TGMCF: a tree-guided multi-modality correlation filter for visual tracking.

LIU, Q., LIU, W., WANG, Y., REN, J., DU, Q., LV, Y. and SUN, H.

2019

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 29, 2019, accepted September 19, 2019, date of publication September 26, 2019, date of current version November 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943917

TGMCF: A Tree-Guided Multi-Modality Correlation Filter for Visual Tracking

QIAOYUAN LIU[1,2], WEIWEI LIU[1], YURU WANG[1], JINCHANG REN[3], QIAO DU[1], YINGHUA LV[1], AND HAIJIANG SUN[2]

[1]Department of Information Sciences and Technology, Northeast Normal University, Changchun 130117, China
[2]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Beijing 130000, China
[3]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, U.K.

Corresponding authors: Yuru Wang (wangyr915@nenu.edu.cn), Yinghua Lv (luyh@nenu.edu.cn), and Haijiang Sun (sunhaijiang@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61300099, in part by the China Postdoctoral Science Foundation Funded Project under Grant 2015M570261, and in part by the Science and Technology Development Plan of Jilin Province under Grant 20170101144JC, and in part by the China Scholarship Council under Gramt 201706625002, and in part by the Key Laboratory of Aeronautical Optical Imaging and Measurement.

ABSTRACT For updating the tracking models, most existing approaches have an assumption that the target changes smoothly over time. Despite their success in some cases, these approaches struggle in dealing with occlusion, illumination changes and abrupt motion which may break the temporal smoothness assumption. To tackle this problem, in this paper we propose a tree-guided visual tracking model based on the multimodality correlation filter which could estimate the target state according to the most reliable information in previous frames. We maintain a representative target state set in a tree model over the whole tracking process. Ideally, the tree model is able to capture all the landmark states of the target, and provides a confident template for the correlation filter. Therefore, we propose an optimal updating strategy to record the most recent stable and representative states for tree updating. By utilizing stable target-states for template training, the multi-modality correlation filter is able to output a more accurate target position than the baseline and the SOTA (state-of-the-art) methods. Tested on the OTB50 (object tracking benchmark) and OTB100 dataset, the proposed TGMCF has demonstrated outstanding performance on several typical tracking difficulties and overall comparative results with the SOTA trackers are obtained on several public tracking benchmarks.

INDEX TERMS Visual tracking, tree-guided, multimodality correlation filter.

I. INTRODUCTION

Visual tracking is a fundamental computer vision task with a wide range of applications such as human tracking [1] and robot perception [2]. In the process of visual tracking, the quality of model is decisive to the tracking performance. Most existing trackers utilize only one model from the beginning to the end, assuming that the tracking methods could continuous output accurate results [3], [4]. However, there exist so many challenges in practical visual tracking that such assumptions are not held true. Actually the tracker can't guarantee the result of every frame to b accurate in the tracking process, drift or temporary loss is inevitable, once they are considered the whole model would be polluted.

Inspired by the tracker TCNN [5] which is based on multiple CNNs [6] and maintained a tree structure to well handle the pollution problem in the training model, we introduce this idea into the correlation filter tracking and achieved a comparable performance with SOTA trackers.

The method of TCNN was ranked the second in the 2016 VOT challenge [7], in which a tree structure with multiple CNNs was built to estimate target states and determine the desired path for online model updates. The target state is estimated by sampling target candidates around the state in the previous frame, and the best sample is identified as the result of a weighted average score from a set of CNNs. Although this method achieves an outstanding accuracy in tracking, it is quite time-consuming, as each modality in the tree structure needs to be trained by CNN. As a result, the TCNN could only achieve 1.5fps tracking speed which is far from real-time tracking.

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

At the same time, another important branch of trackers: CF (correlation filter) [8] has developed into one of the most important technologies in the field of visual tracking due to its fast speed and high accuracy. Similar to SVM (support vector machines) [9], the correlation filter is also a supervised algorithm. Different from SVM that always focuses on binary classification with label 0 or 1, CF would train samples with continuous labels for a confidence map before eventually determination. Finally, the positions with the maximum responses would be set as the center of target. The advantage of the correlation filter is to utilize previous results as training samples for model training, which can easily transmit dynamic target information into subsequent frames. However, most existing correlation filter based trackers only consider the results of adjacent frames in training [10]. When the training set is full, the method would discard the oldest sample to let the new sample in. This strategy seems to make sense as it assumes that recent samples would contribute more to training but actually, it is illogical. Once inaccurate results were integrated into the training set, the trained model would easily get over-fitting on the recent frames and finally cause a drift on the result.

Considering the advantages and disadvantages of both the TCNN and the correlation filter, a TGMCF (tree guided multi-modality correlation filter) based tracker is proposed to estimate the target's position, aiming to achieve better results with higher tracking speed. In the proposed TGMCF approach, the most classic correlation filter KCF (Kernel zed Correlation Filter) [11] was selected as the baseline to ensure the effectiveness of the tree structure, specifically each correlation model is trained by a frameset with a fixed number of frames. Since every correlation model can be regarded as a state of the target, we add them as nodes into the tree structure. The state of the target would keep changing during the tracking process, hence there would be constantly nodes that need to be added into the tree structure. We estimate the reliability of every new node and its response with the existing nodes in the tree structure to determine its parent node and update the model. Whenever a new node is added, the whole model would be updated online. Besides we also introduced an adaptive optimization strategy for the construction of the tree, specifically the reliability of the new node is compared with a threshold to estimate the quality of the new node. The structured correlation model trained with the target stages could effectively avoid the disturbance of unstable target information to tracking and be more reliable than the conventional correlation model trained with all the previous results. To this end, a significant performance is achieved in different tracking challenges with our proposed approach.

The main contributions of the paper can be summarized as follows: i) a correlation filter model built in a tree structure, i.e. TGMCF, is proposed for visual tracking considering the stages of target rather than all the previous results. ii) an adaptive optimization strategy is introduced according to the reliability of each node in the tree structure, which could update the model online and ensure the discrimination of the entire model. Comprehensive experiments on the widely used OTB50, OTB100 and VOT2015 datasets have fully demonstrated the superior performance of the proposed TGMCF method in comparison to a number of SOTA trackers.

The remaining part of this paper is organized as follows. We first review the various correlation filter based visual tracking methods and the multi-model tracking methods in Section 2 and then introduce the baseline correlation filter algorithm KCF in Section 3. Section 4 discusses the proposed TGMCF algorithm in detail and the experimental results are presented in Section 5. Finally, some concluding remarks are drawn in Section 6.

II. RELATED WORK

In the last 2-3 decades, numerous methods have been proposed for visual tracking, including deep learning based tracker [12], [13], correlation filter based tracker [14], [15] and some other traditional trackers [16], [17]. In this section, a comprehensive survey about the SOTA trackers, which relative to the method proposed would be given.

The tracker proposed in this paper is essentially a correlation filter based tracking method. As we know, correlation filters have become one of the most popular technologies in visual tracking, which mainly attempts to minimize the sum of squared error between the desired correlation response and the circular correlation of the object patch. This method considers that the obtained previous results can be helpful for the subsequent tracking, and the model trained by using previous results as training sample would be more reliable. Danelljan *et al.* [8] first introduced the correlation filter into visual tracking problem and proposed the MOSSE tracker in 2010, which uses the grey-scale image to extract a single channel feature with a high speed. Henriques *et al.* [11] proposed a kernelized correlation filter (KCF) by introducing the kernel trick into ridge regression, which could solve for the filter taps efficiently by utilizing a circulant matrix. Since then, the correlation filter became one of the most widely used strategies for visual tracking. Spatially Regularized Discriminative Correlation Filter (SRDCF) [4] utilizes spatial regularization by introducing a spatial regularization component, which can penalize the correlation filter coefficients during learning and lead to not only alleviating the unwanted boundary effects but also allowing the CF to be learned on larger regions. SRDCFDecon [30] decontaminates the training set by estimating weights of all training sample in each frame.

Despite of correlation filter, deep learning is also an important development direction in visual tracking. Due to the great representative ability of the CNN feature, Danelljan *et al.* [13] proposed to apply deep convolutional features in visual tracking, aiming to increase discriminative of the tracking method. The tracker of UPDT [36] system analyzed the effect of shallow and deep features on target tracking. They found that the depth model should be responsible for the robustness of the network, while the shallow model should be responsible for the accuracy of location. The TCNN [5] method proposes to construct a tree based multi-model with CNN to handle the change of target appearance, by storing the deep features in nodes and analyzing the scores of every nodes in tree to locate the target's position. The process of feature extraction in CNN is a multi-layer progressive processing, which is an imitation of the human brain on information processing. Usually, the more data used for training, the more effective the model would be. Although the CNN feature could be great helpful for improving the tracking accuracy, compared with the traditional hand-crafted features, such as Harr like [18], HOG (Histogram Oriented Gradient) [19], and LBP (Local Binary Pattern) [20] which could realize a fast tracking speed with fewer parameters, the CNN feature need much more time for model training.

Among various tracking models, the multi-modality model is a relatively effective strategy for target representation which could well eliminate the errors caused by model pollution [21]–[23]. Trackers based on sparse representation [24] apply multiple target templates could compute the likelihood of each sample by minimizing its reconstruction error while integrating multiple observation models via an MCMC (Markov Chain Monte Carlo) framework. Besides, ensemble classifiers applied to visual tracking problem could also gain a good achievement. Wang *et al.* [25] proposed a hybrid ensemble classifier based on two types of classifier (LDM [26] and SVM [9]) for visual tracking and achieved a good performance on the challenges of complex background and occlusion. Therefore it can be concluded that in visual tracking the jointly decision made by multi-modality could be more reliable than single models.

In this paper, we propose to introduce the tree structure into the correlation filter based tracking. Comparing with existing similar work such as Ensemble_SW_Obs [37] and AECF [35], AECF aims to extract the key CFs from the previous frames to decrease computational burden yet it fails to consider the topological relationship between different CF templates. The Ensemble_SW_Obs did introduce a tree structure into correlation filter, but the tree is a binary one with the edge value of either 0 or 1. Different from these two trackers, our approach takes the relationship between the templates as the value of the edges in the tree. The detailed of TGMCF will be discussed in the Section IV. Moreover, the experimental results show that our method could achieve superior performance over these methods.

III. BASELINE FRAMEWORKS

Although many correlation filter based trackers have shown dominant and impressive results in visual tracking, the KCF [11] method would be the classic one with milestone significance. Actually, most correlation filter based trackers are essentially based on KCF [11]. To illustrate the generalization of the proposed strategy, we take KCF as the baseline tracker in our paper.

In KCF, the circulant matrix is introduced to collect both positive samples and negative samples; besides the ridge, regression is also employed to train the target detector. The aim of this method is to learn a multi-channel correlation filter$f$ from a set of training sample $\{(x_k,y_k)\}_{k=1}^{t}$. Specifically f is represented by a linear regression function as $f\ (x_k) = \omega^T x_k$ and the objective function is designed to minimize the squared error over samples x_i and each element of the regression targets y as follows,

$$\underset{w}{min} \sum_i (f\ (x_i) - y_i)^2 + \lambda \|w\|^2 \tag{1}$$

where ω is the weight coefficient and λ is a regularization parameter that controls over fitting, as same as the ω in SVM. It is worth mentioning that the circulant matrix used in KCF successfully converted the multiplication between matrixes into the dot product in the Fourier domain. This step not only reduces the computational complexity but also improves the tracking speed to a real-time level

The most important strategy in this method is the introduction of kernel tricks in model training, which also forms a significant contribution to speed improvement as detailed below:

$$f\ (x_k) = \omega^T x_k = \sum_{i=1}^{n} \alpha_i k(x_k,x_i) \tag{2}$$

where k is the kernel function used. By applying the correlation filter f trained according to the previous results on the detection region in the current frame, a response map can be obtained, on which the center of the target can be determined as the position with the maximum responses.

IV. THE PROPOSED TGMCF ALGORITHM

This section introduces the construction of our TGMCF for visual tracking, and then detail discusses the procedures of the target estimation for the tracking algorithm, and the method of maintaining multiple KCF models in a tree structure for robust tracking.

The procedure of the proposed TGMCF model is illustrated in FIGURE. 1.

A. TREE CONSTRUCTION

In our TGMCF approach, a tree structure based on KCF is built to describe the target appearances in different stages. The tree structure is built and updated in the whole tracking process, where each node is generated every M frames until the number of the nodes in the tree reach N. When the tree is full, we would only update the existing nodes with new nodes in order to keep the fixed number of nodes. The details about the correlation filter tree are discussed as follows.

In the first frame, the HOG feature of ground truth is extracted to be the first node in the tree structure. In our approach, what stored in every node is the correlation template trained by recent framesets, in which every frameset consist of M frames, except for the first node.

During the tracking process, every tracking result would be obtained according to all the existing nodes in the tree, and each new node is generated every M frames. In the tree structure $T = \{V, E\}$, a vertex $v \in V$ corresponds to a correlation filter trained by KCF, and a directed edge

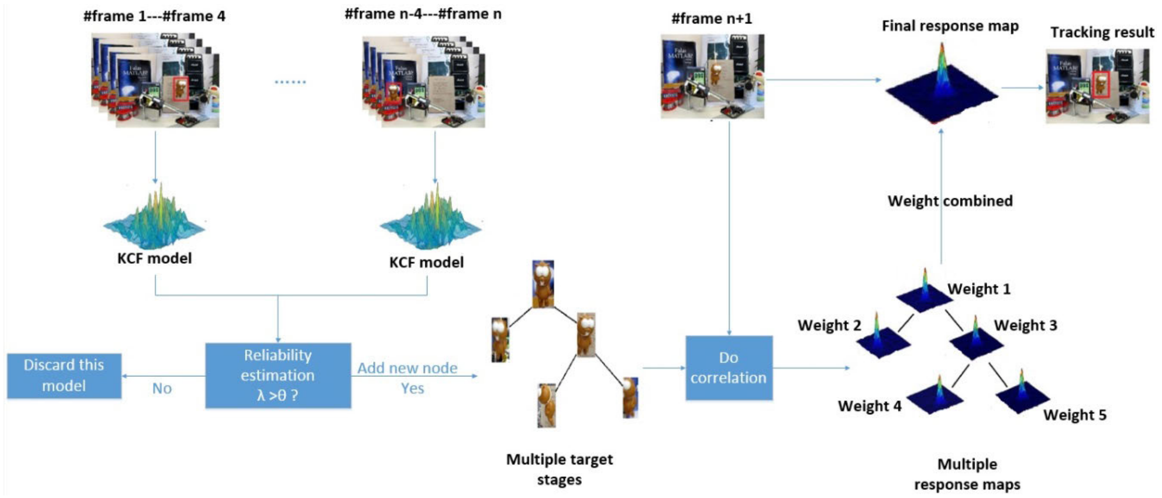


FIGURE 1. The flowchart of the proposed TGMCF model.

$(u, v) \in \varepsilon$ defines the relationship between correlation filters. The score of an edge (u, v) is the determined by its two end vertices as follows,

$$S(u, v) = \frac{1}{|F_v|} \sum\nolimits_{t \in F_v} \delta_{u \to t} \tag{3}$$

where F_v is a set of consecutive frames for KCF model training associated with v, and $\delta_{u \to t}$ is the predicted positive score of the node u while tracking with the t^{th} frame, it can be calculates as follows,

$$\delta_{u \to t} = \frac{1}{|g|} \sum\nolimits_{i \in x} g_t^i \tag{4}$$

where g is a set consists of x response maps with top x response values while tracking the t^{th} frame, in another word we consider the most relevant x templates to evaluate the node u.

With new nodes added, if the tree structure updates only depend on tracking results, the model is also prone to deviations. Once the drift result is utilized for the model update, the discriminate ability of the tree would be decreased and the subsequent tracking results will also be affected. Therefore, despite of evaluating the score of edge, the reliability λ_v of each node also needs to be considered, which represents the score of every correlation filter stored in a node. Specifically, the λ_v can be obtained by

$$\lambda_v = min\left(s\left(p_v, v\right), \lambda_{p_v}\right) \tag{5}$$

where the reliability is computed recursively, and p_v represents the parent node of the node v, which should be chosen before adding a new node.

A parent node is determined by comparing its reliability between all the existing nodes when the new node is added, and the node with the maximum reliability will be the decided parent node below.

$$p_z = arg\, max_{v \in V_+} \lambda_z^v \tag{6}$$

where V_+ represents the existing nodes in the tree, and λ_z^v represents the reliability of node z when it is connected to node v, we also have

$$\lambda_z^v = min\left(s\left(v, z\right), \lambda_v\right) \tag{7}$$

Similar to Eq. 5. We believe that not every node would contribute to tracking; hence a threshold θ is set for the reliability, any new node with its reliability below θ would be abandoned. Otherwise, this new node would be added into the tree structure.

When a new node is added, the correlation filter of the parent node would be updated according to the frameset in both the new node and its parent node

$$\mathrm{mf} = (1 - \eta) * \mathrm{mf} + \eta * \mathrm{f_t} \tag{8}$$

where mf represents the correlation filter in the new node, it is initialized as the correlation filter in its parent node at first, and η is a parameter for updating, t $\in \mathrm{F}_{(\mathrm{p,z})}$ is the sum of frames in the parent node and its child node, so $\mathrm{f_t}$ consists of the features extracted from both the parent frameset and the child frameset,. Especially, when the number of nodes reaches the upper limit N, we will always keep the most recent N nodes in the tree.

B. TARGET STATE ESTIMATION USING MULTIMODALITY TREE MODEL

In the correlation filter based tracker, the tracking result is estimated from the detection region determined by the result of the previous frame. In this part, HOG feature of the detection region is extracted at first then do correlation operations between the detection region and every template in existing nodes V_+ in the tree before determining $|\mathrm{V}_+|$ response maps. Because the filter templates are stored during different states, so every template in nodes would have different effect on target localization. In our approach, every response map is assigned with a weight for analysis, which is estimated in two parts. One is the reliability of node λ_v as discussed in Eq. 5, and the other is the correlation value between the current target and every existing node in the tree structure, which can be given by

$$\gamma_{v\to t} = max\emptyset_{v\to t} \tag{9}$$

where $\emptyset_{v\to t}$ represents the response map of node v in frame t.

Meanwhile, the correlation value can also show the impact of each node on the tracking result. We consider the maximum response value of this node as the correlation value between the current target and each existing node. With the value of $\gamma_{v\to t}$ and λ_v, the weight can be computed by

$$w_{v\to t} = \frac{\min(\gamma_{v\to t}, \lambda_v)}{\sum_{v\in V+}(\gamma_{v\to t}, \lambda_v)} \tag{10}$$

The weight is determinative to the new node associated with frame t when it is updated from the correlation filter of node v. Finally, the target can be localized with the weights and multiple response maps via Eq. 11 and Eq. 12.

$$L_t = \sum_{v\in V+} w_{v\to t}\emptyset_v \tag{11}$$

$$x_t = arg\max L_t \tag{12}$$

To better illustrate our approach, the overall algorithm is also shown in Algorithm 1.

Algorithm 1 Framework of the Proposed TGMCF Tracking Method

```
Input: Video frames $I_1, I_2, \ldots I_t$. Target state $x_0$ at the first frame; Number of vertex n in the tree. Number of frames m used to generate a new node. Threshold $\theta$ for adding the current node.
Output: Target states $x_1, x_2, \ldots, x_t$;
Initialization: Initialize target state $x_0$ according to the ground-truth data; Extract the feature from $x_0$ as the first node of the tree using KCF;
For all time step t do
    i. Determine the detection region according to the result of previous frame $x_{t-1}$, and extract HOG features to do correlation with all the filter templates stored in the tree, then multiple response maps Ø could be obtained;
    ii. Calculate the correlation value $\gamma$ between each node and the current frame using Eq. 9;
    iii. Calculate the weight $w$ of each node for the current frame with $\gamma$ and the reliability $\lambda$ according to the Eq. 10, when there is only one node in the tree, the reliability value is set as null by default;
    iv. Use Eq. 11 to process the final response maps, and ensure the target state $x_t$ according to Eq.12.
    v. Every m frames tracked, one new node and its parent node could be obtained according to Eqs. 5-7. when the reliability of the new node is less than the threshold $\theta$, this node will be discarded; otherwise use Eq. 8 to update the new node and store it in the tree;
    vi. If the tree structure has reached N nodes, just keep the latest n nodes and discard the older ones;
End for
```

V. EXPERIMENT

For performance evaluation, both quantitative and qualitative assessments are adopted using the OTB50, OTB100 and VOT benchmark dataset [27], [31], [32]. In addition, a comprehensive comparison with the representative SOTA trackers is also provided in this part.

A. PARAMETER SETTINGS

All our experiments are tested on MATLAB2016 using an Intel(R) Core(TM) 2.30GHz CPU with 8GB RAM.

We compared the effect of some key parameters including target padding, the parameter σ in Gaussian, number of nodes, threshold to evaluate the coming node, and the number of frames in each node, which are shown in FIGURE. 2. Different parameters are used for tracking the same video sequence "Tiger2", and the optimal values of the parameters are determined by the criterion. While the ACLE is the average Euclidean distance between the centers of the predicted and ground truth bounding boxes. Smaller ACLE represents better performance of the tracker. The AOR is the average overlap ratio. Larger AOR represents better performance of the tracker. Both the ACLE and AOR are the metrics to measure the tracking performance frame by frame.

The parameter of target padding is a scale expansion based on original target frame. If this parameter were too large, too much background would be considered which might distract the overall model, on another hand, if this parameter is too small, it would be difficult for the tracker to adapt for complex tracking conditions. It can be concluded from FIGURE. 2 that when target padding is 1.8, the ACLE is the smallest and AOR is the largest. At this time, better result can be obtained for tracking.

It is easy to know that if too many nodes stored in the tree structure, the long-term target states will inevitable have an impact on the current state analysis, conversely if too few states were considered; the model would easily get over fitting to the neighboring frames. By analyzing the data shown in FIGURE. 2, it can be decided that it is optimal to set seven nodes in the tree structure which would help to gain the best tracking performance.

The threshold measures the quality of the coming nodes, if the threshold is too small, the training samples in the tree structure would be insufficient to reflect the overall change of the target, which would lead to an inaccurate tracking. On the other hand, if the threshold is too large, the qualities of the nodes in tree would be decreased, which will also affect the model's discriminative ability. By analyzing the data shown in FIGURE. 2, it is best to set the threshold for adding the current nodes and updating the entire model to 0.27.

Similarly, if there are too many frames in a frame set, a node may contain not only one state, which may lead to error analysis. If the number of frames in the frame set is too small, it is possible that the next node will still contain the

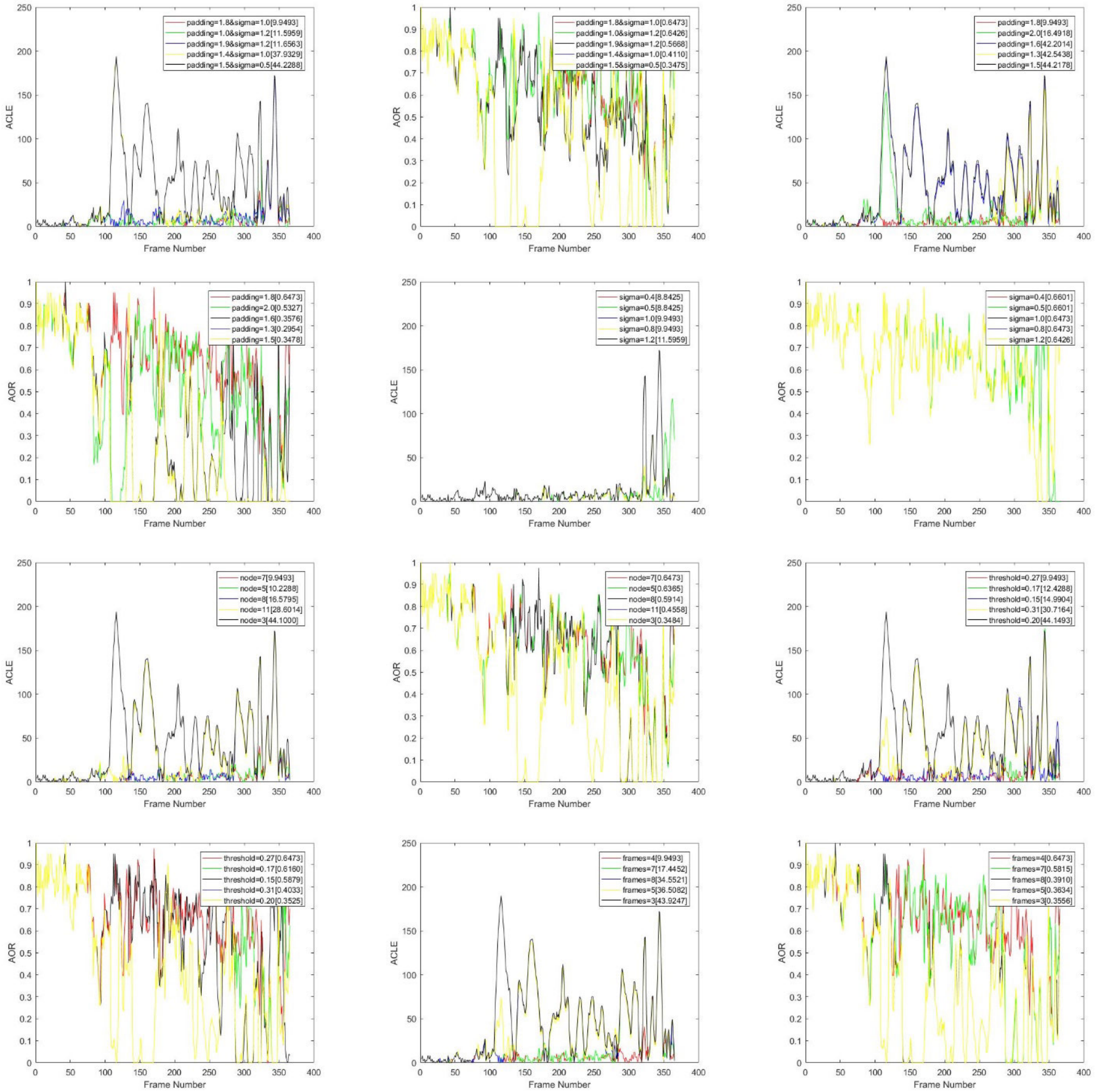


FIGURE 2. ACLE and AOR performance under different parameter settings.

previous target state, resulting in repeated computation that will be no difference from frame-based model updating. After experimental testing as shown in FIGURE 2, it is found that set the frame number in the frame set as 4 would be reasonable to obtain a better performance.

For a fair comparison, we run all trackers using default parameters without manual adjustment.

B. EVALUATION CRITERIA

While testing on the OTB datasets, all the trackers are quantitatively evaluated using two measures: precision plot and success plot.

Precision is defined as the average Euclidean distance between the estimated center location of the target *center_t* and the center of its corresponding ground-truth *center_a*, which can be computed by

$$P = \sqrt{|center\_t - center\_a|^2} \tag{13}$$

The precision plot is defined as the average number of frames per video that are at most 20 pixels away from the ground truth. The success rate measures the IoU (Intersection over Union) between the bounding box of the tracking result r_t and ground truth bounding box r_a, which can be

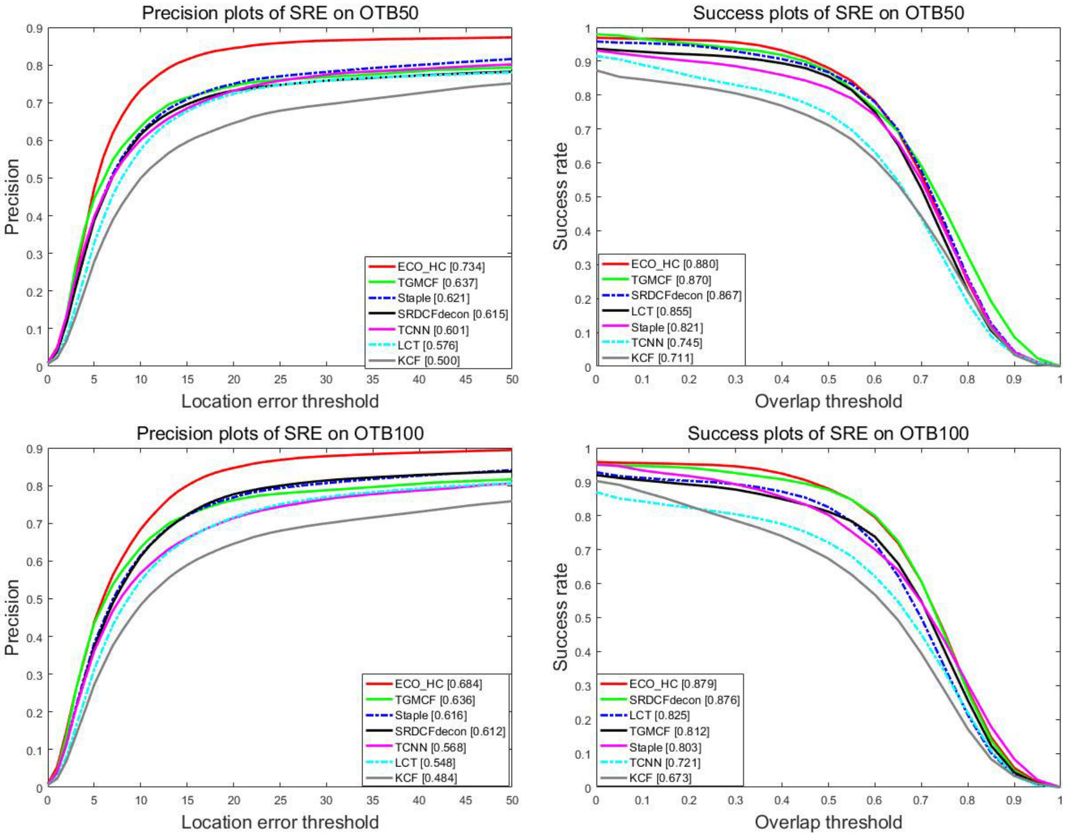


FIGURE 3. Success and precision plot comparisons between our approach and the six SOTA trackers on OTB50 and OTB100, in which our approach is named TGMCF in plots and the AUC values are reported in brackets.

given by

$$S = \frac{r_t \cap r_a}{r_t \cup r_a} \tag{14}$$

where $\cap$ and $\cup$ represent the intersection and union of two regions respectively, and $|\cdot|$ denotes the number of pixels in the region. Specifically, the success plot shows the percentage of bounding boxes whose IoU score is larger than a given threshold. We use the Area under the Curve (AUC) of the success plots to rank the trackers. For a full treatment of these metrics, readers are encouraged to read [27].

While testing on the VOT dataset, all the trackers are also quantitatively evaluated using the average accuracy and failure score. The detailed analysis on accuracy and failure score can be found in [33]. For a predicted object region and its ground truth at frame t, accuracy is defined as same as the success rate in OTB. Average accuracy per sequence is calculated by averaging these accuracy scores over the total time. If a tracker fails, i.e., accuracy score decreases to zero, it will be re-initialized (c.f. [34] for further details).

C. QUANTITATIVE ANALYSIS

FIGURE. 3 compares our approach with six SOTA trackers including ECO_HC [28], SRDCFdecon [30], Staple [29], LCT [14], TCNN [5], and KCF [11] on the OTB50 [31] and OTB100 [27]. It can be observed that our approach achieved an outstanding performance among these trackers.

In detail, we present the success plots and precision plots for SRE (spatial robustness evaluation) on the whole OTB2013 and OTB2015. In the plot of precision rate, our tracker ranks the second, closely follows the most representative tracker ECO_HC, whilst achieves 1.6% higher than the method of Staple in OTB50 and achieves 2% higher than Staple in OTB100. In the plot of success, our tracker ranks the second in OTB50 and the fourth in OTB100, but achieves a great achievement than the baseline tracker KCF and TCNN.

The reason why our tracker's performance in success rate is not as comparable as the precision is due mainly to the baseline tracker KCF used in our method. To this end, KCF's drawbacks still exist in dealing with scale changes. There may be some situations that the tracker has already located the center of target correctly, but with an incorrect scale, a low success rate is produced. Despite of this, compared to the baseline tracker KCF, our tracker gains a great increase in both the precision rate and the success rate. For the precision rate, we achieved 13.7% higher than the baseline tracker on OTB50 and achieved 15.2% higher than on OTB100. Meanwhile, we also gain 15.9% and 13.9% in terms of success rate on OTB50 and OTB100, respectively.

TABLE 1. The OP over 0.5(in %) and the mean OP (in %) results of the trackers on OTB50. The best two results are shown in red and blue fonts, respectively.

	TGMCF	AECF	ECO_HC	KCF	SRDCFdecon	Staple	TCNN	LCT
OP over 0.5	87.0	76.5	88.0	71.1	86.7	82.1	74.5	85.5
Mean OS	63.7	62.5	73.4	50.0	61.5	62.1	60.1	57.6

TABLE 2. Experimental results in VOT2015 [32] for the proposed TGMCF tracker and the seven SOTA trackers.

Methods	TGMCF	UPDT	Ensemble_SW_Obs	ECO_HC	KCF	SRDCFdecon	Staple	TCNN
Acc.	0.55	0.532	0.52	0.54	0.49	0.54	0.54	0.52
Failures	1.29	-	1.97	1.26	2.31	2.26	1.1	1.55

TABLE 3. VOT2015 [32] accuracy results for the proposed TGMCF and the compared trackers. Red, blue and green indicate the first, second and third rankings, respectively.

	label_camera_motion		label_empty		label_illum_change		label_motion_change		label_occlusion		label_size_change		Mean		Weighted mean		Pooled	
	A-Rank	Overlap	A-Rank	Overlap	A-Rank	Overlap	A-Rank	Overlap	A-Rank	Overlap	A-Rank	Overlap	A-Rank	Overlap	A-Rank	Overlap	A-Rank	Overlap
ECO_HC	1	0.58	1	0.57	7	0.4	1	0.47	1	0.47	1	0.42	1.67	0.48	1.67	0.52	1	0.54
KCF	6	0.48	1	0.55	1	0.49	7	0.42	1	0.43	1	0.42	2.5	0.47	2.5	0.48	1	0.49
SRDCFdecon	1	0.56	1	0.58	1	0.5	1	0.49	1	0.43	1	0.47	1	0.51	1	0.53	1	0.55
Staple	1	0.54	1	0.6	1	0.54	1	0.48	1	0.48	1	0.46	1	0.52	1	0.53	1	0.54
TGMCF	1	0.57	1	0.58	1	0.54	1	0.46	1	0.48	1	0.42	1	0.5	1	0.53	1	0.54
TCNN	1	0.58	1	0.59	1	0.41	1	0.52	1	0.51	1	0.47	1	0.51	1	0.55	1	0.56
Ensemble_SW_Obs	1	0.54	1	0.57	1	0.65	1	0.46	1	0.43	1	0.49	1	0.52	1.67	0.52	1	0.53

Table 1 shows the overall performance of trackers by giving the results of the success rate of overlap precision over 0.5 in the sequences (OP over 0.5) and the mean overlap success of all 11 fixed thresholds (Mean OS). Especially, the tracker of AECF [35] is additionally compared due to the lack of source code. As a result, we only compare its performance according to the data provided in its paper in Table 1 rather than Fig. 3. As seen from the experimental results in Table 1, our method performs better than the AECF method proposed in 2019 and can respectively perform the third in OP and the second in OS, closely follows the most representative trackers. However we can achieve a rather improvement than the baseline tracker KCF and TCNN.

Therefore, it can be concluded that the strategy proposed who considers diverse models is effective to the visual tracking problem.

We also provide an attribute based evaluation between our tracker and the six representative trackers on OTB50. Specifically, for performance evaluation, nine most representative tracking challenges are chosen to analyze the tracking performance, which are scale variation, out of view, occlusion, motion blur, illumination, out-of-plane rotation, deformation, fast motion and in-plane rotation. According to the experimental results shown in FIGURE 4, our tracker presents almost the best among the six SOTA trackers, except sometimes only second to ECO_HC. Despite of the significant improvement on the baseline tracker KCF, our tracker can also achieve 5.9% higher than ECO_HC in scale variation, 17.6% higher than TCNN in out of view, 7.6% better than ECO_HC in occlusion, 4.8% better than ECO_HC in motion blur, performs the same as ECO_HC in illumination problem, 1.7% better than Staple secondly to ECO_HC in out-of-plane rotation and 5.7% better than TCNN secondly to ECO_HC in fast motion. This empirically demonstrates the effectiveness of considering multiple KCF models in the tree structure in improving the stability of such trackers against challenging photo-metric and geometric variations.

Apart from this, we also compared our tracker on the VOT dataset with seven SOTA trackers, including UPDT [36], Ensemble_SW_Obs [37], ECO_HC [28], KCF [11], SRDCFdecon [30], Staple [29] and TCNN [5]. Table 2 shows the average results of VOT2015. Among the compared correlation filter based trackers, the trackers of UPDT proposed in 2018 analyses the effect of deep and shallow features on target tracking, using a feature fusion strategy for tracking. The tracker of Ensemble_SW_Obs [37] also introduces a tree structure into the correlation filter based tracking. Different from our TGMCF, only binary tree structure was employed. According to these results, TGMCF has the first ranking in terms of the sum of average accuracy, and the second ranking in terms of the failures score.

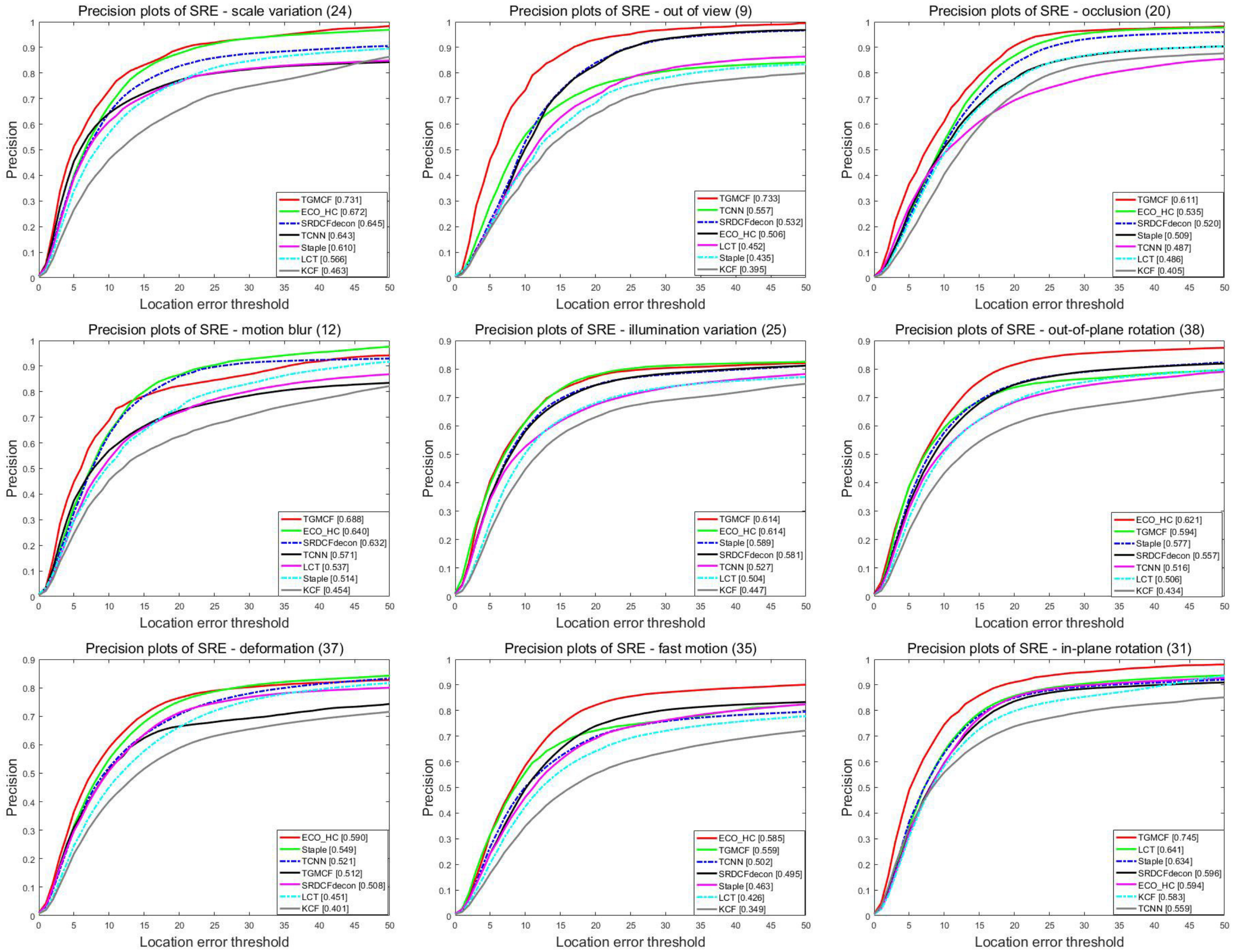


FIGURE 4. Attribute based evaluation. Precision plots compare our method with six SOTA trackers on OTB50. Our method outperformed almost all the trackers especially the baseline method KCF. The precision values are reported in brackets, and the number of videos for each attribute is shown in parenthesis.

Table 3 also compares the trackers in terms of the average accuracy. In detail, the last six columns represent different tracking challenges. Mean and Weighted mean are averaged attributes in an equal or weighted manner, while Pooled corresponds to averaging per-frame results of the super-sequence obtained by concatenating all of the sequences. Among the compared trackers, our TGMCF can perform consistently better among the compared trackers, and shares the first ranking in terms of overlap.

D. QUALITATIVE ANALYSIS

For qualitative comparison, our tracker is also compared with the six SOTA trackers mentioned above on six challenging sequences including Coke, basketball, singer2, soccer, Skating1 and football1. The tracking results are shown in FIGURE. 5 and FIGURE. 6.

In the Coke video sequence, the object experienced several tracking difficulties, such as illumination change, out-of-plane rotation, partial occlusion and complete occlusion. In this paper, we show the experimental results of frame #32, #41, #58, #187, #256 and #267 in FIGURE. 5(a). After the illumination change in frame #58, major of trackers begin to get a drift, however our tracker can still locate the target accurately. At frames #187, #256 and #267, the coke was completely occluded by green leaves, leading to failure feature detection of the coke. All the methods fluctuated greatly, at this frame, where the target position would have to be estimated via prediction. Our method trains a correlation filter template every 4 frames and stores it in the nodes of a tree structure. The reliability of nodes and the correlation between nodes are calculated and compared continuously. For example, in the tracking process, the reliability of nodes corresponding to frames #32, #41 and #58 would be relatively high, while the reliability of nodes corresponding to frames #187, #256 and #267 would be relatively too low to be stored in the tree structure. When occlusion happens, our method would predict the target position by considering the most effective information provided in previous frames. Due to our method is based on KCF, only HOG features are extracted and no boundary effect alleviation is added in the

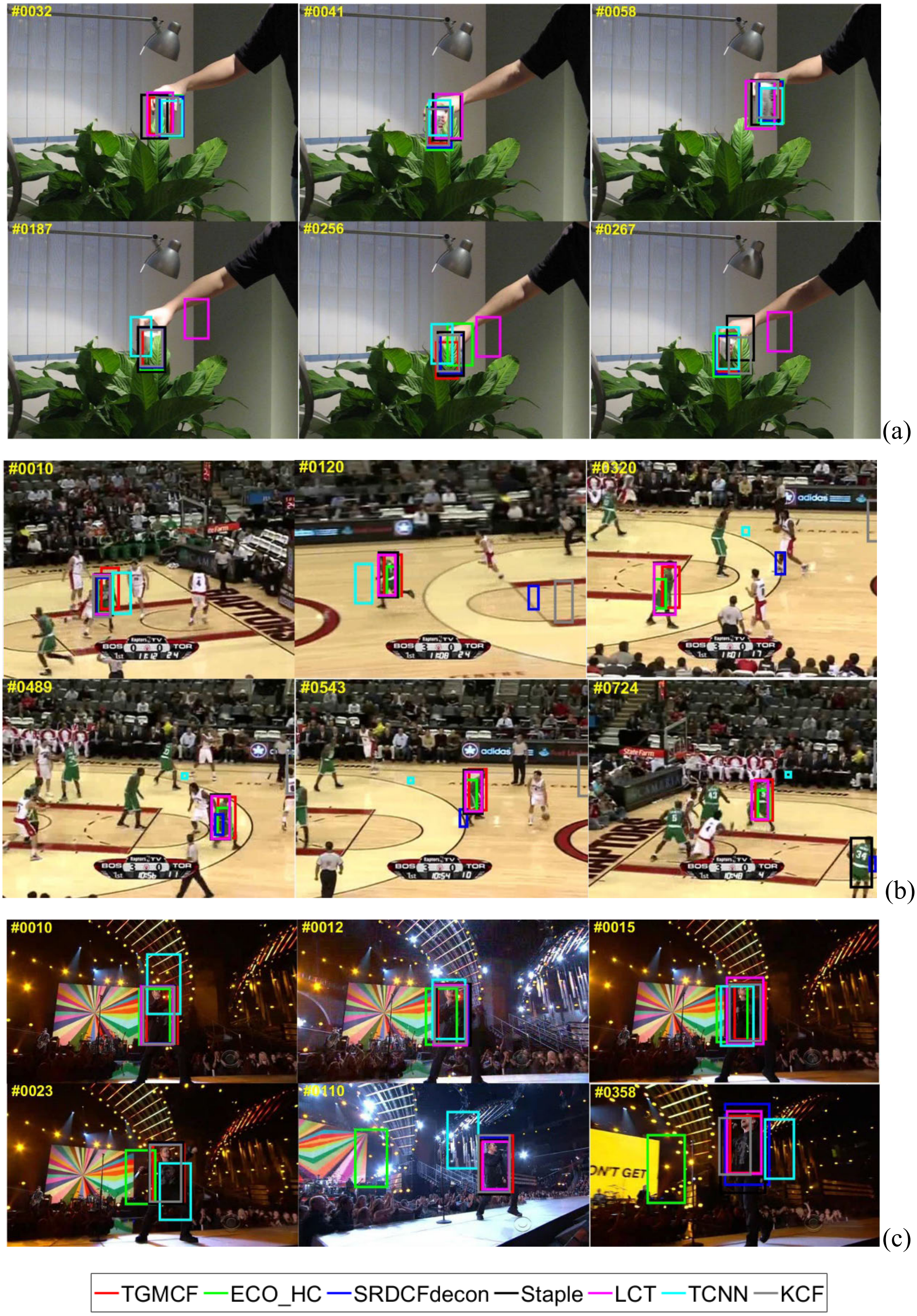


FIGURE 5. Comparison of tracking results as a bounding box in different colors for several tested videos on some key frames, where the sequence names for (a-c) are Coke, Basketball, Singer2.

whole process. However, accurate tracking and prediction can also be achieved, which fully illustrates the effectiveness of the joint correlation filter based on the tree structure.

During tracking the sequence of basketball, the player would run in a high speed with similar appearance person around, which could confused the trackers easily.

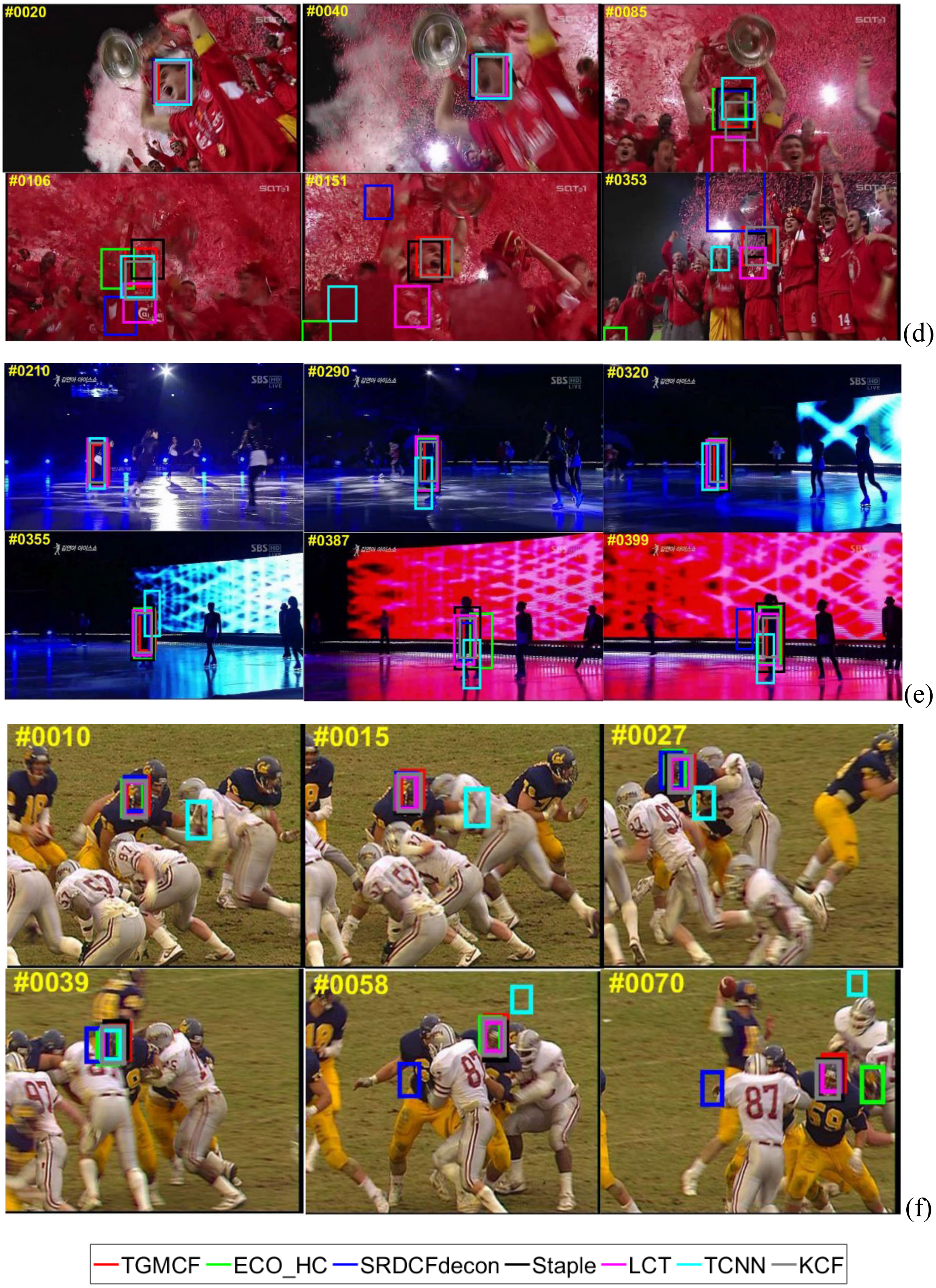


FIGURE 6. Comparison of tracking results as a bounding box in different colors for several tested videos on some key frames, where the sequence names for (d-f) are Soccer, Skating1, and football1.

Additionally, the target would be occluded many times in the whole tracking process, and the trajectory of the target is irregular. These have make the tracking quite challenge hence the high robustness of tracking is required. As can be seen from FIGURE. 5(b), there are many occlusions occurred in the frame of #10 and #489, where most trackers had a

TABLE 4. TGMCF's tracking speed in both OTB50 and OTB100.

OTB50				OTB100			
Sequence name	Tracking speed/fps	Sequence name	Tracking speed/fps	Sequence name	Tracking speed/fps	Sequence name	Tracking speed/fps
Basketball	33.6206	Jump	8.7055	Bird2	22.0705	Football1	84.2024
Biker	132.4418	Jumping	44.5525	BlurCar1	32.3892	Freeman1	51.2491
Bird1	74.0535	Liquor	17.4391	BlurCar3	23.0888	Freeman3	175.6231
BlurBody	13.5242	Matrix	105.1026	BlurCar4	13.5822	Girl2	12.0703
BlurCar2	35.84	MotorRolling	29.3919	Bolt2	42.881	Gym	29.1521
BlurFace	48.6643	Panda	94.7793	Boy	57.7461	Human2	11.2574
BlurOwl	19.5715	RedTeam	108.8816	Car2	34.0114	Human5	69.8516
Bolt	54.7803	Shaking	25.4252	Car24	106.1965	Human7	27.3368
Box	9.9555	Singer2	10.6957	Coke	31.2442	Human7	27.3368
Car1	26.6904	Skating1	22.8492	Coupon	25.0851	Human8	29.4819
Car4	7.9646	Skiing	103.6561	Crossing	108.7658	Jogging [1,2]	38.278
CarDark	74.9532	Soccer	52.1835	Dancer	11.4912	KiteSurf	109.8602
CarScale	51.6962	Surfer	125.9454	Dancer2	13.2614	Lemming	13.0712
ClifBar	49.02	Sylvester	31.6145	David2	71.071	Man	78.1692
Couple	136.6324	Tiger2	16.3553	David3	25.0173	Mhyang	12.0815
Crowds	82.347	Trellis	9.8808	Dog	47.9539	MountainBike	17.6773
David	20.2681	Walking	58.7953	Dog1	29.8821	Rubik	14.3211
Deer	20.4794	Walking2	30.0597	Doll	24.2129	Singer1	13.479
DragonBaby	34.9584	Woman	32.475	FaceOcc1	28.3265	Skater	8.5743
Dudek	10.6206	Girl	26.6723	FaceOcc2	22.1227	Skater2	12.1174
Freeman4	206.2768	Human3	46.2998	Fish	14.439	Subway	86.6561
Ironman	45.3562	Human4	46.2711	FleetFace	8.4054	Suv	14.4555
Human9	27.1477	Human6	68.8602	Vase	26.9098	Tiger1	16.2962
		Trans	13.3101	Twinnings	21.6412	Toy	34.9736
Average: 44.27							

large fluctuation. However, our proposed method can still be able to track the target stably, and the fluctuation is quite small after the first occlusion.

The Singer2 video sequence as shown in FIGURE. 5(c) is a video with more challenges. The target in most frames can be stored into the nodes in the tree structure, but there are differences in terms of reliability and correlation between nodes. Our method stores every stable state into one node, and the six key frames shown in FIGURE. 5(c) can be stored into six different states with different reliability values. For frames #5, #23 and #110 the target is clearer; hence, their reliability will be relatively stronger. Our method decides the location of nodes according to the correlation between nodes, so the correlation filtering templates trained in frames #5, #15, #23 and #338 are possibly on the same branch, and the frames #12 and #110 are more likely to appear on the same branch. Therefore, the filter templates on the branch of frames #5 and #23 will contribution more in making decision on the target in frames #338, rather than the branch of frames #12 and #110. The final target position would be determined by all the information in the tree structure. As shown in the results, only our method, LCT and SRDCFdecon can keep a stable tracking from the beginning to the end, yet the tracking results of our method are more accurate than the other two trackers. The same situation also exists in the sequence of soccer, skating1 and football1, where our tracker can always outperform the methods in peers as clearly illustrated in FIGURE.6(d-f).

In tracking speed, due to selection of the CF tracker as the baseline, the proposed TGMCF can achieve a real-time tracking. Relative to the 1.5fps of TCNN our TGMCF can achieve 44fps tracking speed which greatly improved the tracking speed. The detail of the tracking speed is shown in Table 4.

The success of our tracker can be summed up into two points: One is the stored multiple correlation filter models in a tree structure, which can save the stable and reliable states in the tracking process to provide an effective reference for subsequent tracking more smoothly. The second is the adaptive strategy in managing and updating the entire tree structure according to the reliability of each correlation filter to improve the model, which can effectively avoid the model pollution caused by the sudden change of target during tracking and finally make the tracker more robust.

VI. CONCLUSION

In this paper, we present a novel tree structure based correlation filter for visual tracking to avoid model pollution and enhance tracking performance. The proposed TGMCF approach stores target states in different stages as nodes in the tree structure, and comprehensively consider the multi-modality model according to the reliability of each node during the tracking process. Besides, by adaptive adjusting the nodes in the tree structure and updating the model, the tracker becomes more effective and robust. Finally, our tracker demonstrates a comparative performance against many representative trackers on the OTB50 and OTB100 benchmark datasets. Although there is still a gap between our approach and the deep learning trackers, there is a great improvement compared with the baseline tracker KCF. To illustrate the effectiveness of the strategy proposed, we just take the most classic correlation filter KCF as an example. In the future, more types of correlation filters with better performance would also be applied to the tree structure, and it is believed that better results could be achieved.

REFERENCES

[1] L. Wang, L. Xu, L. Hao, Q. Deng, and M. Q.-H. Meng, "Improving object visual tracking performance by scene occluder estimation for video surveillance," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Ningbo, China, Aug. 2016, pp. 1836–1839.

[2] B. Bayram and G. İnce, "Audio-visual human tracking for active robot perception," in *Proc. 23rd Signal Process. Commun. Appl. Conf. (SIU)*, Malatya, Turkey, May 2015, pp. 1264–1267.

[3] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. S. Torr, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.

[4] 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. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4310–4318.

[5] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNs in a tree structure for visual tracking," 2016, *arXiv:1608.07242*. [Online]. Available: https://arxiv.org/abs/1608.07242

[6] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Antalya, Turkey, Aug. 2017, pp. 1–6.

[7] M. Kristan *et al.*, "The visual object tracking VOT2016 challenge results," in *Proc. ECCV*, Amsterdam, The Netherlands, 2016, pp. 191–217.

[8] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Nottingham, U.K., 2014, pp. 65.1–65.11.

[9] S. Avidan, "Support vector tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1064–1072, Aug. 2004.

[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. (CVPR)*, San Francisco, CA, USA, Jun. 2010, pp. 2544–2550.

[11] 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.

[12] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 3549–3557.

[13] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. ICCV Workshop (ICCVW)*, Santiago, Chile, Dec. 2015, pp. 621–629.

[14] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 5388–5396.

[15] H. Zeng, N. Peng, Z. Yu, Z. Gu, H. Liu, and K. Zhang, "Visual tracking using multi-channel correlation filters," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Singapore, Jul. 2015, pp. 211–214.

[16] Q. Bai, Z. Wu, S. Sclaroff, M. Betke, and C. Monnier, "Randomized ensemble tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 2040–2047.

[17] K. Zhang, L. Zhang, and M.-H. Yang, "Fast compressive tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2002–2015, Oct. 2014.

[18] Z. Yan, F. Yang, J. Wang, Y. Shi, C. Li, and M. Sun, "Face orientation detection in video stream based on Harr-like feature and LQV classifier for civil video surveillance," in *Proc. IET Int. Conf. Smart Sustain. (ICSSC)*, Shanghai, China, 2013, pp. 161–165.

[19] N. He, J. Cao, and L. Song, "Scale space histogram of oriented gradients for human detection," in *Proc. Int. Symp. Inf. Sci. Eng.*, Shanghai, China, Dec. 2008, pp. 167–170.

[20] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 32–39.

[21] X. Mei and H. Ling, "Robust visual tracking using $\ell 1$ minimization," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 1436–1443.

[22] G. Zhu, F. Porikli, and H. Li, "Beyond local search: Tracking objects everywhere with instance-specific proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 943–951.

[23] B. Han and L. Davis, "On-line density-based appearance modeling for object tracking," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, Beijing, China, vol. 2, Oct. 2005, pp. 1492–1499.

[24] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2042–2049.

[25] Y. Wang, Q. Liu, L. Jiang, M. Yin, and S. Wang, "Sequential Monte Carlo-guided ensemble tracking," *PLoS ONE*, vol. 12, no. 4, 2017, Art. no. e0173297.

[26] T. Zhang and Z.-H. Zhou, "Large margin distribution machine," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 313–322.

[27] 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.

[28] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6931–6939.

[29] 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. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1401–1409.

[30] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1430–1438.

[31] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE CVPR*, Portland, OR, USA, Jun. 2013, pp. 2411–2418.

[32] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, ''The visual object tracking VOT2015 challenge results,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1–23.

[33] L. Čehovin, A. Leonardis, and M. Kristan, ''Visual object tracking performance measures revisited,'' *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1261–1274, Mar. 2016.

[34] M. Kristan, ''The visual object tracking VOT2014 challenge results,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2015, pp. 191–217, doi: 10.1007/978-3-319-16181-5_14.

[35] W. Wang, K. Zhang, and M. Lv, ''Robust visual tracking based on adaptive extraction and enhancement of correlation filter,'' *IEEE Access*, vol. 7, pp. 3534–3546, 2018.

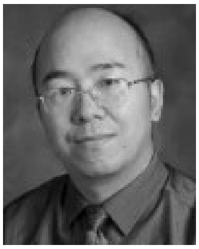
[36] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, ''Unveiling the power of deep tracking,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 483–498.

[37] E. Gundogdu, H. Ozkan, and A. A. Alatan, ''Extending correlation filter-based visual tracking by tree-structured ensemble and spatial windowing,'' *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5270–5283, Nov. 2017.

QIAOYUAN LIU received the bachelor's degree from the Department of Computer Science and Technology, Northeast University, Shenyang, China, in 2014, and the master's degree from the Department of Computer Science and Technology, Northeast Normal University, Changchun, China, where she is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology. Her current research interest includes visual tracking.

WEIWEI LIU received the bachelor's degree from the Department of Computer Science and Technology, Harbin Engineering University. She is currently pursuing the master's degree with the Department of Computer Science and Technology, Northeast Normal University, Changchun, China. Her current research interest includes visual tracking.

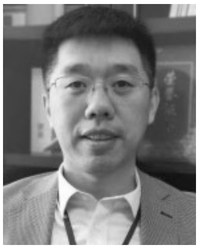
YURU WANG received the Ph.D. degree from the Department of Computer Science and Technology, Harbin Institute of Technology, China, in 2010. Her current research interests include computer vision and pattern recognition.



JINCHANG REN received the Ph.D. degree in electronic imaging and media communication from the University of Bradford, U.K., in 2009. He is currently the Deputy Director of the Hyperspectral Imaging Center, University of Strathclyde. His current research interests include intelligent information processing, visual computing, and multimedia signal processing, especially on content-based image/video analysis, and retrieval.

QIAO DU received the bachelor's degree from the Department of Computer Science and Technology, North China Electric Power University. He is currently pursuing the master's degree with the Department of Computer Science and Technology, Northeast Normal University, Changchun, China. His current research interest includes visual tracking.

YINGHUA LV is currently the Dean of the Humanities College, Northeast Normal University. He is also a Doctoral Supervisor with the Department of Computer Science and Technology, Northeast Normal University.



HAIJIANG SUN received the master's and Ph.D. degrees from the Changchun Institute of Optics and Machinery, Chinese Academy of Sciences. He is the Director of the Research Department of Image Processing Technology, Changchun Institute of Optics and Machinery, Chinese Academy of Sciences. His current research interests include high-speed image processing technology, target automatic recognition, tracking and measurement technology, and optical image enhancement display technology.