squared error (MOSSE) to learn the correlation filters on the gray image sequence and achieve efficient computation. Whereafter, circulant structure used for increasing sample numbers is exploited in [25]. Later, more and more work focus on the feature representations for DCF method. Single channel gray features are replaced by multi-channel color name (CN) and HOG features in color names tracker (CNT) [26] and kernelized correlation filters tracker (KCF) [27], respectively. Staple [28] combines color and gradient information and make full use of their complementarity, thus achieving faster speed. With the rise of deep learning, hierarchical convolutional features tracker (HCF) [15] and multi-hierarchical filters tracker (MFT) [29] extract deep features from VGGNet and ResNet respectively to construct filters and their performances have been further improved. Danelljan *et al.* [30] propose continuous convolution operator to fuse the deep feature maps with multiple resolution and the same strategy is used to combine the deep and hand-crafted features in [31]. [32] adaptively integrates the deep and shallow features by evaluation criterion. To overcome scale changes of target in the tracking process, [33] and [34] exploit different scale selection approaches both of which are widely used up to now. Other DCF approaches mainly concentrate on learning model [35], boundary effect [12], [36] and long-term tracking [37]. In this work, we select different features to construct correlation filters learning model and exploit different properties to localize the target accurately and improve the tracking performance.

C. TRACKING BY MULTIPLE MODELS

Trackers with two or more models can reduce drift in the tracking process, advance the ability to cope with challenges and enhance the tracking stability in the complex environment. Grabner *et al.* [38] proposed online Boosting (OAB) to cascade the weak classifiers into a stronger classifier. A reliable fusion framework [39] is proposed to implement a tracker that can predict the target states by frames accurately. Hedged deep tracker (HDT) [40] take a correlation filter trained by hierarchical convolutional features as a weak classifier. The weighted sum of all weak classifiers is the predicted result and the weights are computed by an adaptive hedge method. Tracking-learning-detection (TLD) tracker [5] uses an online detector to correct the results of tracker. Meanwhile, the classifier is updated by samples, which is suitable for long-term target tracking. Multiple experts using entropy minimization (MEEM) tracker [41] exploits multiple expert models to predict target and solves the problem of sample contamination. Long-term correlation tracker (LCT) [11] combines three correlation filters with online SVM to achieve long-term tracking. Multi-cue correlation filters tracker (MCCT) [42] combines different features to form multiple experts, among which the tracker selects the most reliable result as final result in each frame. In this work, we exploit the information of positive and negative samples contained in the classifier to train the online classifier model and correct target localization by correlation filters. It corrects the target matching errors in the candidate region in the DCFs framework and relieves the drift caused by the background information.

III. OVERVIEW OF OUR METHOD

In this work, we propose dual learning models for accurate visual tracking algorithm based on filter banks constructed by multi-level features. Fig. 1 shows the main framework of our proposed algorithm.

Firstly, the traditional hand-crafted features are rough and it is difficult to fully describe posture changes of target in the movement. Therefore, we use three different feature descriptors to form the low-level features. We adopt the network model of VGG-19 [43] to extract the deep features. Convolutional features of lower layers are helpful for accurate localization of target since they contain more comprehensive spatial details. We take the convolutional features extracted from earlier layers as the middle-level features. At the same time, the features extracted from deeper layers contain rich semantic information, which can express the appearance change of target robustly, so the higher layer features are defined as the high-level features. The target can be comprehensively expressed through descriptions of multi-level features together.

Secondly, we propose HPSR to measure the confidence score of response map in each layer to predict target position. Corresponding response maps are output after the construction of the correlation filter banks based on multi-level features is completed and the final HPSR is obtained through weighting them.

Finally, we utilize HPSR to make alternate predictions for the dual learning model. The estimated position of target is determined by the correlation filter learning model. Then the positive and negative samples are sampled around the estimated position for redetection, and the proposal with the highest classification score is taken as the result of classifier model. In order to predict the location of target more

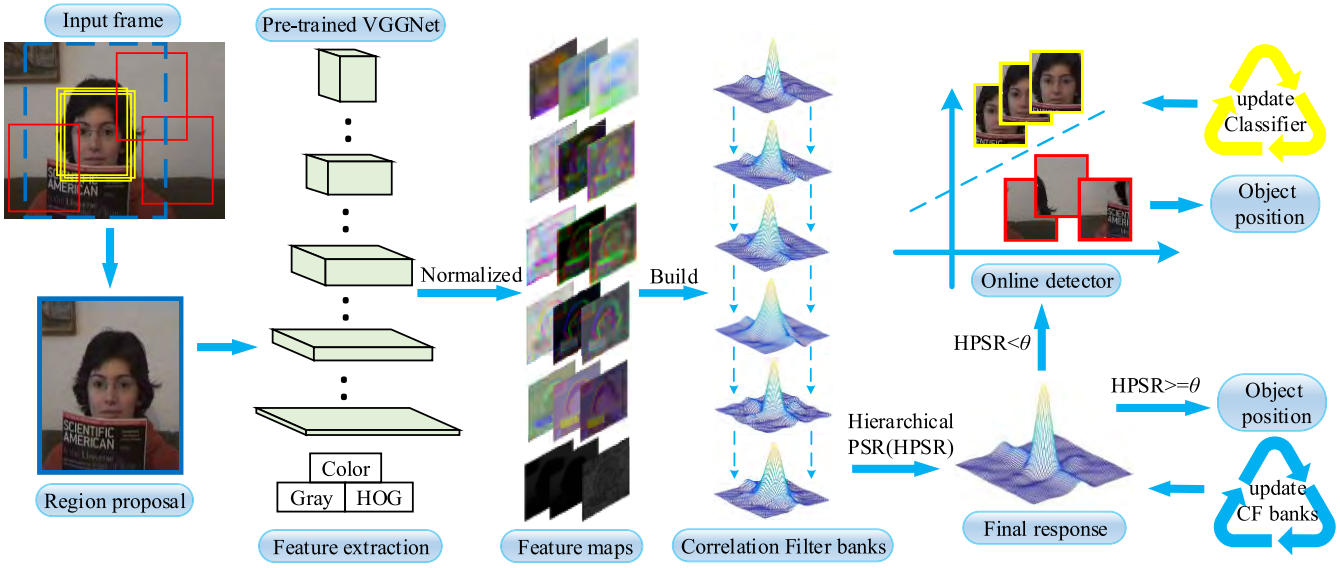

FIGURE 1. Flowchart of the proposed dual model learning combined with multiple feature selection algorithm for accurate visual tracking.

accurately, the results of dual learning models are synthesized to obtain target location. In addition, we exploit the method in [34] to estimate the scales of the target, and finally get the result of target prediction.

IV. PROPOSED METHOD

In this section, we introduce our proposed method including multi-level feature extraction, correlation filter learning model, hierarchical PSR, redetection classifier learning model and dual learning model updating strategy.

A. MULTIPLE LEVEL FEATURE EXTRACTION

In recent years, with the rise of deep learning, many popular deep network models have emerged, including VGGNet [43], AlexNet [44], ResNet [45] and some variant of the network structures. These networks have been applied to large-scale image detection and classification and have achieved remarkable results. We utilize the feature maps extracted by VGGNet to encode the appearance and extract the hand-crafted features of target for auxiliary description. VGGNet is trained by 1.3 million images and has appropriate number of network layers. It can not only provide features from more levels, but also realize fast forward propagation. So it is better for feature extraction.

1) LOW-LEVEL FEATURES

We utilize many hand-crafted feature descriptors to represent target, which are HOG, gray and color name (CN) respectively. As shown in Fig. 2(b)–(d), the HOG features, called Histogram of Oriented Gradient, is a feature descriptor used for object detection in computer vision. The features are constructed by calculating and counting the gradient direction histogram in the local region of the image, which reflect the edge shape information of region block. CN features have rich expressiveness and high identification. It is obtained by transforming RGB space to CN space which can reflect the 11-dimensional thematic color information of the region [26]. The gray features are simple features that contain only brightness information. We concatenate these features as the low-level features of the multi-level features. We extract 31-dimensional HOG feature maps, 1-dimensional CN feature map and 1-dimensional gray feature map from the image patch respectively. Totally 33-dimensional feature maps to represent the low-level features. Remarkably, the sizes of the three kind of feature maps are different from each other and these feature maps should be normalized to a fixed size. Then, we concatenate these features together.

2) MIDDLE-LEVEL FEATURES

With the forward propagation of CNN, the semantic discriminative information of different categories in the image will be strengthened, while spatial details will be gradually lost. Therefore, we retain the earlier layer features as middle-level features. As shown in (e) and (f) of Fig. 2, the facial information and contour of occluded objects in the feature map of (e) and (f) are still clear, and most edges and texture information in the image are preserved. Therefore, we utilize the rich spatial details of these convolutional layers for precise location of target.

3) HIGH-LEVEL FEATURES

Similarly, the deeper layer features with rich semantic information are taken as the high-level features, as shown in (g)–(i) of Fig. 2. A pixel in the feature maps of deeper layers corresponds to a large part of receptive field, which can significantly improve the adaptive ability of target to posture

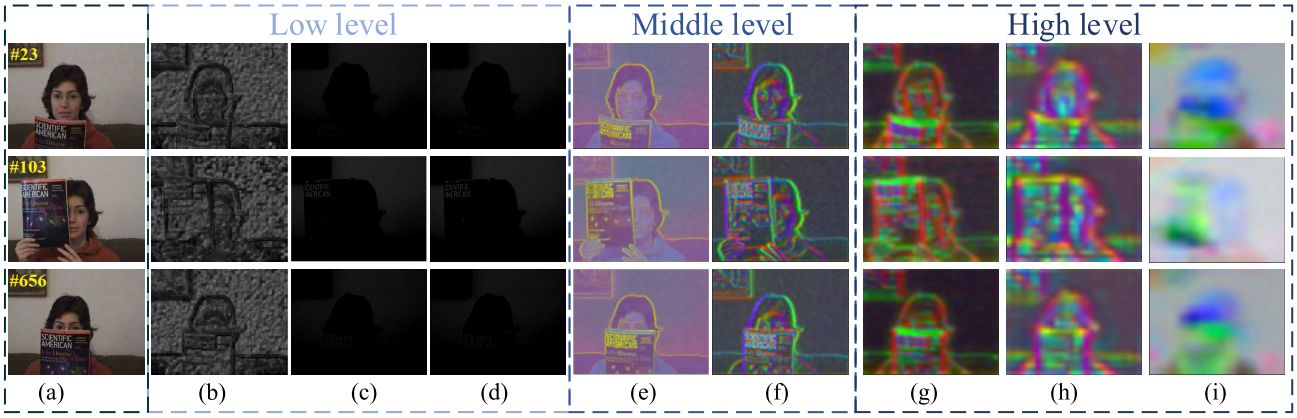

FIGURE 2. Visualization of multi-level features. (a) Three challenging frames on the faceocc1 sequence. (b)–(d) are low-level features which are hand-crafted features including HOG, gray and color name. (e)–(f) are middle-level features extracted from Conv1–2 and Conv2–2. (g)–(h) are high-level features extracted from Conv3–4, Conv4–4 and Conv5–4.

and environmental changes. The high-level feature maps in (i) present the color difference between target and background, which help to describe the dramatic changes in the appearance of target. The feature maps in (g) and (h) combine certain information of surrounding pixels and have a strong discriminative ability to the changes of target appearance.

The resolution of feature maps decreases with the increase of pooling operation in CNN. The input of VGGNet is image with size 224×224. The output convolutional features in pool5 are 7×7 pixels, which are reduced to the original 1/32. Accurate localization on such feature maps with small size is infeasible, so we utilize bilinear interpolation to resize the feature maps to the sizes of correlation filters in the spatial domain.

B. CORRELATION FILTER LEARNING MODEL

In recent years, target tracking algorithms based on correlation filter have been widely used. The trackers based the method can greatly accelerate the processing speed of samples by using the cyclic structure of training and test samples. Let $X_l \in \mathbb{R}^{M\times N\times D}$ denote a set of multi-channel feature maps extracted from the l-th convolutional layer, X_l^d denote the feature maps extracted from the d-th channel of the l-th layer, $d \in\{1,2,\ldots,D\}$. W_l^d denotes the correlation filter established for each channel feature map X_l^d. $Y \in \mathbb{R}^{M\times N}$ is represented as a Gaussian shape label matrix, which obeys the 2D Gaussian distribution and is defined as:

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

where M and N denote the width and height of the convolutional feature map respectively, D denotes the number of feature map channels. The optimal correlation filter W_l^* in the l-th layer needs to minimize the following cost function:

$$W_l^* = \arg\min \| \sum_{d=1}^{D} W_l^d \bullet X_l^d - Y\|^2 + \lambda \|W_l\|^2 \tag{2}$$

where λ ($\lambda \geq 0$) denotes the regularization parameter. Minimization of the above problem (2) is equivalent to utilize FFT to train correlation filters, which can be solved efficiently in the frequency domain. The learned filter on the d-channel in the frequency domain is written as:

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

where$\mathcal{Y} = \mathcal{F}(Y)$,$\mathcal{X}_l^d = \mathcal{F}(X_l^d)$,$\mathcal{F}(\cdot)$ denotes the discrete Fourier transform (DFT), the bar denotes complex conjugation. The operator $\odot$ is the Hadamard (element-wise) product.

In prediction stage for next frame, the candidate patches in the image are obtained and the feature map of the d-th channel in the l-th layer is extracted from them, which is represented as $\mathbf{z}_l^d, z_l^d \in \mathbb{R}^{M\times N\times D}$, its DFT is $\mathcal{Z}_l^d$. The correlation response map of l-th layer is computed as:

$$E_l = \mathcal{F}^{-1}(\sum_{d=1}^{D} \mathcal{W}_l^d \odot \mathcal{Z}_l^d) \tag{4}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform operation. By searching the location of maximum response in correlation response map $E \in M \times N$, we can estimate the target center in the feature map of l-th layer.

We also construct the response maps from the low-level features after obtaining responses of convolutional layer, thus forming correlation response maps with six layers (including five convolutional layer and one-layer fused hand-crafted features).

The response maps of all the multi-layer features are denoted as set as $\{E_1, E_2 \ldots, E_l\}$, and target displacement of each layer is inferred by hierarchical response maps. The final response map $\hat{E}$ is obtained by weighting the response maps of all the multi-layer features. The newly formed response map is defined as:

$$\hat{E}(m,n) = \alpha_1 E_1(m,n) + \alpha_2 E_2(m,n) + \cdots + \alpha_l E_l(m,n) \tag{5}$$

where $E_l(m, n)$ represents the response value of location (m, n) in the response map of the l-th layer, α_l is weight of the l-th layer. The target location is finally estimated by searching for maximum response value of $\hat{E}$ in the t-th frame by:

$$(x_t, y_t) = \arg\max_{m,n} \hat{E}(m, n) \tag{6}$$

Finally, the center position $p_t = (x_t, y_t)$ is the estimated target in the current frame.

C. HIERARCHICAL PSR

The correlation filters determine the position of maximum response value as the center position of target by constructing the response map. Therefore, it is very important to measure confidence of response map. Ref. [10] proposes Peak to Sidelobe Ratio (PSR), which is defined as the ratio of the peak intensity of the main lobe to the strongest side lobe and has been widely used in signal processing. It is formulated as:

$$PSR_l^t = \frac{\max(E_l^t) - \mu_l^t}{\sigma_l^t} \tag{7}$$

where E_l^t is the final response map of the l-th level by Eq. (5), $\max(E_l^t)$ denotes the maximum value of E_l^t, μ_l^t and σ_l^t denote mean and standard deviation of the l-th layer response map in the t-th frame respectively.

As can be seen from the above description, we have used multi-layer features to represent the target and generated response maps. Based on PSR, we propose a new hierarchical PSR (HPSR) with the following formula:

$$HPSR^t = \sum_{l=1}^{L} \beta_l \times PSR_l^t \tag{8}$$

where L denotes the number of feature level, β_l denotes the weight assigned to each layer response map. The larger the value of $HPSR^t$ is, the higher the tracking quality of the t-th frame is.

HPSR can reflect tracking reliability by measuring the fluctuation of the response map. In object tracking, the ideal response map usually have only one sharp peak and are smooth in all other areas, which indicate the tracking result is reliable and the HPSR will be larger. When the target is disturbed by other factors such as occlusion, illumination variation, etc., some non-target responses may be close to the target response value, resulting in intense fluctuation in response map. It indicates that the tracked result is unreliable, with it, HPSR will be smaller. Fig. 3 intuitively shows the HPSR distribution on the *woman* sequence. The target leaves the white vehicle in the 38-th frame, the background changes, and the corresponding HPSR value drops to a low point, such as point A. Both point B and point C are in normal fluctuations, and HPSR can reflect the target changes and environment in which it is located. The target changes dramatically and is partially occluded by the trees from the point D, the corresponding HPSR value drops to the lowest point, that is, point F. At this point, we need to activate our second model to find the target and prevent it from being lost.

D. REDETECTION CLASSIFIER LEARNING MODEL

The classifier learning model is very crucial to correct target location when the target is unstable due to occlusion or background interference. For the classifier learning model, considering the requirement of real-time tracking, we adopt the online SVM classifier with independent training after the prediction of correlation filter learning model.

Different from the long-term tracking algorithms which use the redetection strategy in each frame of sequence, the classifier model to exploits target re-localization improve the tracking accuracy. We utilize the threshold θ to activate the classifier learning model. When the hierarchical maximum response value is less than activation threshold, i.e. $HPSR < \theta$, the classifier model is enabled. The candidate sample set x_{SVM} is collected around the target $p_t = (x_t, y_t)$ obtained by Eq. (6) and the updated SVM is used for classification. The formula $score = w_{SVM}^T x_{SVM} + b$ is classification scores, w_{SVM}^T and b are the parameters and bias of classifier respectively. The target position obtained by finding maximum classification score and the result is obtained by combining the correlation filter learning model.

E. DUAL LEARNING MODEL UPDATING

Visual tracking is a task of target position estimation in dynamic samples, which usually involves model updating. The target and background are constantly changing in the image sequence. It will eventually lead to the tracking failures if the learning model constructed according to the first frame does not change with the target appearance. Therefore, it is necessary to update the learning model with fresh target appearance samples. In this paper, a dual learning models updating strategy is adopted. The correlation filter learning model and the classifier learning model are continuously updated to adapt to the changes of target appearance.

1) UPDATE FOR CORRELATION FILTER LEARNING MODEL

The optimal filter in the l-th layer updated by minimizing the output error of all tracking results. However, this involves solving D correlation filters. If the number of channels are large (e.g. Both Conv5–4 and Conv4–4 have the number of channels of D =512 in VGGNet-19), the price of computation will be very costly. In order to obtain a robust approximation, we update the numerator A_t^d and denominator B_t^d of the t-th frame. The correlation filter W_l^d in the t-th frame can be updated effectively as follows:

$$A_t^d = (1 - \eta) A_{t-1}^d + \eta Y \odot \bar{X}_l^d(t) \tag{9a}$$

$$B_t^d = (1 - \eta) B_{t-1}^d + \eta \sum_{i=1}^{D} X_l^i(t) \odot \bar{X}_l^i(t) \tag{9b}$$

$$W_l^d = \frac{A_t^d}{B_t^d + \lambda} \tag{9c}$$

where t denotes frame index of image sequences, η denotes learning rate. In addition, we update the filters every frame.

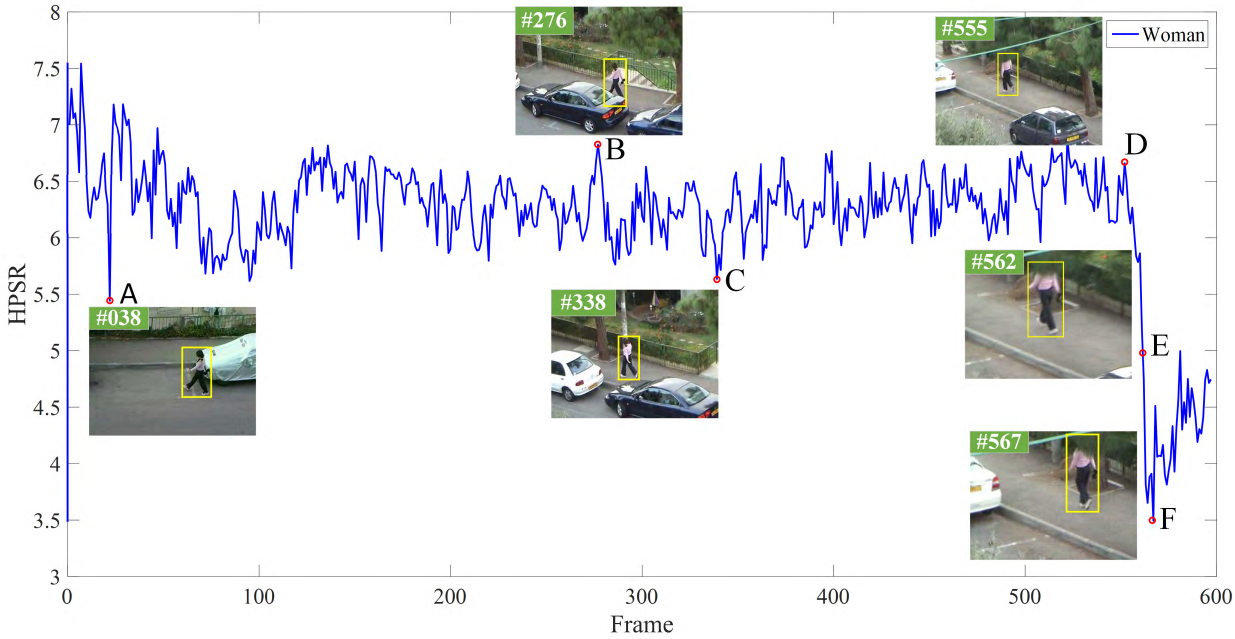

FIGURE 3. Distributions and analysis of HPSR value on the *woman* sequence.

2) CLASSIFIER LEARNING MODEL UPDATE

The online classifier model is updated independently every frame. Positive and negative samples are collected centered around target position of the t-th frame by the dense sampling and only 50 samples are sampled. Sizes of collected samples are same as that of estimated bounding box. Denote the given training data set is $G = \{(x_{SVM,j}, y_{SVM,j}), j = 1, \ldots r\}$, r represents the number of samples, $x_{SVM,j}$ denotes training samples, $y_{SVM,j}$ denotes sample label. The bounding box area of samples is m and bounding box area of target in the t-th frame is n. Their overlap rate can be defined as $s = (m \cap n)/(m \cup n)$, When samples overlap rate s >0.5 are labeled as positive, and when s <0.1 are labeled as negative. The illumination invariant features (IIF) introduced in [41] are extracted from those samples for online SVM training according to the objective function which is as follows:

$$\min \frac{1}{2}\|\boldsymbol{w}_{SVM}\|_2^2 + C_{SVM}\sum_{i=1}^{r} L_h[y_{SVM,j}, \boldsymbol{w}_{SVM}^T \boldsymbol{x}_{SVM,j}] \quad (10)$$

where C_{SVM} denotes penalty parameter, L_h denotes hinge loss.

V. EXPERIMENT RESULTS

In this section, we conduct comprehensive experiments to evaluate the proposed tracking algorithm with dual learning models. Firstly, we describe the detailed implementation of our tracker. Secondly, we analyze effectiveness of each contribution of the proposed algorithm. Finally, we compare our tracker with state-of-the-art trackers. Quantitative, attribute-based and qualitative evaluations are performed on OTB-2013 and OTB-2015 to verify the effectiveness of our proposed tracker.

Object tracking benchmark (OTB) contains two datasets. One is OTB-2013, proposed by Wu *et al.* in 2013, and the other is OTB-2015, proposed in 2015. They contain 50 image sequences and 100 image sequences respectively. OTB involves 11 attributes, including illumination variation (IV), motion blur (MB), deformation (DEF), fast motion (FM), out-of plane rotation (OPR), scale variation (SV), occlusion (OCC), background clutter (BC), out-of-view (OV), in-plane rotation (IPR), low resolution (LR). One-pass evaluation (OPE) proposed in OTB-2013 is used to objectively evaluate the performance of trackers, which mainly adopts two indicators: success plot and precision plot. The success plot represents the percentage of successful frames whose overlap rate between the tracked bounding box and the ground-truth bounding box is larger than the given threshold. Evaluated trackers are ranked by the area under the curve (AUC) of each success plot. The precision plot is defined as the percentage of frames whose average Euclidean distance between the center positions of tracked bounding box and the ground-truth is less than the given threshold.

A. IMPLEMENTATION DETAILS

We adopt VGGNet-19 to extract high-level features (the output of Conv3–4, Conv4–4 and Conv5–4) and middle-level features (output of Conv1–2, Conv2–2) of target candidate regions in the process of network forward propagation. At the same time, we extract low-level hand-crafted features. Then, the multi-level features are formed. The response maps of features contain various information, which provide

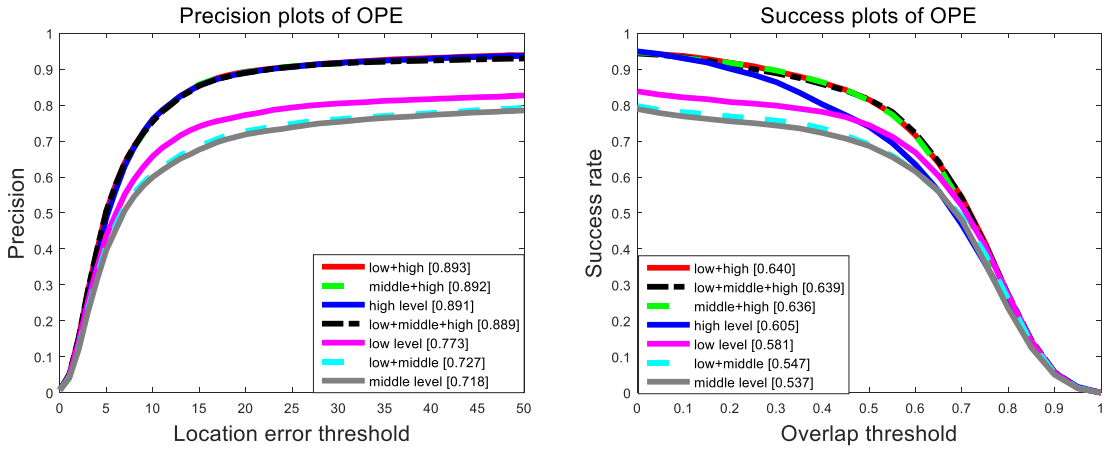

FIGURE 4. Performance comparison of target localization with different multi-level features. The low-level features represent the hand-crafted features formed by the combination of HOG, gray and CN, the middle-level features represent the spatial convolutional features composed of Conv1–2 and Conv2–2, and the high-level features represent the semantic convolutional features consisting of Conv3–4, Conv4–4 and Conv5–4. The rest features are multi-level features combined by features of different levels. The 'low+high', 'middle+high', 'high', 'low+middle+high', 'low', 'low+middle' and 'middle' are represented by red, green, blue, dark, pink, cyan and gray color, respectively.

comprehensive information for target localization. We fuse response map of each level by setting corresponding weights for them. However, we finally select the low-level and high-level features as feature representation. Our experiments are implemented in MATLAB 2015b on a computer with Intel I7-6700K 4.0 GHz CPU, 16GB RAM and a GeForce GTX980Ti GPU card. The version of CUDA is 8.0. Deep learning toolbox named Matconvnet is used in MATLAB to implement our tracker.

In our experiments, some parameters need to be set to fixed values in advance. We set the width of Gaussian kernel σ in Eq. (1) to 0.1, and the regularization parameter λ in Eq. (2) is set to 10^{-4}. Middle-level features include Conv1–2, Conv2–2, and high-level features include Conv3-4, Conv4-4, Conv5-4. In the combinations of ''low+middle+high'', the weights for low-level features, middle-level features and high-level features are set to (0.25, 0.25, 0.25, 0.3, 0.5, 1) from low to high. For the combination of ''low+high'', the weights for low-level features and high-level features are set to (0.25, 0.3, 0.5, 1), respectively.

B. EFFECTIVENESS ANALYSIS

In order to verify the effectiveness of each contribution, we further discuss how to select multi-level features, how to select the thresholds for HPSR to activate the redetection model, and whether incorporating the redetection model is valid.

1) ANALYSIS OF MULTI-LEVEL FEATURES

We compare the low-level, middle-level, high-level features extracted from CNNs and their combinations on the 50 image sequences on OTB-2013. Fig. 4 shows that the combination of low, middle and high features is not ideal on the precision plot while combining low and high features is very prominent. We conclude that the lower features of VGGNet including Conv1–2 and Conv2–2 do not have a significant effect on precise localization of targets. Conv3–4 provides its edge and spatial information, and the semantic information is more important for target localization. Therefore, we select the high-level features integrated by Conv3–4, Conv4–4 and Conv5–4 and the low-level features fused by HOG, gray and CN as the multi-level features of our proposed tracker.

2) THRESHOLD ANALYSIS OF HIERARCHICAL PSR

We have done a lot of experiments to explore how to choose the threshold of HPSR. Firstly, we determine the threshold interval. Then, we compare the threshold in the interval one by one on the OTB-2013 dataset. The distributions of results are shown in Fig. 5. According to the highest precision and success rare, we choose 3.8 as threshold for HPSR. If the value of HPSR is less than this threshold, we activate the classifier learning model to redetect the target. Otherwise, we only rely on the correlation filter model to localize the target.

3) EFFECTIVENESS ANALYSIS OF DUAL LEARNING MODELS

We compare the HPSR with threshold and enable the classifier learning model if the conditions are satisfied. In order to test whether the redetection model works, we incorporate it into the two variations of our tracker for verification, and the results are shown in Table 1. After adding the classifier, Success rates of two variations of our tracker are improved and the precision is improved in tracker combing the classifier with low-level and high features. Therefore, in the case that the HPSR is less than threshold value, it indicates that the single learning model cannot localize target accurately, and certain strategies are needed to correct the predicted position of correlation filter learning model.

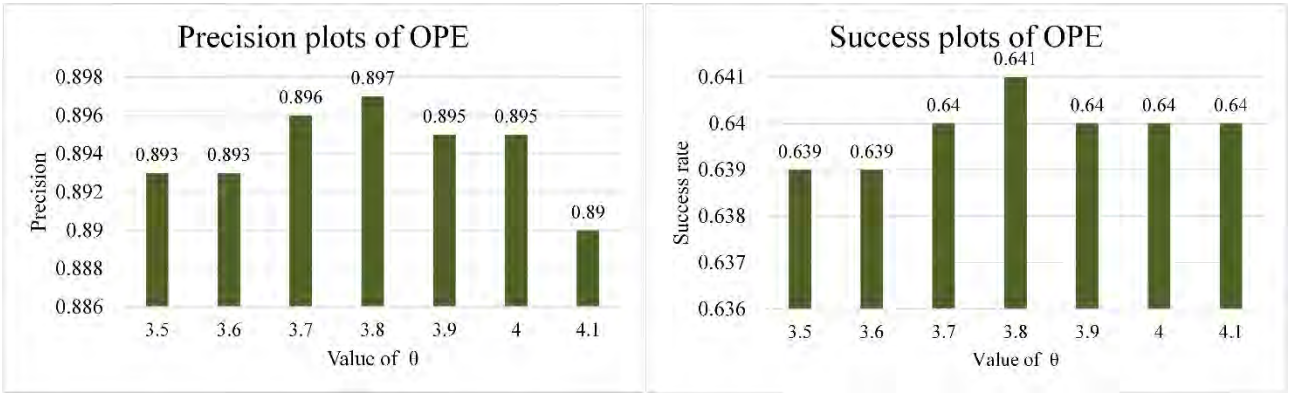

FIGURE 5. Comparison between different thresholds from which we can reasonably select the threshold for comparison with HPSR.

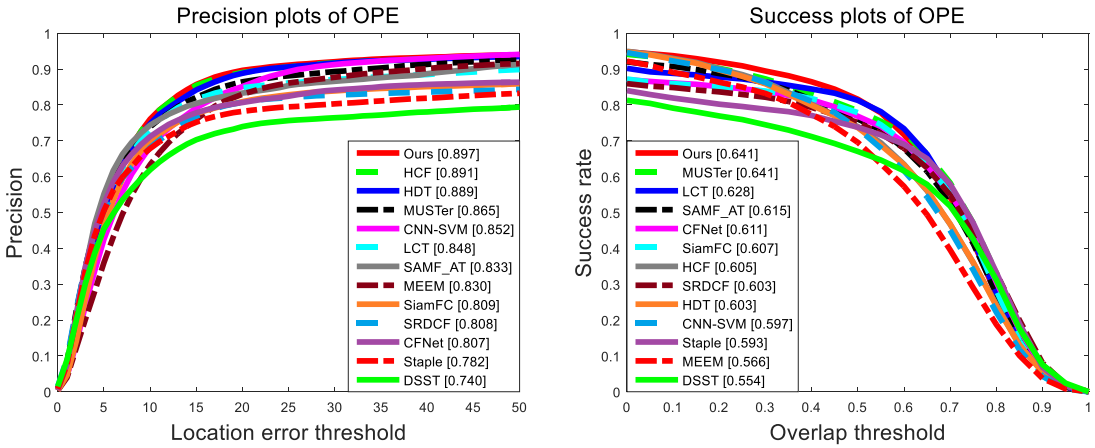

FIGURE 6. Results of precision and success rate on OTB-2013. 50 image sequences are quantitatively analyzed by using the evaluation metrics of OPE. The precision plot with different thresholds for center location error is shown on the left, and the success plot with different overlap thresholds is shown on the right. These trackers are ranked according to the AUC of each tracker. It can be found that our tracker performs favorably against advanced algorithms.

TABLE 1. Comparison between a single learning model and dual learning models. CF model-1 represents the multi-level features composed by low-level and high-level features while CF model-2 represents the multi-level features composed by low-level, middle-level and high-level features. The correlation filter learning model is used for localization, that is, a single model. The validity of the dual models is demonstrated by whether the classifier model is incorporated or not.

Types of models	OTB-2013		OTB-2015	
	Precision	Success rate	Precision	Success rate
CF model-1	0.893	0.640	0.845	0.605
CF model-1 + Classifier	0.897	0.641	0.846	0.606
CF model-2	0.889	0.639	0.830	0.600
CF model-2 + Classifier	0.889	0.641	0.830	0.601

4) EFFICIENCY ANALYSIS

The speed of our algorithm is about 1.23 fps. Inorder to improve the performance of the algorithm, a little speed is lost, but our algorithm has been greatly improved in accuracy and success rate. Deep feature extraction and SVM training are two stages that take more time and the processing time is 0.03973 s and 0.05939 s, respectively. The processing time of hand-crafted feature extraction is 0.00115 s. The time for filter location is 0.00462 s and for relocation of SVM is 0.00298 s. The work in the later stage is to improve the real-time performance of our algorithm.

C. OVERALL PERFORMANCE

In addition, we compare the proposed tracker with 12 state-of-the-art trackers. These trackers can be classified into three types according to the related work:

(1) Tracking by Deep learning. HCF [15], SiamFC [22], FCNT [9], CNN-SVM [6], HDT [40], CFNet [23] are included;

(2) Tracking by correlation filters. SAMF_AT [35], MUSter [37], SRDCF [12], Staple [28], LCT [11], and DSST [34] are included;

(3) Tracking by multiple models. MEEM [41] is included.

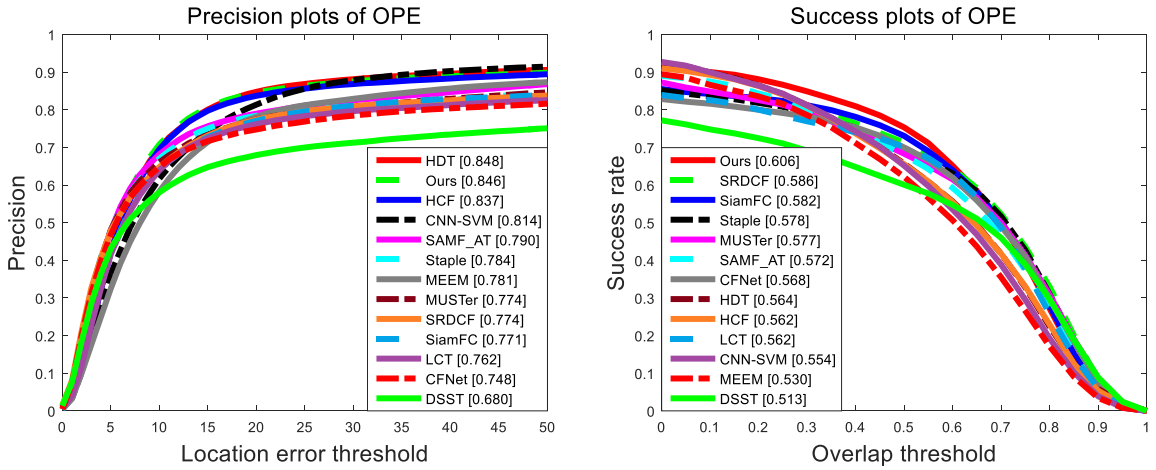

FIGURE 7. Results of precision and success rate on OTB-2015. Quantitative analyses of OPE are performed on 100 image sequences. It can be found that our tracker is 0.2% lower than HDT in precision, but the performance is far ahead in success rate.

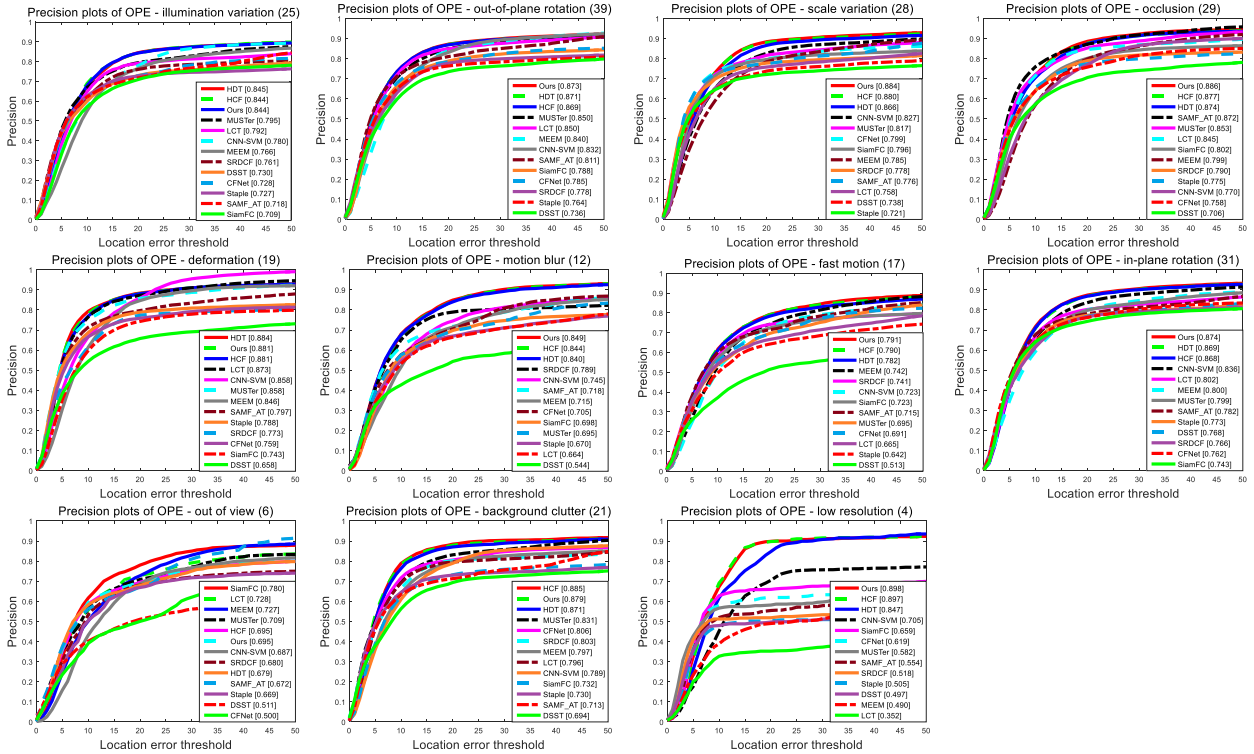

FIGURE 8. Precision plots for attribute-based evaluation on OTB-2013.

1) QUANTITATIVE EVALUATION

In general, according to the evaluation metrics proposed by OTB-2013, our proposed algorithm performs favorably against state-of-the-art trackers. In order to demonstrate the completeness and persuasiveness, the OPE results of proposed tracker and other compared trackers on the OTB-2013 dataset are shown in Fig. 6. Simultaneously, the comparisons on the OTB-2015 dataset are shown in Fig. 7. Note that OTB-2013 is a subset of OTB-2015 dataset, so trackers need to face greater challenges on the OTB-2015 and the performance of trackers on the OTB-2015 is worse than that on the OTB-2013.

2) ATTRIBUTE-BASED EVALUATION

We further use the image sequences annotated by 11 attributes to comprehensively evaluate the performance of trackers. Fig. 8 and Fig. 9 show the accuracy and success rate of our tracker and other 12 trackers respectively on the OTB-2013. We use radar grap to present the attribute-based evaluation of all tracers on OTB-2015 in Fig. 10. The performance of

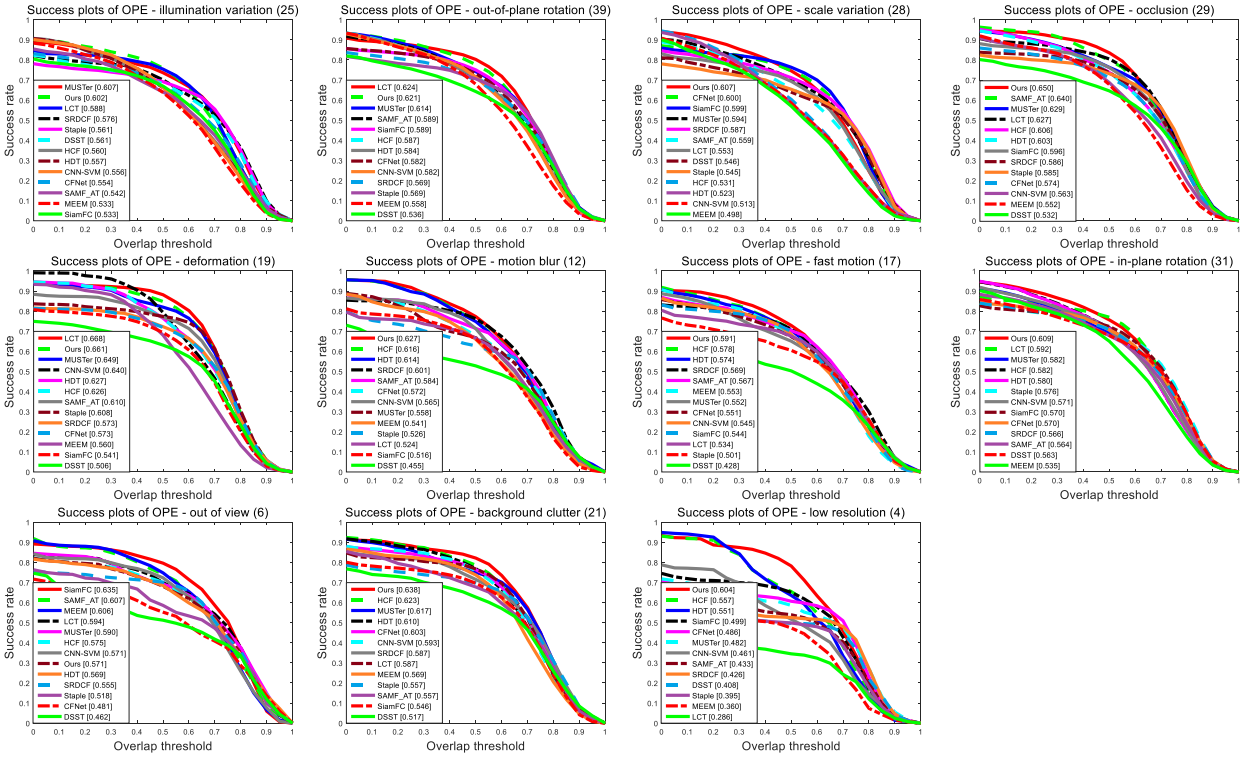

FIGURE 9. Success plots for attribute-based evaluation on OTB-2013.

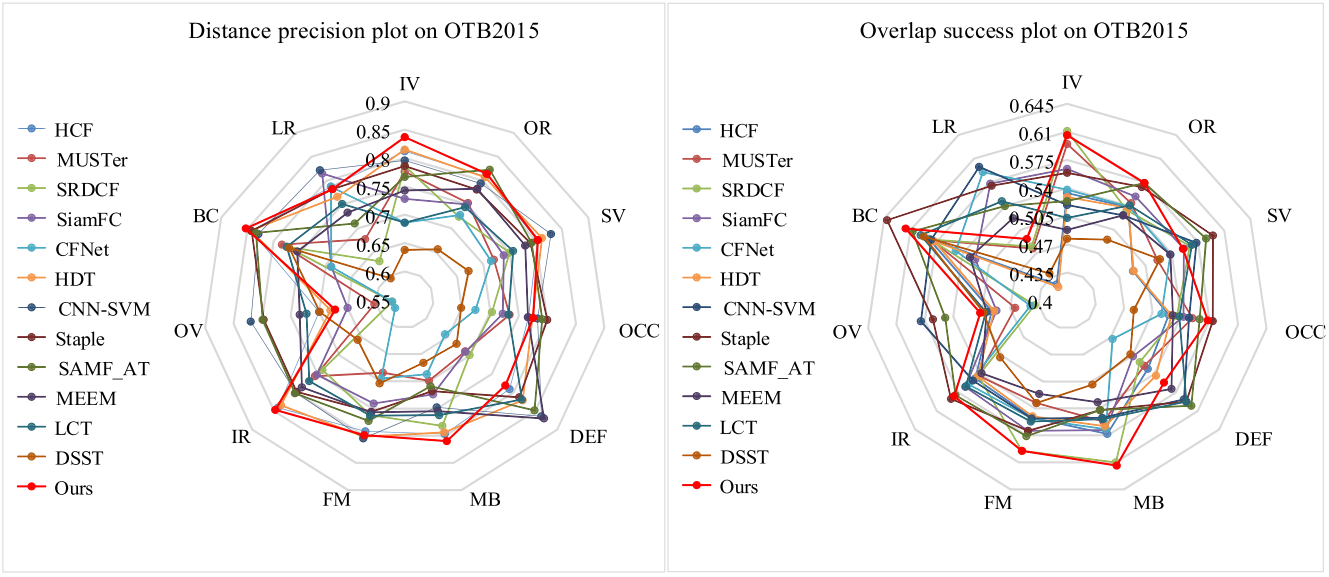

FIGURE 10. Performance evaluations of different attributes on OTB-2015: illumination variation (IV), out-of-plane rotation (OR), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IR), out-of-view (OV), background clutter (BC), and low resolution (LR). Our tracker performs well on all attributes.

each algorithm varies greatly, and there is no algorithm that shows excellent performance on 11 attributes. Our tracker shows excellent performance on most of attributes, but it does not perform well in handling challenges such as OV, DEF and LR. Under the interference of these challenges above, our dual learning models may not track the target again when target disappear completely. However, they show excellent performances in image sequences with complex background and partial occlusion, which are mainly attributed to the multi-level features with rich spatial and semantic details. Moreover, the HPSR can measure the fluctuation of multi-level response maps. If they fluctuate drastically,

FIGURE 11. The display of tracking results. We select the partial tracking results of Staple, CFNet, HDT, SiamFC, HCF and proposed argorithm in 10 chanllenging image sequences (from left to right, top to bottom, respectively,are *soccer*, *carscale*, *clifbar*, *dragonbaby*, *fleetface*, *ironman*, *jogging-2*, *kitesurf*, *skating1*, *singer1*).

the classifier learning model would be activated to perform redetection after the position is estimated by correlation filter model. This process aims at improving the tracking accuracy.

3) QUALITATIVE EVALUATION

Qualitative comparative experiments are performed on several image sequences. Compared trackers include Staple [28], CFNet [23], HDT [40], SiamFC [22], HCF [15], and the proposed tracker. The tracking results, in 10 challenging sequences from the OTB-2015 dataset, are presented in Fig. 11. Staple uses CN and HOG features for image representation and combines their response scores obtained by independently training correlation filters for effective tracking. Although Staple can adapt to scale variations and in-plane-rotation (e.g. *carscale*, *singer1*), it does not perform well in the presence of occlusion, background clutter and fast motion (e.g. *soccer*, *jogging*-2, *dragonbaby*). SiamFC and CFNet are partly similar, because they both use Siamese networks to extract features from training samples and match them to the region of interest in the first frame. However, SiamFC merely uses deep learning for classification, thus it cannot effectively track target in the presence of complex background, blurred motion, and illumination changes (e.g. *clifbar*, *fleetface*, *skating1*). CFNet incorporates correlation filters into the deep network and the classification is more robust. It is adaptive to the occlusion and deformation (e.g. *jogging*-2, *skating1*), but not adaptive to rotation and fast motion (e.g. *dragon baby*, *kitesurf*). HCF and HDT use hierarchical convolutional features learned from large-scale datasets, which are more effective than traditional hand-crafted features. The two algorithms are similar and their performances are almost the same. They can reduce the interference of occlusion, complex background (e.g. *jogging*-2, *skating1*), but still cannot handle the scale changes and rotation. The proposed algorithm shows superior performance on these challenging image sequences. The superior performance is not only due to multi-level features, but also due to the dual learning models which improves the tracking accuracy and makes our tracker well adaptive to the moving process of the target.

VI. CONCLUSIONS

In this work, we propose dual learning models and multiple features selection to solve the problems caused by the integration of multi-layer convolutional features and correlation filters. Our method can effectively overcome the following

two situations: 1) Semantic features doped with the invalid information lead to unreasonable description of target, subsequently causing transient drift and even loss of target under various challenging factors. 2) Correlation filters have limitations on targets with fast motion and rapid appearance changes. Unreasonable model updating causes tracking failure as well. Our method fuses hand-crafted features with features extracted from the multi-layer convolutional network, which adds basic features that the deep features lacks and makes the descriptions of target improved. The proposed HPSR index can comprehensively reflect the fluctuation of correlation filters at all levels. When the fluctuation is abnormal, the online classifier learning model is activated to predict the target position again and eliminate the error accumulation of correlation filters. The experiments demonstrate that the proposed algorithm with dual learning models shows superior accuracy and performance on OTB-2013 and OTB-2015.

REFERENCES

[1] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 798–805.
[2] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 213–228, Jan. 2015.
[3] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.
[4] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.
[5] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.
[6] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, Feb. 2015, pp. 597–606.
[7] Y. Tu, Y. Lin, J. Wang, and J.-U. Kim, "Semi-supervised learning with generative adversarial networks on digital signal modulation classification," *Comput. Mater. Continua*, vol. 55, no. 2, pp. 243–254, 2018.
[8] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2013, pp. 2553–2561.
[9] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 825–841.
[10] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.
[11] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5388–5396.
[12] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4310–4318.
[13] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
[14] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
[15] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3074–3082.
[16] J. Zhang, X. Jin, J. Sun, J. Wang, and A. K. Sangaiah, "Spatial and semantic convolutional features for robust visual object tracking," *Multimed. Tools. Appl.*, pp. 1–21, Aug. 2018. doi: 10.1007/s11042-018-6562-8.
[17] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2013, pp. 809–817.
[18] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 749–765.
[19] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4293–4302.
[20] C. Huang, S. Lucey, and D. Ramanan, "Learning policies for adaptive tracking with deep feature cascades," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 105–114.
[21] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2574–2583.
[22] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2016, pp. 850–865.
[23] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5000–5008.
[24] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1781–1789.
[25] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 702–715.
[26] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1090–1097.
[27] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[28] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1401–1409.
[29] S. Bai, Z. He, T.-B. Xu, Z. Zhu, Y. Dong, and H. Bai. (2018). "Multi-hierarchical independent correlation filters for visual tracking." [Online]. Available: https://arxiv.org/abs/1811.10302
[30] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 472–488.
[31] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6931–6939.
[32] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 493–509.
[33] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 254–265.
[34] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014, pp. 1–5.
[35] A. Bibi, M. Mueller, and B. Ghanem, "Target response adaptation for correlation filter tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 419–433.
[36] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler. (2018). "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking." [Online]. Available: https://arxiv.org/abs/1807.11348
[37] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "Multi-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 749–758.
[38] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2006, pp. 47–56.
[39] O. U. Khalid, J. C. SanMiguel, and A. Cavallaro, "Multi-tracker partition fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 7, pp. 1527–1539, Jul. 2017.
[40] Y. Qi *et al.*, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4303–4311.
[41] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 188–203.

[42] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4844–4853.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 1–13.
[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

JIANMING ZHANG received the B.S. degree from Zhejiang University, in 1996, the M.S. degree from the National University of Defense Technology, China, in 2001, and the Ph.D. degree from Hunan University, China, in 2010. He is currently an Associate Professor and the Deputy Dean with the School of Computer and Communication Engineering, Changsha University of Science and Technology, China. He has published more than 70 research papers. His main research interests include the areas of computer vision, data mining, and wireless ad hoc and sensor networks. He is a Senior Member of CCF.

XIAOKANG JIN received the B.S. degree from the Changsha University of Science and Technology, China, in 2016, where he is currently pursuing the M.S. degree in computer science and technology. His research interests include computer vision, deep learning, and object tracking.

JUAN SUN received the B.S. degree from the Changsha University of Science and Technology, China, in 2018, where she is currently pursuing the M.S. degree in computer science and technology. Her research interests include computer vision and object tracking.

JIN WANG received the B.S. and M.S. degrees from the Nanjing University of Posts and Telecommunications, China, in 2002 and 2005, respectively, and the Ph.D. degree from Kyung Hee University, South Korea, in 2010. He is currently a Professor with the School of Computer and Communication Engineering, Changsha University of Science and Technology. His research interests mainly include wireless communications and networking, performance evaluation, and optimization. He is a member of ACM.

KEQIN LI is currently a SUNY Distinguished Professor of computer science with the State University of New York. He has published over 620 journal articles, book chapters, and refereed conference papers. His current research interests include cloud computing, fog computing and mobile edge computing, energy-efficient computing and communication, embedded systems and cyber-physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU–GPU hybrid and cooperative computing, computer architectures and systems, computer networking, machine learning, and intelligent and soft computing. He received several best paper awards. He currently serves or has served on the editorial boards of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, the IEEE TRANSACTIONS ON SERVICES COMPUTING, and the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING.