IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 30, 2020, accepted May 19, 2020, date of publication May 25, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997072

Robust Visual Object Tracking With Multiple Features and Reliable Re-Detection Scheme

HAIJUN WANG[1,2], WENLAI MA[1,2], SHENGYAN ZHANG[1], GUO CHEN[2], HONGJUAN GE[2], AND YUJIE DU[1]

[1]Aviation Information Technology Research and Development Center, Binzhou University, Binzhou 256600, China
[2]College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Corresponding author: Haijun Wang (whjkyx@163.com)

This work was supported in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J17KA088 and Grant J16LN02, in part by the Natural Science Foundation of Shandong Province under Grant ZR2019PF021, in part by the Key Research and Development Program of Shandong Province under Grant 2016GGX101023, in part by the Dual Service Projects of Binzhou University under Grant BZXYSFW201805, and in part by the Research Fund Project of Binzhou University under Grant 2019ZD03 and Grant BZXYL1803.

ABSTRACT In recent years, correlation filter based trackers have seen widespread success because of their high efficiency and robustness. However, a single feature based tracker cannot deal with complex scenes such as serious occlusion, motion blur and illumination variation. In this paper, we develop a novel tracking method combining color feature, Hog feature and motion feature. The motion feature is estimated between adjacent frames by large displacement optical flow. Besides, in order to cope with boundary effect existing in traditional correlation filter based trackers, an adaptive cosine window is introduced in our method, which can highlight the target region, suppress the background region and enlarge search region. Meanwhile, a novel judge scheme combining Hog correlation response and color response is adopted to evaluate the reliability of tracking result. Finally, inverse sparse representation is presented to locate coarse positions of target in case of tracking failures. Extensive experiments on five famous tracking benchmarks including OTB100, TColor-128, UAVDT, UAV123 and VOT2016 demonstrate our proposed method outperform other sate-of-the-art methods in terms of robustness and accuracy.

INDEX TERMS Visual tracking, correlation filter, motion feature, adaptive cosine window, reverse sparse representation.

I. INTRODUCTION

Visual object tracking, one of the classical and fundamental research topics in computer vision, has long been widely used in traffic monitoring, medical image processing, automatic driving and video surveillance. Although great breakthrough has been made in the past decade [1]–[13], designing a general and robust tracker remains a challenging task, due to many unpredictable factors including illumination change, scale variation, serious occlusion, motion blur, and so on.

Recently, trackers based on correlation filter (CF) have been proposed and obtained promising performance on many challenging benchmarks. The core of CF trackers is to train a discriminative classifier to separate the target from its surrounding background [14]–[16]. Through exploiting Fast Fourier Transform (FFT) on the circulant shifted training samples, the target in a new frame is able to be located at a very low computational complexity. Although CF trackers have achieved promising performance with an extremely high speed, there still exist some factors which severely hamper the tracking performance. First, there are undesired boundary effects in CF trackers due to periodic assumption. Discriminative correlation classifiers are trained with the circulant shifted version of the target and only the detection scores near the center of searching region are accurate. Therefore, only a restricted search area is used to train the correlation filter, which makes CF trackers easily drift to the background in the presence of heavy occlusion and motion blur. Second, historical information of target in video frames is not considered in most CF trackers. The position of target from continuous frames rarely changes greatly which can be used to improve tracking accuracy. Third, there is no re-detection scheme in traditional CF trackers in case of tracking failures. When the

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

target is heavily occluded or is out of view, most CF trackers are not able to locate the position of target again.

In this paper, we address the above-mentioned problems by several aspects. First, in order to solve the boundary effect, we introduce an adaptive cosine window to enlarge search region, which is composed of a traditional cosine window and an adaptive target likelihood map. The target likelihood map is computed for each frame and can estimate the probability of each pixel belonging to the target or the background. Thus, this scheme can highlight the target and suppress the background. Second, we use large displacement optical flow to estimate the motion feature in adjacent frames. Then we combine Hog correlation response, response from color feature and motion feature to obtain a robust response of target. Finally, we use a re-detection module to judge the reliability of tracking result. If the tracking result is regarded as unreliable, we introduce reverse sparse representation to refine the candidates and get coarse locations of target. The flowchart of our proposed method is shown in FIGURE 1.

The main contributions of this paper are summarized as follows,

- We combine traditional fixed cosine window and the target likelihood map of each frame to form an adaptive cosine window, which can effectively cope with boundary effect.
- We introduce large displacement optical flow to predict the motion feature of adjacent frames, which can enhance the robustness of final response of target.
- We adopt a novel judgement scheme to estimate the reliability of tracking results. If the reliability of tracking result is considered to be unreliable, we propose a reverse sparse representation scheme to locate coarse positions of target in case of tracking failures.
- We conduct extensive experiments on OTB100 [17], TColor-128 [18], UAVDT [19], UAV123 [20] and VOT2016 [21] to demonstrate the superiority of our method.

The remainder of this paper is organized as follows. In section II, we give a brief description of related work on visual tracking. Section III describes the detailed introduction of our proposed method. In section IV, we report the experimental results and discussions. Finally, we draw conclusions of this paper in section V.

II. RELATED WORK

A. TRACKERS BASED ON CORRELATION FILTER

CF trackers have obtained promising tracking results owing to the dense sampling and efficient computation in the Fourier domain. Bolme *et al*. [22] firstly applied correlation filter into the field of visual tracking using minimum output sum of square error (MOSSE). MOSSE attracted a huge amount of interest with a tracking speed of more than 600 fps. Henriques *et al*. [23] exploited the circulant structure of training samples in the kernel space (CSK). Later, Henriques *et al*. [24] further extended CSK tracker from one channel feature to multiple features, named kernelized correlation filter (KCF). In order to deal with scale variation, Danelljan *et al*. [25] proposed a discriminative scale space tracker (DSST) to accurately estimate the scale of target. Li and Zhu [26] developed a scale adaptive tracker with multiple features (SAMF) including Hog and color-naming. To alleviate the adverse effect of boundary effect, Danelljan *et al*. [27] presented a spatially regularized discriminative correlation filter (SRDCF) tracker, which can train correlation filter on a huge set of negative training samples. Galoogahi *et al*. [28] developed background aware correlation filters which are trained with real background patches instead of shifted patches. Li *et al*. [29] proposed spatial-temporal regularized correlation filters (STRCF) which can be solved via the alternating direction method of multipliers. Inspired by the successful application of features from convolutional neural networks (CNN) in image classification [30], image segmentation [31] and image denoising [32], [33], Ma *et al*. [34] utilized hierarchical convolutional features instead of handcrafted features in the framework of correlation filter to improve tracking performance. Qi *et al*. [35] developed a novel adaptive weighted method to hedge each weak CNN based tracker into a stronger one. Danelljan *et al*. [36] presented a unified learning framework for down-weighting the corrupted training samples and up-weighting the accurate ones. Furthermore, Danelljan *et al*. [37] developed a generic formulation to learn discriminative convolution operators for visual tracking in the continuous spatial domain. Efficient convolutional operators (ECO) is developed by Danelljan *et al*. [38] to reduce the computation complexity.

B. TRACKERS BASED ON SPARSE REPRESENTATION

Sparse representation is widely used in the field of computer vision, such as face recognition, image classification, object detection and so on. Motivated by the successful application in face recognition, Mei and Ling [39] firstly introduced sparse representation into visual tracking, called ℓ_1 tracker. Each candidate in ℓ_1 tracker is linearly represented by both target template and trivial template. The candidate with the least reconstruction error is chosen as tracking result of current frame. However, the sparse coefficients corresponding to each candidate need to be computed by ℓ_1 minimization, which makes ℓ_1 tracker quite computational expensive. In order to accelerate the tracking speed of ℓ_1 tracker, Bao *et al*. [40] proposed an improved ℓ_1 norm using an efficient gradient descent optimization approach. Xiao *et al*. [41] found that tracking methods with ℓ_2-regularized least square are able to achieve almost the same performance as methods with ℓ_1-regularized least square, but the computational complexity is much lower. To improve the tracking performance of tracker based on sparse representation, Zhong *et al*. [42] builded a sparse collaborative tracking model which takes account of both holistic templates and local representations. An efficient and robust tracking method is developed by Jia *et al*. [43], which considers both structural and partial information of target object. Zhuang *et al*. [44] developed a

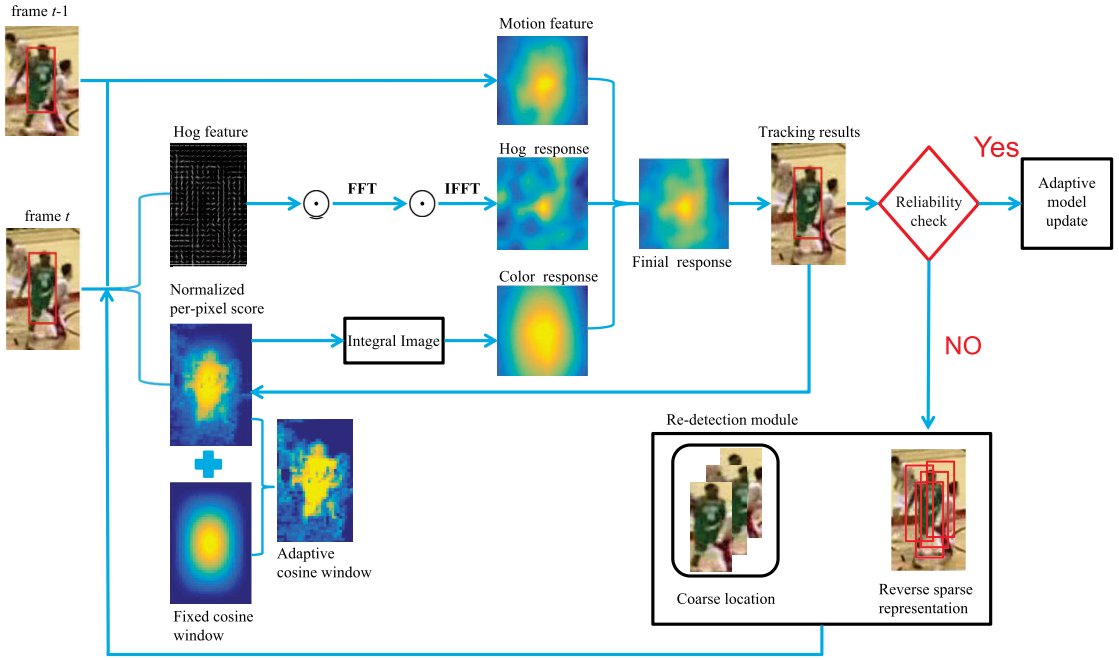


FIGURE 1. The flowchart of our proposed method.

novel reverse sparse representation formulation, which allows multiple templates to be reconstructed simultaneously by the whole candidate set. Wang *et al*. [45] proposed a novel inverse sparse tracer which uses a locally weighted distance metric to replace traditional Euclidean distance metric. Sun *et al*. [46] viewed visual tracking as a two-stage optimization problem which taking account of both temporal discontinuity and continuity information of target appearance.

III. PROPOSED METHODS

A. RESPONSE FROM CORRELATION FILTER

A standard tracker based on correlation filter learns a discriminative classifier to estimate the position of target by finding the maximum value of response map. The classifier $\mathbf{w}$ is trained using an image patch $\mathbf{x}$ with the size of $M \times N$. All the training samples are centered around the target image patch $\mathbf{x}$. Each training sample $\mathbf{x}_{m,n}(m,n) \in \{0,1,\ldots,M-1\} \times \{0,1,\ldots,N-1\}$, corresponding to a Gaussian function label $y(m,n) = e^{-\frac{(m-M/2)^2+(n-N/2)^2}{2\sigma^2}}$, is derived from the circularly shifted version of image patch $\mathbf{x}$ along the horizontal and vertical directions. The classifier $\mathbf{w}$ is derived from the solution of the following minimization function:

$$\mathbf{w}^* = \min_{\mathbf{w}} \sum_{m,n} \|\varphi(\mathbf{x}_{m,n}) \cdot \mathbf{w} - y(m,n)\|^2 + \lambda \|\mathbf{w}\|_2^2 \quad (1)$$

where $\varphi(\cdot)$ means the mapping from the linear feature space to the nonlinear one and $\lambda > 0$ is a regularization parameter. Using a kernel $\kappa(\mathbf{x},\mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$, the classifier $\mathbf{w}$ can be expressed as

$$\mathbf{w} = \sum_{m,n} \alpha(m,n) \kappa(\mathbf{x}_{m,n}, \mathbf{x}) \quad (2)$$

Here α is the dual coefficients of $\mathbf{w}$ and can be learned by

$$\hat{\alpha}^* = \frac{\hat{\mathbf{y}}}{\hat{\mathbf{k}}^{\mathbf{xx}'} + \lambda} \quad (3)$$

where $\hat{\alpha}$ means the fast Fourier Transform (FFT) of α and $\hat{\alpha}^*$ denotes the complex-conjugate of $\hat{\alpha}$. $\mathbf{k}^{\mathbf{xx}'}$ stands for a Gaussian kernel and is defined as

$$\mathbf{k}^{\mathbf{xx}'} = \exp\left(-\frac{1}{\sigma^2}\left(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2\mathcal{F}^{-1}\left(\hat{\mathbf{x}}^* \odot \hat{\mathbf{x}}'\right)\right)\right) \quad (4)$$

where $\mathcal{F}^{-1}$ means the inverse FFT transform and $\odot$ denotes element-wise product. Given an new image patch $\mathbf{z}$ in the next frame, the correlation response map f_h is calculated by

$$f_h = \mathcal{F}^{-1}\left(\left(\hat{\mathbf{k}}^{\mathbf{xz}}\right)^* \odot \hat{\alpha}\right) \quad (5)$$

The tracking result in new frame can be located by searching for the maximum value of correlation response map f_h using Hog feature. In order to improve tracking performance, the model is online updated by the following formulation,

$$\hat{\mathbf{x}}_t = (1-\eta_1)\hat{\mathbf{x}}_{t-1} + \eta_1 \hat{\mathbf{x}}'_t \quad (6)$$

$$\hat{\boldsymbol{\alpha}}_t = (1-\eta_1)\hat{\boldsymbol{\alpha}}_{t-1} + \eta_1 \hat{\boldsymbol{\alpha}}'_t \quad (7)$$

where η_1 means the learning rate and the subscript t is the index of current tracking frame.

B. RESPONSE FROM MOTION ESTIMATION

Optical flow scheme, estimating the motion feature of objects from consecutive frames, is widely applied in computer vision, such as image understanding, image analysis and image registration. Motion estimation, in the form of optical flow, can be used to estimate the motion information of

target across frames. In this section, we use large displace optical flow estimation scheme [47] to estimate the motion information of target in consecutive frames to promote the tracking performance.

Let $\mathbf{I}_t, \mathbf{I}_{t+1}$ be the t-th and the $(t+1)$-th frame, $\mathbf{s} : (s_x, s_y)$ is the location in a rectangular image domain and $\mathbf{u} := (u, v)^{\mathrm{T}}$ be the searched optical flow between the t-th frame and the $(t+1)$-th frame. Then, the optical flow field can be calculated by minimizing the following energy functional:

$$E(\mathbf{u}) = E_{color}(\mathbf{u}) + \gamma E_{gradient}(\mathbf{u}) + \zeta E_{smooth}(\mathbf{u}) + \beta E_{match}(\mathbf{u}, \mathbf{u}_t) + E_{desc}(\mathbf{u}_t) \quad (8)$$

where γ, ζ and β are tuning parameters. $E_{color}(\mathbf{u})$ encodes color constancy, $E_{gradient}(\mathbf{u})$ encodes gradient constancy, $E_{smooth}(\mathbf{u})$ encodes robust smoothness constraint, $E_{match}(\mathbf{u}, \mathbf{u}_1)$ and $E_{desc}(\mathbf{u}_1)$ bias the displacement field.

$E_{color}(\mathbf{u})$ assumes that the color value of a pixel is not changed over time by the displacement. The formulation of $E_{color}(\mathbf{u})$ is expressed by

$$E_{color}(\mathbf{u}) = \int_{\Omega} \Psi\left(|\mathbf{I}_{t+1}(\mathbf{s}+\mathbf{u}(\mathbf{s})) - \mathbf{I}_t(\mathbf{s})|^2\right) d\mathbf{s} \quad (9)$$

where $\Psi(s^2) = \sqrt{s^2+\epsilon^2}$, $\epsilon = 0.001$.

$E_{gradient}(\mathbf{u})$ allows slight changes in the color value and can decide the displacement vector by a rule that is invariant under color value changes. The expression of $E_{gradient}(\mathbf{u})$ is given by

$$E_{gradient}(\mathbf{u}) = \int_{\Omega} \Psi\left(|\nabla\mathbf{I}_{t+1}(\mathbf{s}+\mathbf{u}(\mathbf{s})) - \nabla\mathbf{I}_t(\mathbf{s})|^2\right) d\mathbf{s} \quad (10)$$

where $\nabla = (\partial_{s_x}, \partial_{s_y})^{\mathrm{T}}$ stands for the spatial gradient.

$E_{smooth}(\mathbf{u})$ takes account of interaction between neighbouring pixels and introduces the smoothness of the flow field. The formulation is described as

$$E_{smooth}(\mathbf{u}) = \int_{\Omega} \Psi\left(|\nabla u(\mathbf{s})|^2 + |\nabla v(\mathbf{s})|^2\right) d\mathbf{s} \quad (11)$$

To enforce the smooth flow, the term $E_{match}(\mathbf{u})$ is integrated from descriptor matching into the variational formulation, which is described by

$$E_{match}(\mathbf{u}) = \int_{\Omega} \delta(\mathbf{s})\rho(\mathbf{s})\Psi\left(|\mathbf{u}(\mathbf{s}) - \mathbf{u}_t(\mathbf{s})|^2\right) d\mathbf{s} \quad (12)$$

where $\delta(x)$ is a delta function indicating. $\delta(x)$ is 1 if a descriptor match is available in location $\mathbf{s}$. $\rho(x)$ describes the confidence of match.

Assuming the descriptors have been matched, the matching task can be formulated by $E_{desc}(\mathbf{u}_t)$,

$$E_{desc}(\mathbf{u}_t) = \int_{\Omega} \delta(\mathbf{s})|f_{t+1}(\mathbf{s}+\mathbf{u}_t(\mathbf{s})) - f_t(\mathbf{s})|^2 d\mathbf{s} \quad (13)$$

where $f_t(\mathbf{s})$ and $f_{t+1}(\mathbf{s})$ mean the field of feature vectors in frame t and frame $t+1$, respectively.

C. RESPONSE FROM COLOR HISTOGRAM

To effectively cope with shape deformation, a color histogram model is introduced in this section. Including the correct position as a positive example, the color histogram score can be learnt from a huge set of rectangular image patches $\mathbf{x}$ extracted from each frame. Then the histogram weight vector $\boldsymbol{\beta}$ should be obtained by solving the following ridge regression problem,

$$\boldsymbol{\beta}^* = \min_{\boldsymbol{\beta}} \sum_{(\mathbf{x},\ell)\in\mathbf{W}} \left(\boldsymbol{\beta}^{\mathrm{T}}\left[\sum_{\tau\in\mathcal{H}} \psi_{\mathbf{x}}(\tau)\right] - \ell\right)^2 + \varrho\|\boldsymbol{\beta}\|^2 \quad (14)$$

where $\psi_{\mathbf{x}}(\tau)$ denotes the feature pixels of image patch $\mathbf{x}$ in finite region $\mathcal{H}$, $\mathbf{W}$ represents a set of pairs $(\mathbf{x}, \ell)$ and ℓ denotes the labels of image patch $\mathbf{x}$. Then, histogram score can be regarded as an average vote and can be calculated by

$$\min_{\boldsymbol{\beta}} \frac{1}{|\mathcal{O}|}\sum_{\tau\in\mathcal{O}}\left(\boldsymbol{\beta}^{\mathrm{T}}\psi[\tau] - 1\right)^2 + \frac{1}{|\mathcal{B}|}\sum_{\tau\in\mathcal{B}}\left(\boldsymbol{\beta}^{\mathrm{T}}\psi[\tau]\right)^2 \quad (15)$$

where $\mathcal{O}$ denotes the object region and $\mathcal{B}$ means the background region. By introducing the one-hot assumption, the above objective can be decomposed into the following independent terms

$$\sum_{j=1}^{\mathcal{M}}\left[\frac{\mathcal{N}^j(\mathcal{O})}{|\mathcal{O}|}\cdot(\boldsymbol{\beta}^j - 1)^2 + \frac{\mathcal{N}^j(\mathcal{B})}{|\mathcal{B}|}\cdot(\boldsymbol{\beta}^j)^2\right] \quad (16)$$

here $\mathcal{N}^j(\mathcal{O})$ denotes the number of pixels belonging to the region $\mathcal{O}$ for which feature $j \neq 0$. Then, the solution of above ridge regression problem is

$$\begin{aligned}\boldsymbol{\beta}^j &= \frac{p^j(\mathcal{O})}{p^j(\mathcal{O}) + p^j(\mathcal{B}) + \varrho} \\ &= \frac{\frac{\mathcal{N}^j(\mathcal{O})}{|\mathcal{O}|}}{\frac{\mathcal{N}^j(\mathcal{O})}{|\mathcal{O}|} + \frac{\mathcal{N}^j(\mathcal{B})}{|\mathcal{B}|} + \varrho}\end{aligned} \quad (17)$$

After getting the histogram weight vector, the response from color histogram of an image patch $\mathbf{x}$ can be achieved by an integral image. For a new image, the color histogram over the target area $\mathcal{O}$ and background area $\mathcal{B}$ are recomputed and linearly updated as follows,

$$\begin{aligned}p_t(\mathcal{O}) &= (1-\eta_2)p_{t-1}(\mathcal{O}) + \eta_2 p'_t(\mathcal{O}) \\ p_t(\mathcal{B}) &= (1-\eta_2)p_{t-1}(\mathcal{B}) + \eta_2 p'_t(\mathcal{B})\end{aligned} \quad (18)$$

where $p_t(\mathcal{O})$ and $p_t(\mathcal{B})$ are the vectors of $p^j_t(\mathcal{O})$ and $p^j_t(\mathcal{B})$ for $j = 1, 2, \ldots, \mathcal{M}$, respectively.

D. ADAPTIVE COSINE WINDOW

Recently, CF trackers have been reported efficient and excellent tracking performance. However, due to the underlying boundary effect produced by the periodic assumption, the detection scores of CF trackers are only accurate around the center of target. Thus, boundary effect easily leads to a restricted searching area and hampers the tracking performance. In order to alleviate the adverse effect of boundary

effect, a fixed cosine window $\mathcal{C}$ is introduced in the traditional CF trackers. Though the fixed cosine window $\mathcal{C}$ can suppress the some contamination of background region, it also shrinks the searching area when finding the true position of target.

In this section, we proposed an adaptive cosine window to overcome the harmful effect of boundary effect, which can highlight target region and suppress background region better than traditional cosine window. The formulation of adaptive cosine window $\mathcal{W}_{adap}$ is given by

$$\mathcal{W}_{adap} = (1 - \eta_3)\mathcal{C} + \eta_3\mathcal{W} \tag{19}$$

where $\mathcal{W}$ is target likelihood map and is computed by equation (17) in section C. The target likelihood map is able to effectively distinguish target region and background region.

E. TARGET LOCATION

In this paper, we use the combination of correlation response of Hog feature f_h, the response from color histogram f_c and the motion map f_m to locate the position of target, which can enhance the robustness of our method. The final response is a weighted linear combination of f_h, f_c and f_m,

$$f^{(t)} = \zeta_1 \cdot f_h^{(t)} + \zeta_2 \cdot f_c^{(t)} + \zeta_3 \cdot f_m^{(t)} \tag{20}$$

where the superscript t is the frame index, ζ_1, ζ_2 and ζ_3 are weighted paramters. The position of the t-th frame can be searched by finding the maximum value of the final response $f^{(t)}$.

F. RE-DETECTION MODULE

In this section, we first check the reliability of tracking results using correlation response of Hog feature and response of color histogram. Then, if the current tracking result is considered to be unreliable, we will launch the re-detection module to refine the target location.

For the response map of Hog feature, we define $\mathcal{S}_h^i = \frac{max(f_h^i) - \mu_h^i}{\sigma_h^i}$ be the i-th peak-to-sidelobe ratio (PSR). f_h^i denotes the i-th correlation response of Hog feature, μ_h^i and σ_h^i are the i-th mean and standard deviation of f_h^i, respectively. We also define the PSR ensemble pool $\mathcal{C}_h = \{\mathcal{S}_h^2, \mathcal{S}_h^3, \ldots, \mathcal{S}_h^i\}$ with its mean value $\mathcal{M}_h$. For the response map from color histogram, we define $\mathcal{S}_c^i = \frac{\sum_\tau \boldsymbol{\beta}^{\mathrm{T}}\psi_i(\tau)}{\sum_\tau \boldsymbol{\beta}^{\mathrm{T}}\psi_1(\tau)}$ be the color score of i-th frame. We also define the color score ensemble pool $\mathcal{C}_c = \{1, \mathcal{S}_c^2, \mathcal{S}_c^3, \ldots, \mathcal{S}_c^i\}$ with its mean value $\mathcal{M}_c$. If $\mathcal{S}_h^i < o_h \cdot \mathcal{M}_h$ or $\mathcal{S}_c^i < o_c \cdot \mathcal{M}_c$, we consider tracking result of the i-th frame is unreliable and we will launch the re-detection module. At the same time, $\mathcal{S}_h^i$ or $\mathcal{S}_c^i$ will not be put in the ensemble pool $\mathcal{C}_h$ and $\mathcal{C}_c$, respectively. o_h and o_c are constant parameters for $\mathcal{M}_h$ and $\mathcal{M}_c$, respectively.

If $\mathcal{S}_h^i$ or $\mathcal{S}_c^i$ are discarded, we first coarsely locate the position of target based on multitask reverse sparse representation scheme. Then we use CF tracker with multiple features mentioned in our above paper to refine the tracking results.

In order to obtain reliable candidates, a discriminative reverse sparse representation based method is adopted to determine the rough scope efficiently. We construct the positive template sets $\mathbf{T}_{pos} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p]$ around the object within a small circular area and the negative template sets $\mathbf{T}_{neg} = [\mathbf{t}_{p+1}, \mathbf{t}_{p+2}, \ldots, \mathbf{t}_{p+n}]$ far away from the object within an annular region. p and n are the number of positive and negative templates, respectively. If the current tracking result is considered to be unreliable, each template is represented by the candidate set $\mathbf{Y}$ with coefficients $\mathbf{c}$, which can be computed by Eq.(21).

$$\begin{cases} \arg\min_{\mathbf{c}_1} \|\mathbf{t}_1 - \mathbf{Y}\mathbf{c}_1\|_2^2 + \tau\|\mathbf{c}_1\|_1, \\ \qquad \cdots\cdots \\ \arg\min_{\mathbf{c}_p} \|\mathbf{t}_p - \mathbf{Y}\mathbf{c}_p\|_2^2 + \tau\|\mathbf{c}_p\|_1, \\ \arg\min_{\mathbf{c}_{p+1}} \|\mathbf{t}_{p+1} - \mathbf{Y}\mathbf{c}_{p+1}\|_2^2 + \tau\|\mathbf{c}_{p+1}\|_1, \\ \qquad \cdots\cdots \\ \arg\min_{\mathbf{c}_{p+n}} \|\mathbf{t}_{p+n} - \mathbf{Y}\mathbf{c}_{p+n}\|_2^2 + \tau\|\mathbf{c}_{p+n}\|_1, \end{cases} \tag{21}$$

where $\mathbf{c}_i$ denotes the nonnegative sparse coefficients of the i-th template and reflects the similarity between candidate and the corresponding template. We construct similarity map matrix $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{p+n}]$, which can be computed as a whole. Then, Eq.(21) is reformulated as the following equation.

$$\arg\min_{\mathbf{C}} \|\mathbf{T} - \mathbf{Y}\mathbf{C}\|_2^2 + \tau \sum_i \|\mathbf{c}_i\|_1 + \frac{\delta}{2} \sum_{ij} \|\mathbf{c}_i - \mathbf{c}_j\|^2 B_{ij} \tag{22}$$

where $\frac{\delta}{2}\sum_{ij}\|\mathbf{c}_i - \mathbf{c}_j\|^2 B_{ij}$ is the customized Laplacian regularization term. δ is the regularization parameter and B is a binary matrix.

For the i-th candidate, by introducing the weighted discriminative sparse similarity map and the additive pooling scheme [44], the reliability score R_i is calculated by

$$R_i = s_{i-pos} - s_{i-neg} \tag{23}$$

where s_{i-pos} and s_{i-neg} represent what extent can the i-th candidate be related to the positive and negative template sets. The i-th candidate with the higher value of R_i is more possible to be the target. In order to reduce the computation cost, we discard 90% of the candidates through the reliability score. The rest candidates will be put in the correlation filter framework using multiple features with adaptive cosine window to refine the tracking result. The position of target will be located by searching for the maximum value of correlation response map from all the remaining candidates.

G. UPDATE SCHEME

Inspired by [48], we use the same adaptive update scheme to adapt to various changes. For the correlation filter based tracker, the update parameter η_1 is set by

$$\eta_1 = \begin{cases} Q, & \textit{if tracking result is reliable} \\ \upsilon_1 \left(\frac{S_h}{\mathcal{M}_h}\right)^{\upsilon_2} Q, & \textit{otherwise} \end{cases} \tag{24}$$

TABLE 1. Parameter setting in our paper.

Parameter	Value
Hog cell size	4
η_1	0.01
η_2	0.02
η_3	0.5
ζ_1	0.7
ζ_2	0.2999
ζ_3	0.0001
Q	0.01
P	0.02
υ_1	0.8
υ_2	3

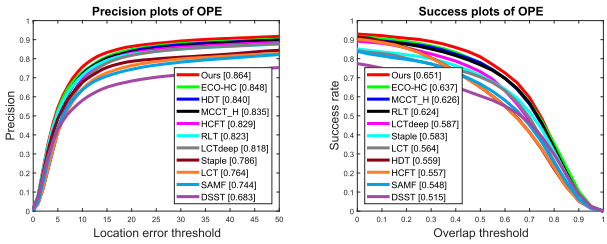


FIGURE 2. Precision plots and success plots of OPE of our method against other 10 state-of-the-art methods on OTB100.

where Q, υ_1 and υ_2 are constants. For color histogram model, the update parameter η_2 is set by

$$\eta_2 = \begin{cases} P, & if\ tracking\ result\ is\ reliable \\ 0, & otherwise \end{cases} \tag{25}$$

where P is a constant.

As for the positive template sets in coarse location process, if the tracking result is reliable, we update the positive template sets according to a threshold θ. We define the similarity vector $d = [d_1, d_2, \ldots, d_p]$. d_i describe the similarity between the i-th positive template and the current tracking result. For the i-th positive template, if max $d_i > \theta$, ($i = 1, 2, \ldots$,m), we replace it with the tracking result. Otherwise, we keep the i-th positive template unchange. For the negative template set, we sample negative templates around the tracking result in the last frame.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

In this section, we demonstrate a comprehensive evaluation of our proposed method. We conduct extensive experiments on five popular benchmarks, i.e., OTB100 [17], TColor-128 [18], UAVDT [19], UAV123 [20], VOT2016 [21].

1) IMPLEMENTATION DETAILS

Our method is implemented on Matlab 2017 and is conducted on a computer with an Inter(R) Xeon(R) 2.4GHz CPU and 128G RAM. Table 1 gives some parameters used in our method, which are fixed for all the experiments.

2) EVALUATION METHODOLOGY

In order to assess different tracking methods fairly, two evaluation metrics: precision rate and success rate, are introduced in this paper. The precision plot is defined as the percentage of tracking frames whose center location error is less than 20 pixels. Here, the center location error is the difference between the estimated positions and ground truth. The success plot denotes the overlap ratio between the predicted bounding box and ground truth. The overlap ratio is defined as $S = \frac{Area(R_t \bigcap R_g)}{Area(R_t \bigcup R_g)}$. Here, R_t means the predicted bounding box and R_g denotes the ground truth. $\bigcap$ and $\bigcup$ are the union and intersection operators, respectively.

B. RESULTS ON OTB100 DATASET

OTB-100 dataset consists of 100 challenging video sequences annotated with 11 different attributes, including low resolution (LR), background clutter (BC), out of view (OV), in-plane rotation (IPR), fast motion (FM), motion blur (MB), deformation (DE), occlusion (OC), scale variation (SV), out-of-plane rotation (OPR), illumination variation (IV).

1) QUANTITATIVE EVALUATION

We compare our method with other 10 state-of-the-art trackers including HCFT [49], LCTdeep [50], HDT [51], MCCT_H [52], RLT [48], Staple [53], LCT [54], SAMF [26], DSST [25], ECO-HC [38]. FIGURE 2 demonstrates the precision plots and success plots of one pass evaluation (OPE) on OTB100. It is obvious that our method achieves the best tracking performance in both precision and success rates. Compared with the baseline RLT tracker, our tracker obtains a concerning improvement (4.1% in precision rate and 2.7% in success rate). HCFT, LCTdeep, HDT using deep features can greatly improve tracking performance compared with trackers using handcrafted features. However, due to the time-consuming feature extraction process, these trackers can only run with a speed of about 1fps at the CPU platform, which can not meet the real-time requirement. Although LCTdeep and LCT trackers have re-detection module, they use a fixed threshold to assess the reliability of tracking result, which can not obtain promising results for all the challenging sequences. Our method using traditional handcrafted features, with the help of novel re-detection module, achieves the precision score of 86.4% and the success score of 65.1%.

In order to further demonstrate the superiority of our proposed method, we report tracking results for eleven challenging attributes in overlap rates in FIGURE 3. It is obvious that our method achieves the best tracking performance in almost all the eleven attributes except for low resolution (ranks second), fast motion (ranks second) and scale variation (ranks second). Besides, the tracking performance of our proposed method has been greatly improved in occlusion (63.2% vs 54.8%) and out-of-plane rotation (61.9% vs 53.8%) when compared with Staple tracker. This improvement is mainly due to the adaptive cosine window scheme, which can enlarge the searching region to alleviate the adverse effect of

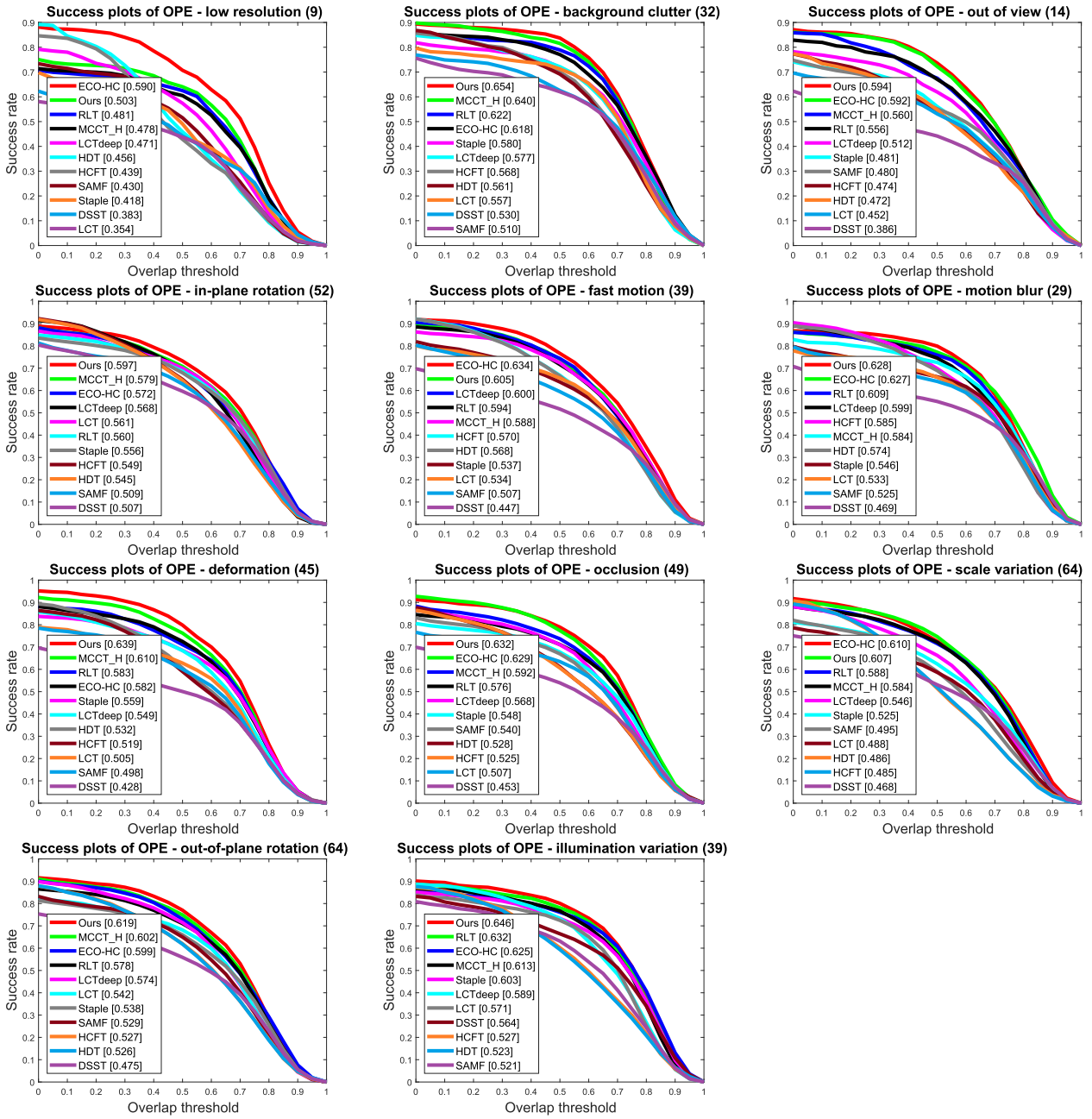


FIGURE 3. Success plots of OPE with different attributes on OTB100.

boundary effect. Since using multiple features, the tracking performance of our method under low resolution has been improved by 2.2% as compared with the baseline RLT tracker. Besides, our method is able to outperform methods using deep features in all the attributes, such as HCFT, HDT and LCTdeep.

2) QUALITATIVE EVALUATION

In order to explicitly demonstrate the comparison results between our method and other 10 state-of-the-art trackers, we visually show the bounding boxes of 11 trackers on several key frames of 10 representative video sequences in FIGURE 4. For both *Girl*2 and *Lemming* sequences, the major challenge is serious occlusion. Since HCFT, HDT, SAMF, Staple, ECO-HC and DSST do not consider the re-detection scheme, it can be seen that these methods are not able to cope with serious occlusion. Although LCTdeep and LCT have failure recovering mechanism, they can not deal with full occlusion and background clutter in *Lemming* sequence. The targets in *Singer*1 and *Singer*2 video sequences have drastic illumination variation, while scale variation is included in the *Singer*1 sequence. It can be

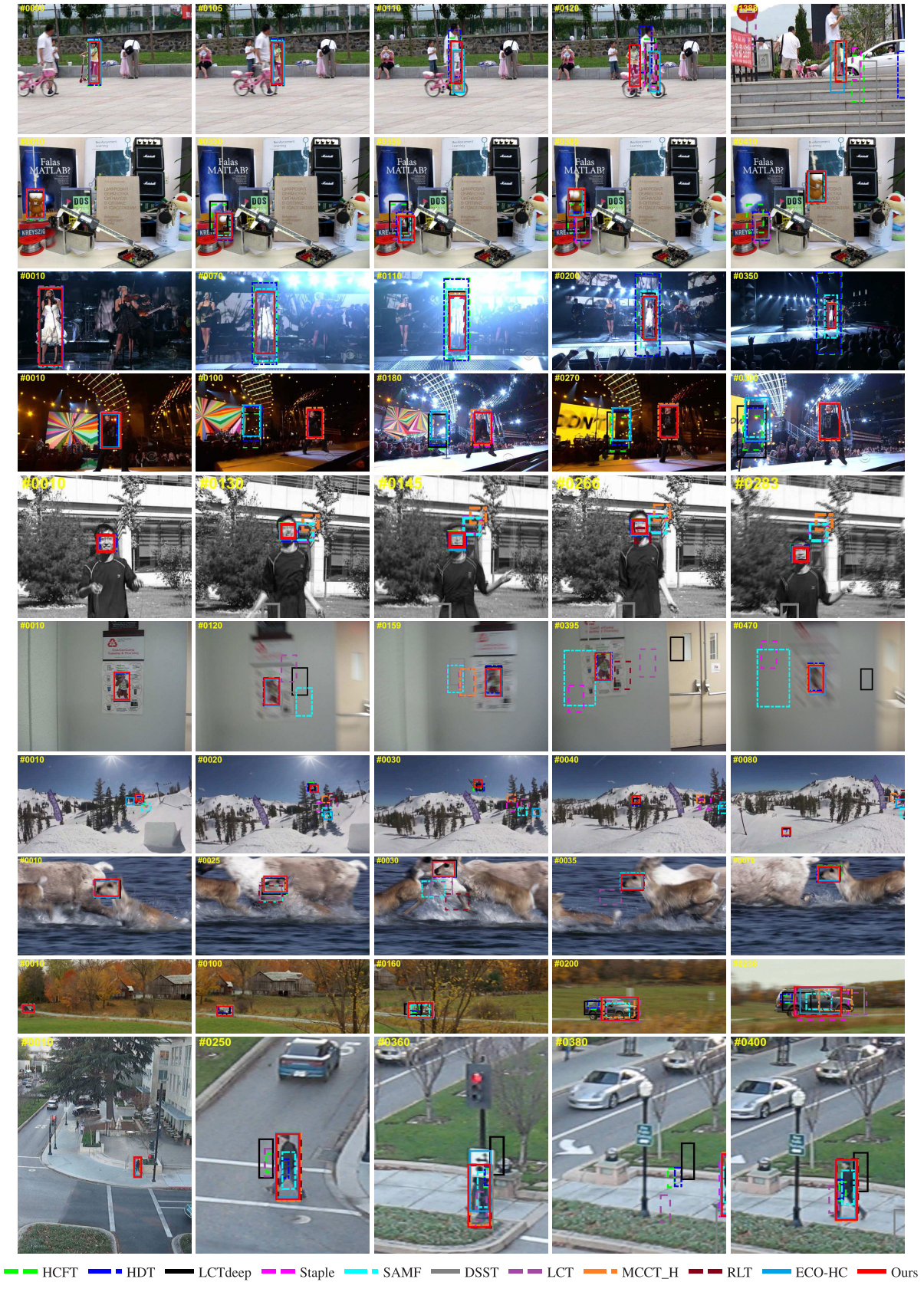


FIGURE 4. Tracking results obtained by HCFT, HDT, LCTdeep, Staple, SAMF, DSST, LCT, MCCT_H, RLT, ECO-HC and Our method on ten sequences (From up to bottom: *Girl*2, *Lemming*, *Singer*1, *Singer*2, *Jumping*, *BlurOwl*, *Skiing*, *Deer*, *CarScale* and *Human*6).

observed from the *Singer*2 sequence that LCTdeep, HDT, ECO-HC and SAMF are sensitive to illumination variation and drift in frame 100. Our proposed method has strong robustness to both illumination variation and scale variation. The sequences in *Juming* and *BlurOwl* are low in quality because of dramatic motion blur. Consequently, MCCT_H, SAMF, RLT, DSST, Staple and LCT have a strong tendency to lose the target. Our method with adaptive cosine window can highlight the target region, suppress the background region and locate the position of target precisely in the whole tracking process. The targets in both the *Skiing* and *Deer* sequences undergo large displacement due to fast motion. SAMF, MCCT_H, RLT, LCT, ECO-HC and Staple learn a great deal of background features and drift to the surrounding background in the process of tracking. Our method considers the motion feature between adjacent frames and can cope with fast motion easily. The targets in *CarScale* and *Human*6 sequences have significant scale variation during tracking, and occasionally have serious occlusion. The bounding boxes of HCFT and HDT keep the same during tracking, as they do not have the scale estimation component. Our method use the same scale estimation scheme as RLT and is able to predict precise scale of target all the time.

3) COMPARISON OF DIFFERENT METHODS IN SPEED ON OTB100

Tracking speed is very important for industrial application of visual tracking algorithm. It is difficult to apply trackers with a slow speed into industrial products. This section demonstrates the comparison of different methods in speed on OTB100. In order to show the comparison of different visual tracking methods validly and fairly, all the three methods with deep features (including HCFT, HDT and LCTdeep) in Table 2 are conducted on a PC equipped with an Inter Xeon CPU E5-2640 v4 with 128G RAM and a single NVIDIA TITAN Xp. The other eight methods with handcrafted features in Table 2 are conducted on the same PC platform without GPU. From Table 2, it can be observed that our tracker in a CPU platform is able to get better tracking results than trackers with deep features in a GPU platform at almost the same speed.

C. RESULTS ON TColor-128 DATASET

To further testify the validity of our proposed tracker, we conduct a comprehensive experiment on TColor-128 dateset. TColor-128 dataset consists of 128 color sequences, which have the same 11 challenging attributes as OTB100 dataset. We compare our method with other 11 state-of-the-art trackers including RLT [48], SAMF_AT [55], SRDCF [27], CoKCF [56], HDT [51], HCFT [49], SAMF [26], LCT [54], CFNET [57], DSST [25] and ECO-HC [38]. FIGURE 5 shows the comparison results with 11 trackers in terms of precision plots and success plots of OPE. It can be seen that compared with trackers using deep features such as CoKCF, HDT, HCFT and CFNET, our method achieves 7.3%, 8.7%, 8.2%, 23.8% improvement respectively in precision plot and

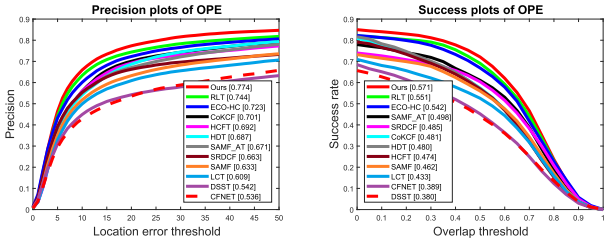


FIGURE 5. Precision plots and success plots of OPE of our proposed method against other state-of-the-art methods on TColor-128.

9%, 9.1%, 9.7%, 18.2% improvement respectively in success plot. Our tracker outperforms the other 11 trackers and ranks the first place in both precision rate and success rate due to the complementary handcrafted features. As using the motion feature of two continuous frames, compared with the baseline RLT tracker, the performance of our tracker has been improved by a margin of 3% in terms of precision plot and 2% in terms of success plot, respectively. Table 3 and 4 demonstrates the precision rate and success rate of 12 trackers with 11 challenging attributes on TColor-128 dataset. The best, second best and third best tracking results are represented in red, blue and green, respectively. It can be observed that our method obtains the best tracking performance in terms of precision plot except for motion blur (ranks second) and low resolution (ranks second). Besides, our method gets the first place in almost all the 11 challenging attributes in terms of success rate except for low resolution (ranks second) and out of view (ranks second). Thus, we can conclude that our method is effective for all the 11 attributes compared with 11 trackers mentioned in this section.

D. RESULTS ON UAVDT DATASET

UAVDT dataset contains 50 challenging video sequences captured from UAV, which are fully annotated with bounding boxes and focus on complex scenarios. These sequences covers 9 complex challenging attributes including background clutter (BC), camera motion (CM), object motion (OM), small object (SO), illumination variation (IV), object blur (OB), scale variation (SV), long-term tracking (LT) and large occlusion (LO). FIGURE 6 illustrates precision plots and success plots of OPE of our proposed method against other 14 state-of-the-art tracking methods including SRDCF [27], PTAV [58], MCPF [59], C-COT [37], FCNT [60], STCT [61], CREST [62], RLT [48], CN [63], HCFT [49], HDT [51], KCF [24], SINT [64] and ECO-HC [38]. It is obvious that our method obtains the first place in both precision rate and success rate, even better than trackers using deep features, such as PTAV, MCPF, C-COT, FCNT, STCT, CREST, FCNT, HDT and SINT. Although PTAV has the verification module and can correct the tracker when needed, this method just obtains the fourth place in precision rate and the ninth place in success rate. Compared with the baseline RLT tracker, our proposed method obtains an improvement (3.7% in precision rate and 2.2% in success rate), which proves that the adaptive

TABLE 2. Comparison of different methods in speed on OTB100.

Method	HCFT	HDT	LCTdeep	Staple	SAMF	DSST	LCT	MCCT_H	RLT	ECO-HC	Ours
Speed(fps)	12.2	5.5	8.6	56.7	18.7	10.5	12.9	30.5	37.1	37.4	7.1

TABLE 3. The precision rates of 12 trackers with 11 challenging attributes on TColor-128 dataset. The best, second best and third best tracking results are represented in red, blue and green, respectively.

Attribute	Ours	RLT	SAMF_AT	SRDCF	CoKCF	HDT	HCFT	SAMF	LCT	CFNET	DSST	ECO-HC
BC	0.815	0.782	0.681	0.711	0.767	0.767	0.744	0.617	0.679	0.620	0.552	0.791
DEF	0.878	0.811	0.757	0.756	0.784	0.767	0.803	0.735	0.759	0.562	0.508	0.774
FM	0.691	0.643	0.598	0.565	0.631	0.608	0.635	0.567	0.516	0.435	0.435	0.611
IPR	0.719	0.679	0.603	0.601	0.653	0.646	0.635	0.515	0.526	0.509	0.505	0.651
IV	0.771	0.716	0.633	0.669	0.726	0.692	0.722	0.625	0.671	0.542	0.589	0.669
LR	0.612	0.555	0.513	0.549	0.607	0.589	0.583	0.432	0.457	0.428	0.405	0.710
MB	0.640	0.599	0.597	0.581	0.642	0.586	0.636	0.562	0.525	0.469	0.455	0.611
OC	0.763	0.742	0.639	0.636	0.637	0.641	0.623	0.584	0.567	0.465	0.491	0.683
OPR	0.760	0.726	0.645	0.610	0.672	0.662	0.674	0.616	0.587	0.500	0.515	0.668
OV	0.635	0.540	0.507	0.523	0.507	0.469	0.492	0.421	0.461	0.313	0.384	0.619
SV	0.774	0.716	0.665	0.655	0.701	0.682	0.688	0.619	0.571	0.591	0.538	0.684

TABLE 4. The success rates of 12 trackers with 11 challenging attributes on TColor-128 dataset. The best, second best and third best tracking results are represented in red, blue and green, respectively.

Attribute	Ours	RLT	SAMF_AT	SRDCF	CoKCF	HDT	HCFT	SAMF	LCT	CFNET	DSST	ECO-HC
BC	0.586	0.563	0.494	0.499	0.505	0.509	0.489	0.433	0.462	0.415	0.384	0.558
DEF	0.627	0.588	0.558	0.528	0.532	0.535	0.542	0.537	0.533	0.401	0.361	0.547
FM	0.532	0.508	0.483	0.447	0.487	0.471	0.485	0.451	0.414	0.353	0.352	0.498
IPR	0.541	0.520	0.480	0.453	0.464	0.475	0.460	0.409	0.401	0.387	0.367	0.510
IV	0.583	0.559	0.496	0.512	0.507	0.486	0.500	0.484	0.485	0.410	0.421	0.524
LR	0.382	0.349	0.313	0.354	0.311	0.322	0.312	0.242	0.246	0.235	0.253	0.483
MB	0.462	0.434	0.446	0.424	0.446	0.430	0.441	0.410	0.370	0.337	0.333	0.452
OC	0.571	0.566	0.487	0.476	0.464	0.473	0.450	0.446	0.417	0.348	0.346	0.526
OPR	0.568	0.546	0.497	0.456	0.485	0.482	0.483	0.469	0.434	0.389	0.376	0.514
OV	0.443	0.404	0.416	0.395	0.401	0.372	0.389	0.339	0.366	0.247	0.289	0.474
SV	0.558	0.526	0.492	0.481	0.449	0.450	0.439	0.450	0.383	0.437	0.343	0.517

cosine window and motion feature are effective. To further validate the effectiveness of our method, tracking results on UAVDT dataset with 9 different challenging attributes are reported in FIGURE 7. We can see that our method achieves the first place in almost all the attributes except for small object (ranks second), scale variation (ranks third), long-term tracking (ranks fifth) and large occlusion (ranks fourth). Although our method is second to C-COT in small object, the difference is quite small, at only 0.1%. The adaptive cosine window can decrease the effect of boundary effect attributing to the enlarging searching region. As a result, our method is able to perform well in object blur and object motion.

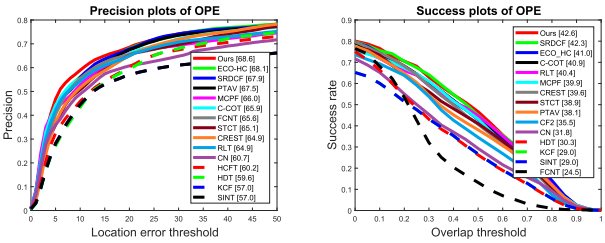


FIGURE 6. Precision plots and success plots of OPE of our proposed method against other state-of-the-art methods on UAVDT.

E. RESULTS ON UAV123 DATASET

UAV123 dataset contains 123 aerial high-resolution video sequences with more than 110K frames. These sequences covers 12 complex challenging attributes including scale variation (SV), aspect ratio change (ARC), low resolution (LR), fast motion (FM), full occlusion (FO), partial occlusion (PO), out-of-view (OV), background clutter (BC), illumination variation (IV), viewpoint change (VC), camera motion (CM) and similar object (SO). As these challenging attributes in UAV123 dataset, the tracking performance of various trackers decrease drastically compared with the OTB100 dataset in terms of the same evaluation metric. FIGURE 8 demonstrates the precision plots and success plots of OPE of our proposed method against other 16 state-of-the-art methods including RLT [48], SRDCF [27], CSDCF [65], Staple_CA [66], BACF [28], CoKCF [56], SRDCFdecon [36], KCC [67], SAMF_CA [66], Staple [53], SAMF [26], DSST [25], fDSST [68], DCF [23], EOC-HC [38] and KCF [24]. From FIGURE 8 we can

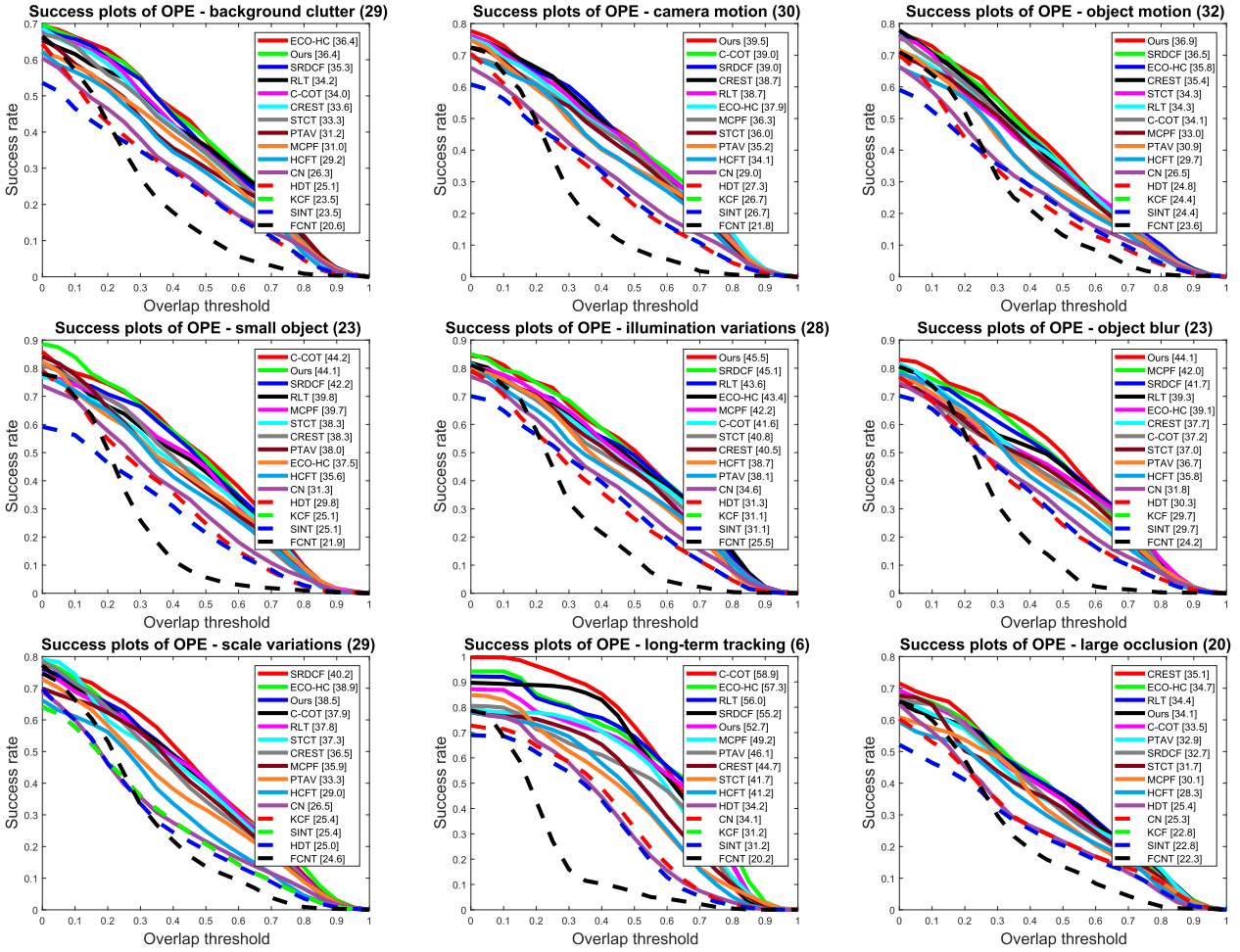


FIGURE 7. Success plots of OPE with different attributes on UAVDT.

see that EOC-HC obtains the first place both in precision rate and success rate. Since the re-detection module used in our tracker can solve the drifting problem, our method achieves the second best tracking performance with a precision score of 68.2% and with a success score of 48.1%. The baseline RLT tracker follows our tracker and achieves the third place. Owing to the adaptive cosine window and motion feature, our method improves the tracking performance by 3% in precision rate and 4% in success rate compared with RLT. To further demonstrate the comparative results, success plots of OPE with 12 different attributes on UAV123 are reported in FIGURE 9. It can be observed that our method achieves the second performance only to ECO-HC in almost all the attributes except for fast motion (ranks fourth), full occlusion (ranks third), background clutter (ranks second), illumination variation (ranks third) and similar object (ranks third). Compared with CoKCF using mutual deep features, our method only using handcrafted features and motion feature obtains better tracking results in all the attributes. Because of the adaptive cosine window, our tracker is able to enlarge the searching region and deal better with boundary effect than BACF.

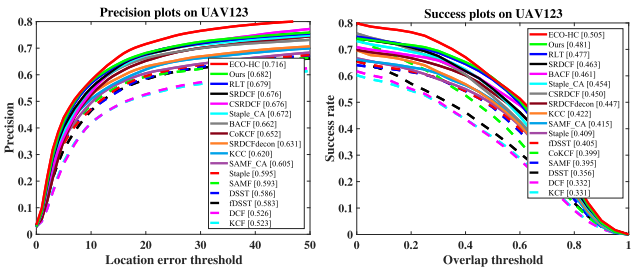


FIGURE 8. Precision plots and success plots of OPE of our proposed method against other state-of-the-art methods on UAV123.

F. RESULTS ON VOT2016 DATASET

In order to further evaluate our method, a comparison with other state-of-the-art trackers which participated in VOT2016 is demonstrated on FIGURE 10 and FIGURE 11. VOT2016 dataset consists of 60 challenging videos. The tracking performance is assessed in terms of accuracy and robustness, which consider the average overlap ratio and failure times, respectively. VOT2016 also introduces the expected average overlap to rank trackers which takes account of the raw values of per-frame accuracies and failures in a principled manner. FIGURE 10 gives the

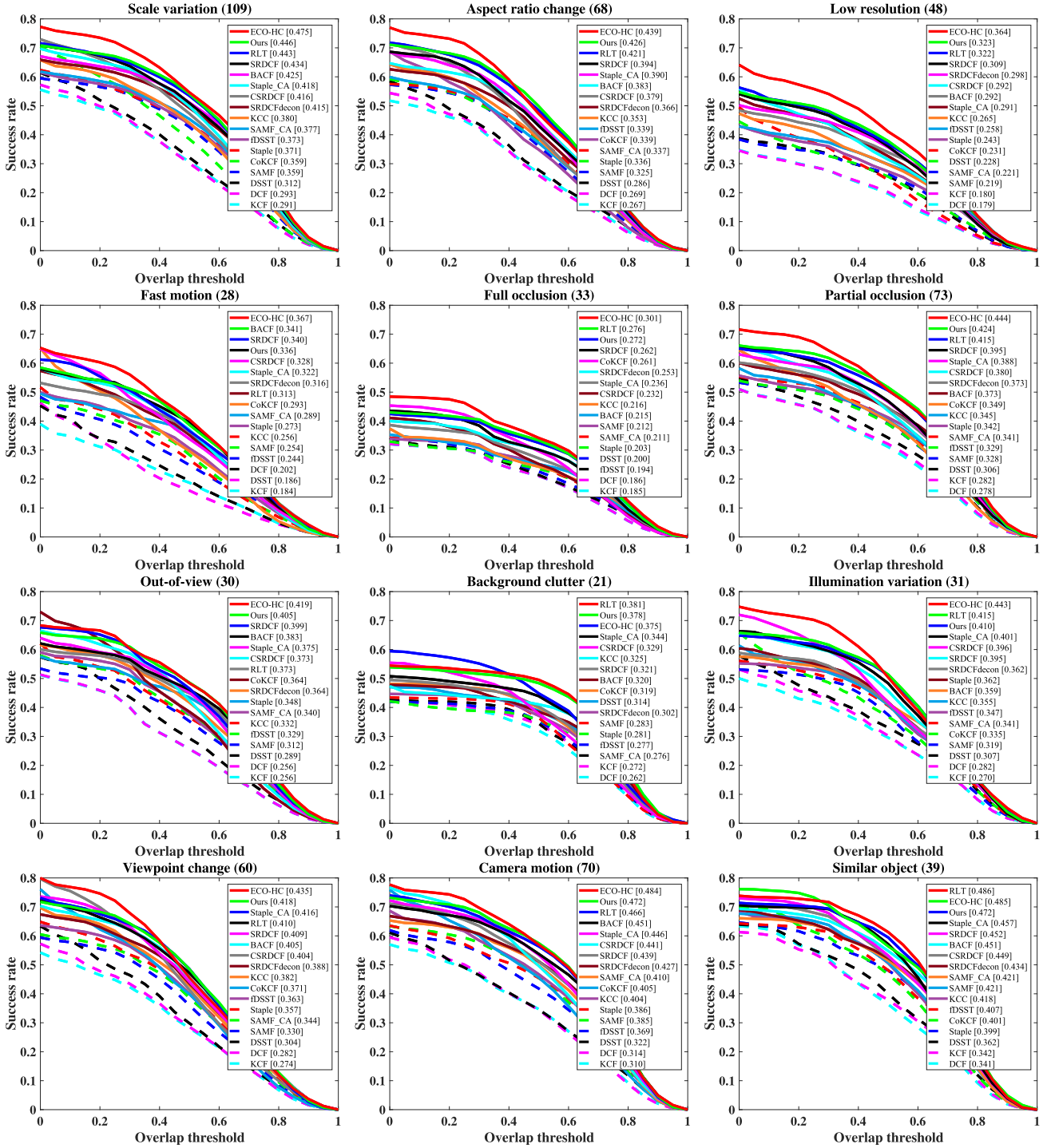


FIGURE 9. Success plots of OPE with different attributes on UAV123.

robustness-accuracy ranking plots under the baseline on VOT2016. FIGURE 11 demonstrates the expected average overlap graph on VOT2016. Trackers closer to the top-right of the plot perform better. We can see that our proposed method achieves the thirteenth place among the 50 trackers and performs better than most of trackers participated in VOT2016.

G. ABLATION ANALYSIS

In this section, we conduct extensive experiments to illustrate the effectiveness of each proposed component including adaptive cosine window, motion feature and re-detection scheme.

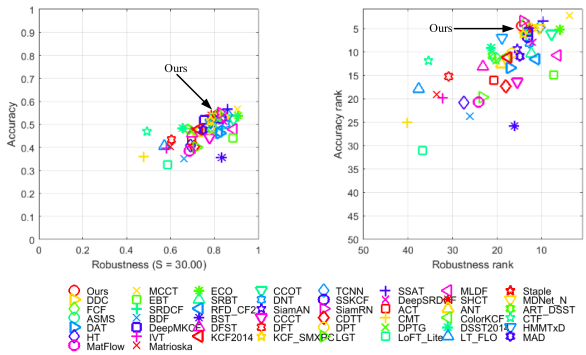


FIGURE 10. Robustness-accuracy ranking plots under the baseline on VOT-2016.

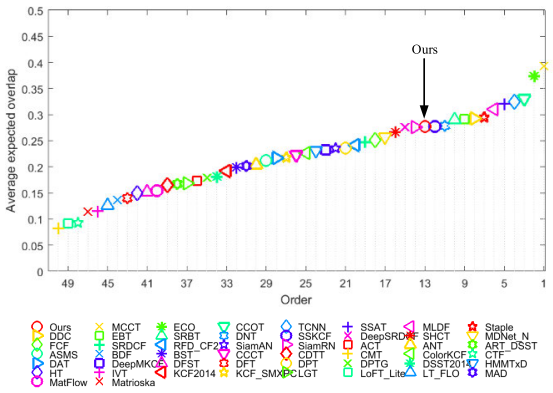


FIGURE 11. Expected average overlap graph on VOT2016.

1) EFFECT OF ADAPTIVE COSINE WINDOW

The adaptive cosine window plays an important role in our proposed method. It has the ability to highlight the target, suppress the background and can cope with boundary effect effectively. The first chart of FIGURE 12 gives the precision plots by RLT method with adaptive cosine window and fixed cosine window on OTB100. The second chart of FIGURE 12 shows the success plots by RLT method with adaptive cosine window and fixed cosine window on OTB100. Figure 14 reflects the effectiveness of adaptive cosine window with our proposed method on OTB100 in terms of precision plots and success plots, respectively. It is clear that the adaptive cosine window scheme improves the precision rate and success rate significantly. Both FIGURE 13 and FIGURE 15 report the tracking results of RLT and our method on OTB100 in terms of 11 challenging attributes with adaptive cosine window and fixed cosine window, respectively. It can be observed that both RLT method and our method with adaptive cosine window perform better than that with fixed cosine window in almost all the eleven challenging attributes.

Our proposed adaptive cosine window combines traditional cosine window and the target likelihood map of each frame with a fixed parameter η_3. η_3 determines the extent of each part acted on the final adaptive cosine window. Figure 16 shows the tracking performance in terms of precision

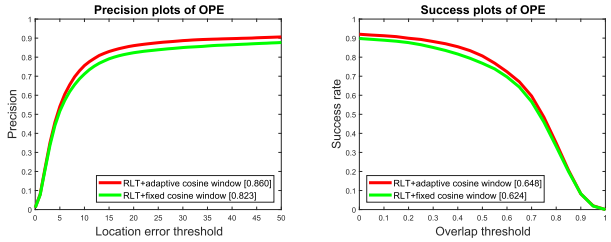


FIGURE 12. Precision plots and success plots of OPE by RLT method with adaptive cosine window and fixed cosine window on OTB100.

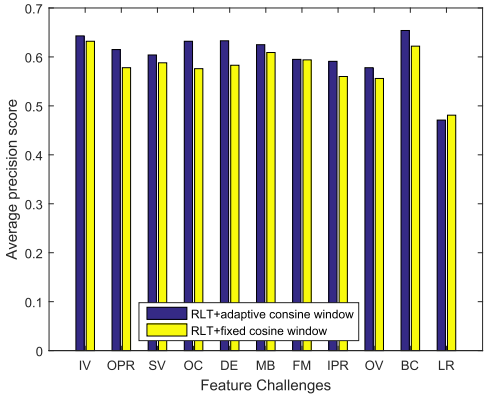


FIGURE 13. Tracking results of RLT on OTB100 in terms of 11 challenging attributes with adaptive cosine window and fixed cosine window.

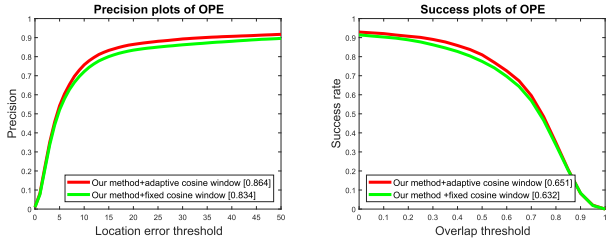


FIGURE 14. Precision plots and success plots of OPE by our method with adaptive cosine window and fixed cosine window on OTB100.

and success rate by our proposed method on OTB100 with different parameter η_3. We can see that when the parameter η_3 is set to 0.5, both precision value and success rate obtain the maximum value.

2) EFFECT OF MOTION FEATURE

Motion feature is important for our method and can be estimated by large displacement optical flow. The motion trajectory of target can be obtained and promotes the tracking performance greatly. The first picture of FIGURE 17 shows the precision plots of OPE by RLT method with motion feature and without motion feature on OTB100. The second picture of FIGURE 17 demonstrates the success plots of OPE by RLT method with motion feature and without motion feature on OTB100. Figure 19 shows the effectiveness of motion feature on our method. It can be observed that RLT with motion feature achieves a precision score of 82.7% and

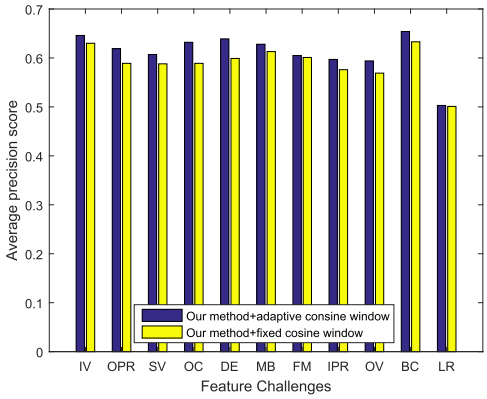


FIGURE 15. Tracking results of our method on OTB100 in terms of 11 challenging attributes with adaptive cosine window and fixed cosine window.

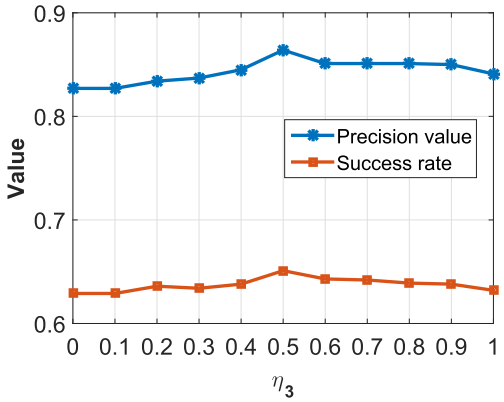


FIGURE 16. The effect of parameter η_3.

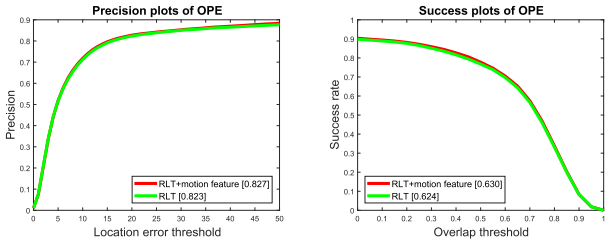


FIGURE 17. Precision plots and success plots of OPE by RLT method with motion feature and without motion feature on OTB100.

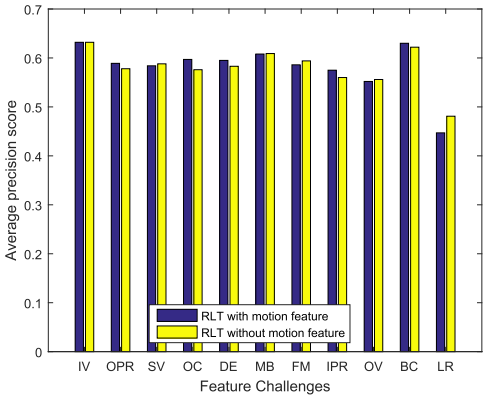


FIGURE 18. Tracking results of RLT on OTB100 in terms of 11 challenging attributes with motion feature and without motion feature.

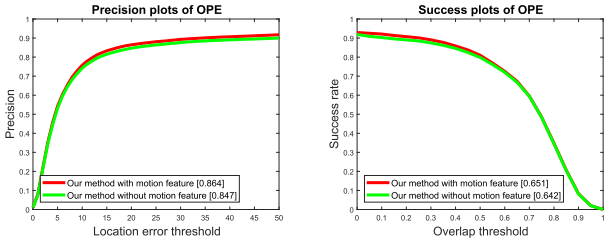


FIGURE 19. Precision plots and success plots of OPE by our method with motion feature and without motion feature on OTB100.

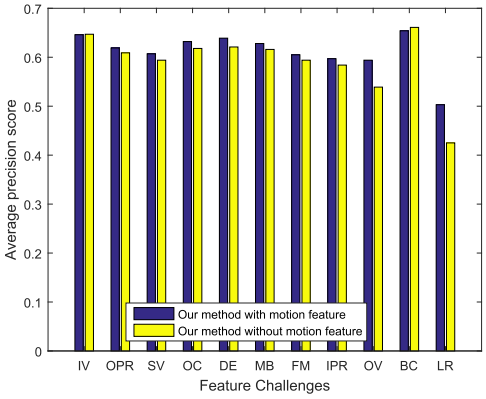


FIGURE 20. Tracking results of our method on OTB100 in terms of 11 challenging attributes with motion feature and without motion feature.

a success score of 63.0%, while RLT without motion feature only obtains a precision score of 82.3% and a success score of 62.4%. At the same time, our method combining three features (including Hog feature, color feature and motion feature) is able to get better performance than without motion feature. FIGURE 18 and FIGURE 20 demonstrate tracking results of RLT method and our method on OTB100 in terms of 11 challenging attributes with motion feature and without motion feature, respectively. It can be easily seen that our method using motion feature can get better performance in almost all the attributes except for IV and BC.

3) EFFECT OF RE-DETECTION COMPONENT

In this section, our method uses the same scheme as RLT to judge the reliability of tracking results. If the tracking results are regarded as unreliable, RLT method use traditional sparse representation to refine the candidates. On the contrary, our method uses reverse sparse representation to find coarse locations efficiently. FIGURE 21 and FIGURE 23 give the precision plots and success plots of OPE by RLT method and our method with traditional sparse representation and reverse sparse representation on OTB100, respectively. The reverse sparse representation reconstructs the template sets with a few candidates and builds a discriminative sparse similarity map to refine the candidates. FIGURE 22 and FIGURE 24 give the tracking results of RLT and our method on OTB100 in terms of 11 challenging attributes with traditional

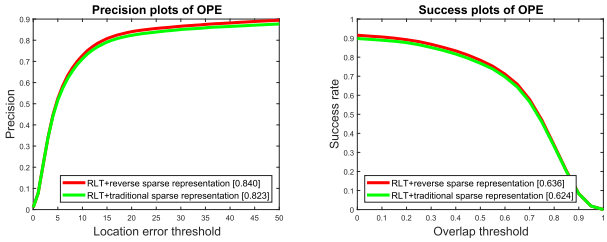


FIGURE 21. Precision plots and success plots of OPE by RLT method with traditional sparse representation and reverse sparse representation on OTB100.

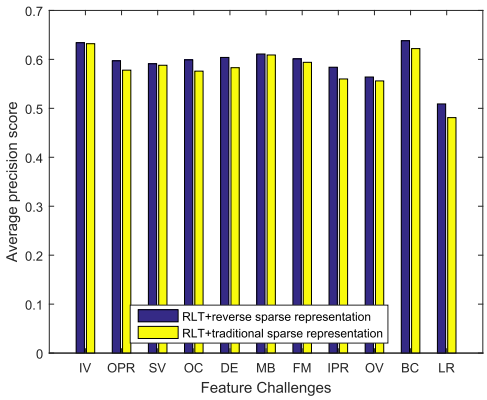


FIGURE 22. Tracking results of RLT on OTB100 in terms of 11 challenging attributes with traditional sparse representation and with reverse sparse representation.

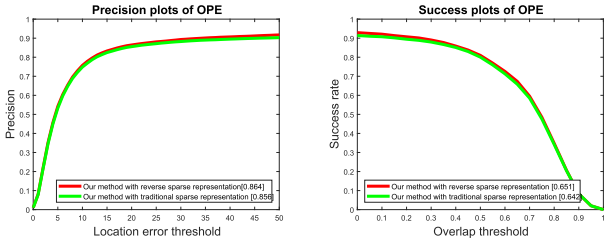


FIGURE 23. Precision plots and success plots of OPE by our method with traditional sparse representation and reverse sparse representation on OTB100.

sparse representation and with reverse sparse representation, respectively. It can be easily seen that our reverse sparse representation can promote the tracking performance significantly in terms of all the attributes.

H. FAILURE CASES

We give some tracking failure cases by our method in FIGURE 25. For the video *Jump* and *Diving*, targets undergo fast motion as well as serious deformation. As our method do not consider in-plane rotation and fails to locate the true position of target. In the *Person* sequence, target person passes through the pavilion and disappears for a long time. Although the re-detection module is activated due to full occlusion, our method can not locate the true position of small target person when target reappears in the far place.

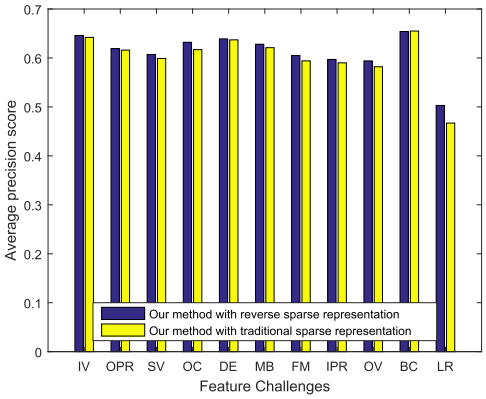


FIGURE 24. Tracking results of our method on OTB100 in terms of 11 challenging attributes with traditional sparse representation and with reverse sparse representation.

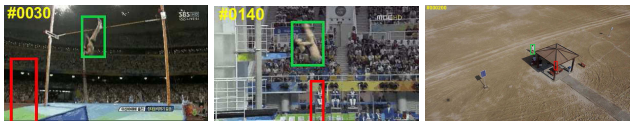


FIGURE 25. Failure cases on the *Jump*, *Diving* and *Person* sequences. Our results are shown in red and the ground truth in green.

V. CONCLUSION

In this paper, we propose a robust visual object tracking method with motion estimation and reliable re-detection scheme. First, motion feature is estimated by large displacement optical flow through adjacent frames, which is combined with color response and Hog correlation response to promote the tracking performance significantly. Second, an adaptive cosine window is adopted to deal with boundary effect, which has the ability to highlight target and suppress background effectively. Third, the color response and Hog correlation response is introduced to judge the reliability of tracking results. Fourth, reverse sparse representation is adopted to refine the candidates in case of tracking failures. At last, extensive experiments are conducted on five popular benchmarks to demonstrate the superiority of our proposed method.

REFERENCES

[1] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.
[2] D. Yuan, X. Zhang, J. Liu, and D. Li, "A multiple feature fused model for visual object tracking via correlation filters," *Multimedia Tools Appl.*, vol. 78, no. 19, pp. 27271–27290, Oct. 2019.
[3] C. Sun, D. Wang, H. Lu, and M.-H. Yang, "Correlation tracking via joint discrimination and reliability learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 489–497.
[4] S. Zhang, Y. Qi, F. Jiang, X. Lan, P. C. Yuen, and H. Zhou, "Point-to-set distance metric learning on deep representations for visual tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 187–198, Jan. 2018.
[5] Y. Kuai, G. Wen, and D. Li, "Learning fully convolutional network for visual tracking with multi-layer feature fusion," *IEEE Access*, vol. 7, pp. 25915–25923, 2019.
[6] D. Yuan, W. Kang, and Z. He, "Robust visual tracking with correlation filters and metric learning," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105697.

[7] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4665–4674.

[8] W. Ou, D. Yuan, Q. Liu, and Y. Cao, "Object tracking based on online representative sample selection via non-negative least square," *Multimedia Tools Appl.*, vol. 77, no. 9, pp. 10569–10587, May 2018.

[9] D. Yuan, N. Fan, and Z. He, "Learning target-focusing convolutional regression model for visual object tracking," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105526, doi: 10.1016/j.knosys.2020.105526.

[10] D. Yuan, X. Li, Z. He, Q. Liu, and S. Lu, "Visual object tracking with adaptive structural convolutional network," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105554, doi: 10.1016/j.knosys.2020.105554.

[11] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," 2020, *arXiv:2003.06761*. [Online]. Available: http://arxiv.org/abs/2003.06761

[12] B. Zhong, B. Bai, J. Li, Y. Zhang, and Y. Fu, "Hierarchical tracking by reinforcement learning-based searching and coarse-to-fine verifying," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2331–2341, May 2019.

[13] B. Zhong, H. Yao, S. Chen, R. Ji, T.-J. Chin, and H. Wang, "Visual tracking via weakly supervised learning from multiple imperfect oracles," *Pattern Recognit.*, vol. 47, no. 3, pp. 1395–1410, Mar. 2014.

[14] D. Yuan, X. Lu, D. Li, Y. Liang, and X. Zhang, "Particle filter re-detection for visual tracking via correlation filters," *Multimedia Tools Appl.*, vol. 78, pp. 14277–14301, Jun. 2019.

[15] J. Fan, H. Song, K. Zhang, Q. Liu, and W. Lian, "Complementary tracking via dual color clustering and spatio-temporal regularized correlation learning," *IEEE Access*, vol. 6, pp. 56526–56538, 2018.

[16] K. Zhang, J. Fan, Q. Liu, J. Yang, and W. Lian, "Parallel attentive correlation tracking," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 479–491, Jan. 2019.

[17] 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.

[18] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5630–5644, Dec. 2015.

[19] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2018, pp. 375–391.

[20] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2016, pp. 445–461.

[21] M. Kristan *et al.*, "The visual object tracking VOT2016 challenge results," in *Proc. Int. Conf. Comput. Vision Workshops (ECCVW)*. Cham, Switzerland: Springer, Oct. 2016, pp. 777–823.

[22] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[23] 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. (ECCV)*. Cham, Switzerland: Springer, Oct. 2012, pp. 702–715.

[24] 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.

[25] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Nottingham, U.K., 2014, pp. 1–11.

[26] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*. Cham, Switzerland: Springer, 2014, 254–265.

[27] M. Danelljan, G. Hager, 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.

[28] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Oct. 2017, pp. 1144–1152.

[29] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4904–4913.

[30] X. Sun, N. M. Nasrabadi, and T. D. Tran, "Supervised deep sparse coding networks for image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 405–418, Jul. 2020.

[31] D. Guo, Y. Pei, K. Zheng, H. Yu, Y. Lu, and S. Wang, "Degraded image semantic segmentation with dense-gram networks," *IEEE Trans. Image Process.*, vol. 29, pp. 782–795, Aug. 2020.

[32] C. Tian, Y. Xu, and W. Zuo, "Image denoising using deep CNN with batch renormalization," *Neural Netw.*, vol. 121, pp. 461–473, Jan. 2020.

[33] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided CNN for image denoising," *Neural Netw.*, vol. 124, pp. 117–129, Apr. 2020.

[34] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Robust visual tracking via hierarchical convolutional features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2709–2723, Nov. 2019.

[35] Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, J. Lim, and M.-H. Yang, "Hedging deep features for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1116–1130, May 2019.

[36] M. Danelljan, G. Hager, 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)*, Jun. 2016, pp. 1430–1438.

[37] M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2016, 471–488.

[38] 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)*, Jul. 2017, pp. 6931–6939.

[39] X. Mei and H. Ling, "Robust visual tracking using ℓ_1 minimization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Kyoto, Japan, Sep./Oct. 2009, pp. 1436–1443.

[40] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1830–1837.

[41] Z. Xiao, H. Lu, and D. Wang, "L2-RLS-based object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1301–1309, Aug. 2014.

[42] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparse collaborative appearance model," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2356–2368, May 2014.

[43] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via coarse and fine structural local sparse appearance models," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4555–4564, Oct. 2016.

[44] B. Zhuang, H. Lu, Z. Xiao, and D. Wang, "Visual tracking via discriminative sparse similarity map," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1872–1881, Apr. 2014.

[45] D. Wang, H. Lu, Z. Xiao, and M.-H. Yang, "Inverse sparse tracker with a locally weighted distance metric," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2646–2657, Sep. 2015.

[46] C. Sun, D. Wang, and H. Lu, "Occlusion-aware fragment-based tracking with spatial-temporal consistency," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3814–3825, Aug. 2016.

[47] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[48] N. Wang, W. Zhou, and H. Li, "Reliable re-detection for long-term tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 730–743, Mar. 2019.

[49] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3074–3082.

[50] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Adaptive correlation filters with long-term and short-term memory for object tracking," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 771–796, Mar. 2018.

[51] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4303–4311.

[52] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4844–4853.

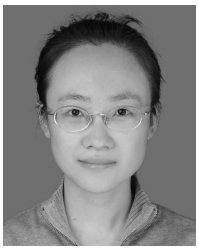
[53] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1401–1409.

[54] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5388–5396.

[55] A. Bibi, M. Mueller, and B. Ghanem, "Target response adaptation for correlation filter tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*. Cham, Switzerland: Springer, Sep. 2016, pp. 419–433.
[56] L. Zhang and P. N. Suganthan, "Robust visual tracking via co-trained kernelized correlation filters," *Pattern Recognit.*, vol. 69, pp. 82–93, Sep. 2017.
[57] 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. (CVPR)*, Jul. 2017, pp. 5000–5008.
[58] H. Fan and H. Ling, "Parallel tracking and verifying," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4130–4144, Aug. 2019.
[59] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 365–378, Feb. 2019.
[60] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3119–3127.
[61] L. Wang, W. Ouyang, X. Wang, and H. Lu, "STCT: Sequentially training convolutional networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1373–1381.
[62] 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. (ICCV)*, Oct. 2017, pp. 2574–2583.
[63] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1090–1097.
[64] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1420–1429.
[65] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4847–4856.
[66] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1387–1395.
[67] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel cross-correlator," in *Proc. Conf. Assoc. Adv. Artif. Intell. (AAAI)*, Feb. 2018, pp. 4179–4186.
[68] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.

HAIJUN WANG received the M.S. degree from the School of Information Science and Engineering, Shandong University, China, in 2007. He is currently pursuing the Ph.D. degree with the College of Civil Aviation, Nanjing University of Aeronautics and Astronautics. He is also with the Flying College, Binzhou University. His research interests include visual tracking and image segmentation.

WENLAI MA received the M.S. degree from the College of Information Science and Engineering, Northeastern University, China. He is currently pursuing the Ph.D. degree with the Nanjing University of Aeronautics and Astronautics, China. He is also with the Flight College, Binzhou University. His research interest includes abnormal behavior and conflict of UAV.



SHENGYAN ZHANG received the M.S. degree from the School of Information Science and Engineering, Dalian University of Technology, China, in 2008. She is currently with the Flight College, Binzhou University, China. Her research interest includes visual tracking.

GUO CHEN received the Ph.D. degree in vehicle engineering from Southwest Jiaotong University, Chengdu, China, in 2000. He is currently a Professor with the College of Civil Aviation, Nanjing University of Aeronautics and Astronautics. His current research interests include machine vibration and fault diagnosis.

HONGJUAN GE received the Ph.D. degree in electric machine and electric appliance from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2007. She is currently a Professor with the College of Civil Aviation, Nanjing University of Aeronautics and Astronautics. Her current research interests include electric machine appliance and airplane equipment design research.

YUJIE DU received the Ph.D. degree in electronic science and technology from the Nanjing University of Science and Technology, China, in 2012. He is currently a Professor with the Flight College, Binzhou University, China. His research interests include visual tracking and image classification.