An improved visual objects tracking algorithm based on features fusion with neural networks, Discrete Cosine Transform, and Histogram of Oriented Gradients

[1]Hanane NEBBAR, [2]Nadjiba TERKI, [3]Mohammed BOURENNANE, [4]Fouaze MOUSSI, [4]Saloua OUARHLENT

[1]Department of Electrical Engineering, LESIA Laboratory, University of Mohamed Khider - Biskra, Biskra, Algeria, Hanane.nebbar@univ-biskra.dz
[2]Department of Electrical Engineering, LESIA Laboratory, University of Mohamed Khider - Biskra, Biskra, Algeria, n.terki@univ-Biskra.dz
[3]Department of Physics, University of Ziane Achour, Djelfa, Algeria, m.bourennane@univ-djelfa.dz
[4]Department of Electrical Engineering, University of Mohamed Khider - Biskra, Biskra, Algeria, fouaze.moussi@univ-biskra.dz
[5]Department of Electrical Engineering, University of Mohamed Khider - Biskra, Biskra, Algeria, saloua.ouarhlent@univ-biskra.dz

Abstract

This paper presents a novel method for Visual Object Tracking (VOT), which treats some challenging problem such as significant changes in appearance caused by occlusion and variations in illumination. The proposed approach combines the Deep Convolutional Neural Networks (DCNN), Discrete Cosine Transform (DCT), Histograms of Oriented Gradients (HOG) features and HSV energy condition. Firstly, an HSV-based energy condition is employed to enhance the learning process by incorporating both RGB and HSV color bases. Instead of using the image template, the technique utilizes the coefficients of the image DCT to handle high saturation images in the Convolutional Neural Networks (CNN's) input. The Inverse Discrete Cosine Transform (IDCT) is used to extract the CNN features. Secondly, the multichannel correlation maps generated by the CNNs are utilized to determine the target position. This is achieved by combining convolutional features. Newton's method is also employed in this process to enhance the long-term memory of the target's appearance and assist in recovering from tracking failures. The updating parameter for the correlation filters is calculated as the highest value among the output maps generated by correlation filters using convolutional features derived from the HOG features of the image template. Finally, the results obtained undeniably prove that the proposed method surpasses most recent tracking techniques.

Keywords Convolutional neural network, Discrete Cosine Transform (DCT), Correlation filter, Visual tracking, Newton's method.

DOI Number: 10.48047/nq.2023.21.7.nq23012 **NeuroQuantology2023;21(7):100-113**

I. INTRODUCTION

Visual tracking poses as one of the most complex problems in computer vision, finding applications in diverse fields like human-computer interaction, video surveillance, and unmanned driving. The main objective in generic visual tracking algorithm is to anticipate the path of a target through a series of images, starting from its initial position. Nonetheless, creating a fast and dependable tracker is a formidable undertaking due to numerous hurdles, such as occlusion, swift motion, and distortion. Moreover, the scarcity of training samples further complicates the development of an efficient and robust tracking system. To address these challenges,

various pioneering trackers have been proposed, leading to significant advancements in tracking performance and robustness. Notably, discriminative-filter-based trackers [1] have garnered considerable attention owing to their competitive performance. Typically, visual tracking methods can be divided into two main types: generative approaches and discriminative methods [2]. Discriminative approaches have witnessed significant progress based on correlation filters, and Examples of successful tasks include object identification, image segmentation, and image classification, which have been effectively achieved [3]. In recent times, DCNNs have gained significant popularity as a prominent method in visual tracking [4]. The utilization of CNN for human tracking was introduced in [5]. The VGG-Net-19 model was utilized to train three adaptive correlation filters in [6]. The effectiveness of this approach was evaluated using contemporary methods. However, despite its advantages, the sustainability of long-term tracking was found to be limited [7].

Huang et al [8] employed reinforcement learning to train an early decision policy, resulting in improved speed for object tracking using CNN. Similarly, Wang et al [9] introduced an approach that involves utilizing an automatic denoising encoder stack to learn generic features for visual tracking. Furthermore, He et al. [10] developed a two-part Siamese network consisting of a semantic branch and an appearance branch, aiming to enhance the discrimination capabilities of SiameFC in tracking. Nevertheless, despite the advantages offered by these techniques, the challenge of sustaining long-term tracking effectiveness persists. In an attempt to overcome this challenge, several researchers have endeavored to improve tracking performance by integrating feature representations from various CNN layers with correlation filters [11, 12]. In [13], the authors have presented an effective hybrid image fusion method that combines the Integer Lifting Wavelet Transforms (ILWT) and the DCT. This method is capable of generating fused images with superior visual quality, making it a potential solution to mitigate certain visual tracking issues.

The utilization of the DCT in visual tracking has received limited attention in the existing literature [14], despite its effectiveness in diverse visual applications like image retrieval [15], face recognition [16], and video object segmentation [17]. In [18], authors introduced a particle filter framework that integrated a sparse appearance model based on structural local DCT, which included occlusion detection for visual tracking.

In recent times, Histograms of Oriented Gradients (HOG) features have emerged as a valuable tool for addressing various challenges in detection and classification. Notably, the successful identification of faces [19] has been accomplished by leveraging the magnitudes and orientations of image derivatives. Y. Wei et al. [20] have introduced a Haar-HOG-based technique, which has shown promise by delivering remarkable speed and efficiency compared to algorithms relying solely on Haar-like features or isolated HOG descriptors. Additionally, this proposed method demonstrated a lower false positive rate and a higher detection rate when compared to techniques that solely rely on the HOG descriptor.

The key contributions outlined in this paper are as follows:

- By applying the HSV energy condition, we tackle the issue of light variation in individual color frames. Our approach allows to opt for either RGB or HSV color bases. Additionally, the DCT has the capability to capture pertinent spatial frequency information. Notably, in the top left corner of the corresponding 2-D DCT matrix, a concentrated cluster of low-frequency coefficients is observed.

- Given the significance of integrating feature representations from multiple CNN layers, as exemplified in [21], a HCF model has been devised.

- HOG features derives from the 2D-DCT coefficients, utilizing them as a basis instead of the original image. This introduction of 2D-DCT coefficients aims to enhance the performance of HOG features. Moreover, a technique has been formulated to counteract model drift, aiding in the identification of alterations in appearance and demonstrated superior empirical results for both object detection and real-time tracking [22]. This involves the utilization of Newton's method to compute the maximum value within the maps generated through correlation filters. Unlike [23], we compute the convolutional feature products derived from the HOG features extracted from an existing image template. Subsequently, this computed value is employed as a parameter for updating the correlation filters.

- The proposed approach evaluates using a comprehensive benchmark dataset known as OTB50, which consists of 50 challenging image sequences.

The presented paper is organized as follows. Section II provides a detailed explanation of the proposed approach. The Section III focuses on the DCT. In Section IV, we elucidate the fusion of the RGB and HSV color bases. In Section V, the effectiveness of the tracker derived from the proposed approach is demonstrated by comparing the obtained results with those presented in various previous references. Finally, conclusions are summarized in Section VI.

II. PROPOSED ALGORITHM

In this section, our method is presented, focusing on effectively handling diverse challenging appearance changes of the target, such as substantial occlusion,

illumination variations, and scale variations. Figure 1 showcases the different stages of the tracking algorithm. The fundamental algorithm can be succinctly described in three essential steps.

Initially, following a similar methodology as presented in [21], we utilize CNN features to train four two-dimensional correlation filters to estimate the target's location.

Next, we introduce a novel approach that involves integrating RGB and HSV color transformations with DCT decomposition. This innovative technique allows us to enhance the tracking process further.

Lastly, we calculate the maximum value from the resultant maps utilizing the correlation filters, Newton's method, and the convolutional features extracted from the HOG feature-based image template. This computed value plays a crucial role as a parameter in the correlation filters' update process.

II.1 Convolution Features

Convolutional Neural Networks have exhibited remarkable success across a range of computer vision tasks. In this investigation, we introduce a novel approach involving translation estimation by leveraging a CNN model to extract features and establish a translation model. Specifically, we harness four layers from the VGGNet-19 model to extract convolutional features.

In the context of visual object tracking, the precise determination of the target object's position takes precedence over its semantic category. As a result, we employ bilinear interpolation [11] to resize each input frame to dimensions of 224 × 224. Subsequently, we collect the outputs from pool 1, pool 2, pool 3, and pool 4 layers to create a multichannel feature map.

As the depth of the CNN increases, the spatial resolution of the target object gradually diminishes due to the pooling operations. To mitigate this challenge, we address each feature map's size using bilinear interpolation, as outlined in equation (1), ensuring that they are resized to a consistent spatial resolution of $M/4 \times N/4$. Here, M and $N/4$ denote the dimensions of the feature vector x. This approach guarantees uniform spatial resolution across the pooling layers.

$$x_i = \sum_k \alpha_{ik} . h_k \tag{1}$$

In this context, x_i stands for the up sampled feature vector at the i^{th} location, and h_k represents the feature map corresponding to the k^{th} feature. Meanwhile, α_{ik} is a weight interpolation factor that relies on the specific positions of the i^{th} and k^{th} vectors within the adjacent features.

II.2 Correlation Filters

The correlation filters showcase a proficient encoding of the visual attributes of the target object [24]. The procedure of acquiring the correlation filter models W entails addressing the subsequent minimization challenge:

$$W^* = \arg\min_W \sum_{m,n} \left\| W . x_{m,n} - y(m,n) \right\|^2 + \lambda \|W\|^2 \tag{2}$$

The learned correlation filter model is denoted as W^*. The feature vector x is characterized by its dimensions, which are M, N and D, with M representing width, N representing height, and D representing the number of channels. The regularization parameter λ assumes values that are non-negative.

where

$$W . x_{m,n} = \sum_{d=1}^{D} W_{m,n,d}^T . x_{m,n,d}$$

with $W_{m,n}^T$ and d the transposed weight for each channel d at position (m,n). The correlation filter model's dimensions are $M \times N$ [25]. Each shifted sample of $x_{m,n}(m,n) \in \{0,1,...,M-1\} \times \{0,1,...,N-1\}$ is associated with a Gaussian function label $y(m,n)$ through the regression process:

$$y(m,n) = e^{-\frac{\left(m-\frac{M}{2}\right)^2 - \left(n-\frac{N}{2}\right)^2}{2\sigma^2}} \tag{3}$$

where σ is the standard deviation.

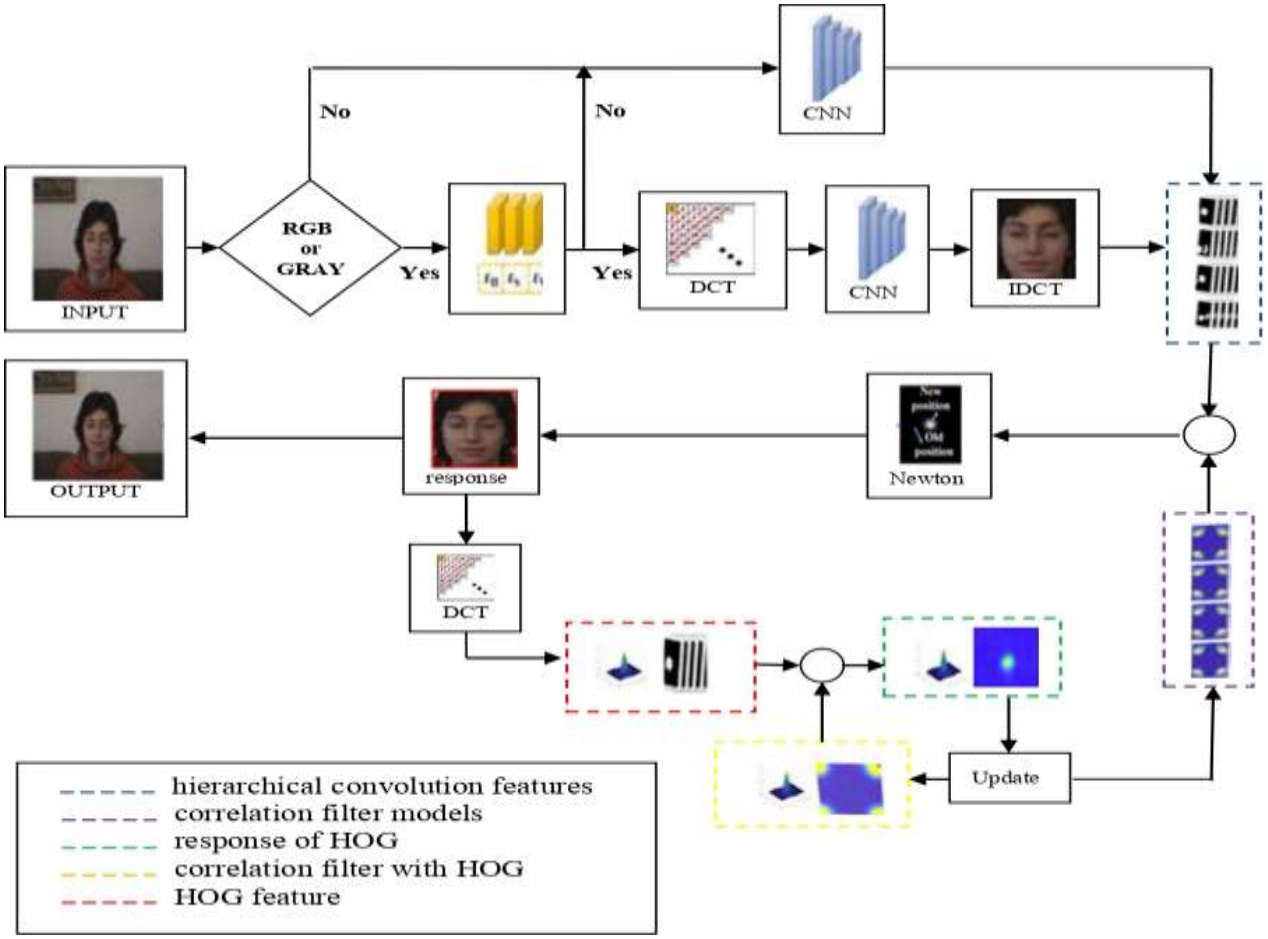


Fig. 1. Key Phases in the Proposed Algorithm Implementation.

The optimization problem outlined in equation (2) can be separately addressed for each feature channel by utilizing fast Fourier transformation (FFT), akin to the vector correlation filter training method described in [26]. In the frequency domain, the learned filter for the d^{th} channel (where d takes values from 1 to D) is defined according to equation (4).

$$W^d = \frac{Y \odot \overline{X}^d}{\sum_{i=1}^{D} X^i \odot \overline{X}^i + \lambda} \tag{4}$$

where y is the Fourier transformation form of $y = y(m,n) \mid (m,n) \in \{0,1,...,M-1\} \times \{0,1,...,N-1\}$ and the bar refer to the complex conjugation. The operator $\odot$ is the Hadamard product.

To compute the d^{th} correlation response map f_l , the Inverse Fast Fourier Transform (IFFT) is employed, as expressed by the equation:

$$f_l = \mathcal{F}^{-1}\left(\sum_{d=1}^{D} W^d \odot \overline{Z}^d\right), \text{where } l = 1,2,...,3$$

During the tracking process, a multi-channel vector Z is employed to compute the value of f_l . The uppercase letters indicate the Fourier transform signals associated with it, the inverse FFT operation is represented by $\mathcal{F}^{-1}$, and the complex conjugation is indicated by the bar symbol.

II.3 Estimation of Coarse-to-Fine Translation

To determine the target translation within the correlation response maps of each layer, denoted as f_l , a search is performed to locate the maximum value in the previous layer $(l-1)^{th}$. The corresponding location in the current layer l^{th} is taken as a reference point for regularization. The most suitable position of the target in the $(l-1)^{th}$ layer is subsequently identified by maximizing the weighted summation of response s from the $(l\text{-}1)^{th}$ and l^{th} layers, while adhering to certain constraints.

$$\arg\min f_{l-1}(m,n) + \gamma f_l(m,n), \quad |m-\hat{m}| + |n-\hat{n}| \le r \tag{5}$$

Within a region of size $r \times r$ centered around $(\hat{m}-\hat{n})$, the search is confined to neighboring areas, ensuring limitations on the search range. Progressing from the outermost to the innermost layers, each response value is subject to multiplication by a regularization factor γ and subsequently propagated back to the response map of preceding layers [11]. Ultimately, through the maximization of equation (5) on the layer boasting the highest spatial resolution, the estimation of the target location is achieved.

Furthermore, by employing Eqs.(2), (4), (5), and Newton's method, the utmost response of the correlation filter derived from HOG can be calculated for $l=1$, and $\gamma=1$.

Newton's method, a technique in the field of optimization, is employed to discover global extreme. By calculating both the gradient and the hessian [27], this method seeks the highest score during each iteration. The process achieves convergence with only a limited number of iterations.

II.4 Model Update

During the tracking process, a significant change in the object's appearance between two consecutive images is evident, leading to potential tracker drifts [26]. To address this issue, it becomes crucial to update the correlation filter model obtained through equation (2) by incorporating a learning rate denoted as η, as demonstrated in equation (6).

$$\begin{cases} \hat{x}^t = (1-\eta)\,x^{t-1} + \eta x^t \\ \hat{W}^t = (1-\eta)\,W^{t-1} + \eta W^t \end{cases} \tag{6}$$

III. DISCRETE COSINE TRANSFORM

The DCT is a mathematical technique that converts a signal from its spatial representation to the frequency domain. By utilizing the DCT, important spatial frequency information in a 2-D signal can be efficiently captured using a small set of low-frequency coefficients, which typically group together in the upper left corner of the corresponding 2-D DCT matrix. This exceptional energy compaction characteristic has led to widespread adoption of the DCT in various applications, including data compression and image quality evaluation. The 2-D DCT of an $M \times N$ image matrix f can be defined as follows:

$$F(u,v) = \alpha_u \alpha_v \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} f(i,j) \times \cos\left(\frac{(2i+1)u.\pi}{2M}\right) \times \cos\left(\frac{(2j+1)v.\pi}{2N}\right) \tag{7}$$

where the 2-D IDCT transform is defined as follows:

$$f(i,j) = \alpha_u \alpha_v \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} f(u,v) \times \cos\left(\frac{(2i+1)u.\pi}{2M}\right) \times \cos\left(\frac{(2j+1)v.\pi}{2N}\right)$$

In this transformation, the indices u and v are constrained within the range of 0 to $(M-1)$ and 0 to $(N-1)$, respectively. The pixel intensity at coordinates (i,j) in the original signal is denoted by $f(i,j)$, while the corresponding transform coefficient located at row u and column v in the DCT matrix is represented by $F(u,v)$. To ensure appropriate normalization during the DCT calculation, vital scalar values α_u and α_v are defined as normalization coefficients. These coefficients play a crucial role in the normalization process of the DCT transformation.

$$\alpha_u = \begin{cases} 1/\sqrt{M}, & u = 0 \\ \sqrt{2/M}, & 1 \le u \le M-1 \end{cases} \tag{8}$$

$$\alpha_v = \begin{cases} 1/\sqrt{N}, & v = 0 \\ \sqrt{2/N}, & 1 \le v \le N-1 \end{cases} \tag{9}$$

The DCT coefficient $F(0,0)$ located at the top left corner of the matrix is referred to as the DC term. As for the other DCT coefficients, they represent AC terms and correspond to high spatial frequency coefficients arranged in increasing order.

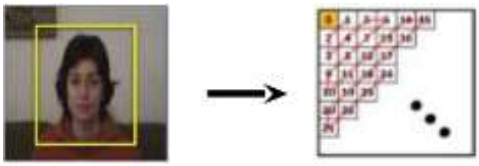


Fig. 2. displays image patches on the left and the DCT coefficient matrix on the right. The yellow color highlights the dc term, while the remaining terms represent the selected ac terms.

IV. CONDITION BASED ON HSV-ENERGY

In this section, we present an innovative strategy to tackle the issue of managing fluctuations in illumination, which proves to be a formidable obstacle for numerous benchmark trackers. The technique revolves around harnessing the energy constituents within the HSV color space. The notion of energy finds broad utility across various domains, encompassing wireless sensor networks [28], image reconstruction [29], and beyond.

For every input RGB frame, we utilize the energy utilization of individual components within the HSV color space to categorize the frame into two groups: low light and high light. The initial category encompasses frames with low energy consumption and minor alterations in lighting, while the subsequent category encompasses

frames with elevated energy consumption and notable fluctuations in lighting.

Figure 3 the depiction showcases how the HSV color space establishes the fundamental framework for computing the energy consumption of every input frame.

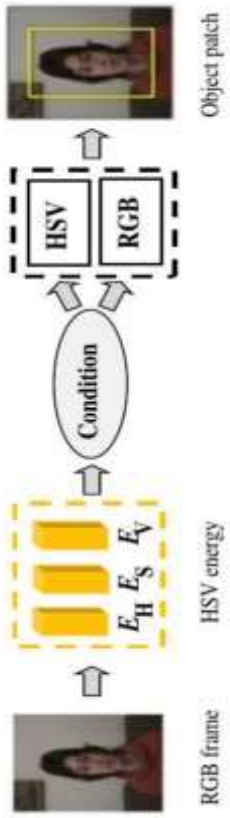


Fig. 3. Condition based on HSV-Energy.

The energy associated with the k^{th} component is represented by E_k. The proportion of energy consumption attributed to each individual HSV component is defined in the subsequent manner:

$$E_k = 100 \times \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(S_{ij}\right)^2}{E_T} \tag{10}$$

$$E_T = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(H_{ij}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}\left(S_{ij}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}\left(V_{ij}\right)^2 \tag{11}$$

If E_k is greater than $23\times100\%$, it indicates that the illumination is very weak. In such situations, the coefficients of the image's DCT are utilized as input for the CNN. Conversely, in the opposite scenario, the input image is decomposed into its RGB components. This step follows the approach proposed in this paper.

V. EXPERIMENTS

The presented algorithm underwent validation and assessment using the OTB50 benchmark dataset, which comprises 50 videos. The tracking algorithm was coded in MATLAB and operated on an Intel I5-12400F 2.50 GHz CPU equipped with 16 GB RAM, with additional assistance from the MatConvNet toolbox. Feature extraction involved carrying out CNN forward propagation on a GeForce GTX1060 GPU.

Convolutional neural network introduced by the Visual Geometry Group in 2012, VGG-Net-19, has been Exploited in this study. This network was made of 19 layers, featuring 16 convolutional layers, 3 fully connected layers, 5 MaxPool layers, and 1 SoftMax layer [30]. To extract features, the network underwent training on the comprehensive hierarchical image repository, Image Net [31].

During the process of feature extraction, only the outputs from pool 1, pool 3, pool 4, and pool 5 were employed. The search window size was remained constant at 1.8 times the target size. A regularization parameter λ of 10^{-4} has been chosen, and the kernel width for generating Gaussian function labels has been set at 0.1. Furthermore, the learning rate η in equation (6) was established as 0.01, also, the control updating parameter was fixed to 0.3. Additionally, the value of γ was varied across different layers: 1 for conv5-4, 0.5 for conv4-4, 0.25 for conv3-4, and 0.15 for conv1-4 layers.

Method evaluation employs Distance Precision (DP) and Overlap Success (OS) metrics. A comparison is conducted against other reference methods [11, 32, 33]. The outcomes for the two performance metrics are presented via two curves within One-Pass Evaluation (OPE). The first curve depicts the distance precision rate based on the location error threshold, indicating the portion of frames where tracking results lie within a specific number of pixels from the ground truth. The second curve portrays the success rate relative to the overlap threshold, signifying the percentage of frames where tracking was successful. The location error threshold spans from 0 to 50, while the overlap threshold is adjusted across the range of 0 to 1. Figure 4 displays the comparative outcomes on the OTB-50 dataset. Among these, the HCFTs tracker attains the second-highest performance levels, registering a distance precision of 88.6% and an overlap success rate of 79.3% . It is worth highlighting that the newly proposed approach showcases its effectiveness by achieving enhancement gains of 2.7% in distance precision and 4.4% in overlap success.

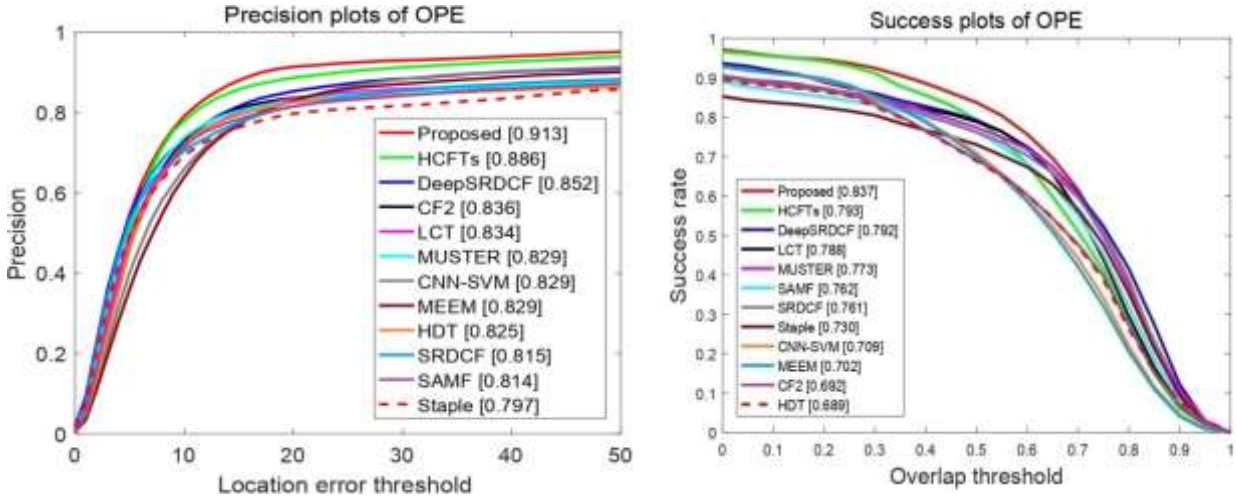
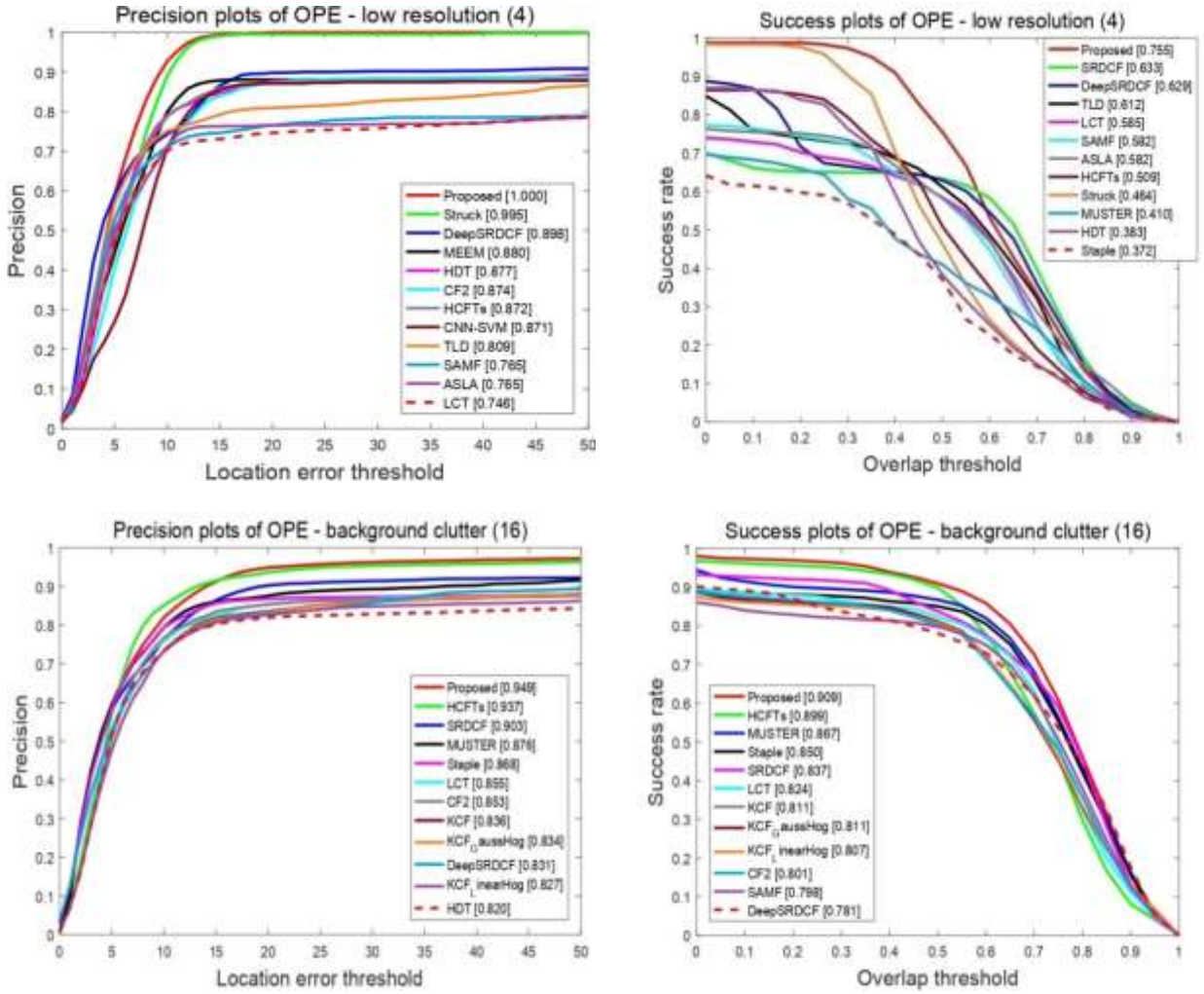
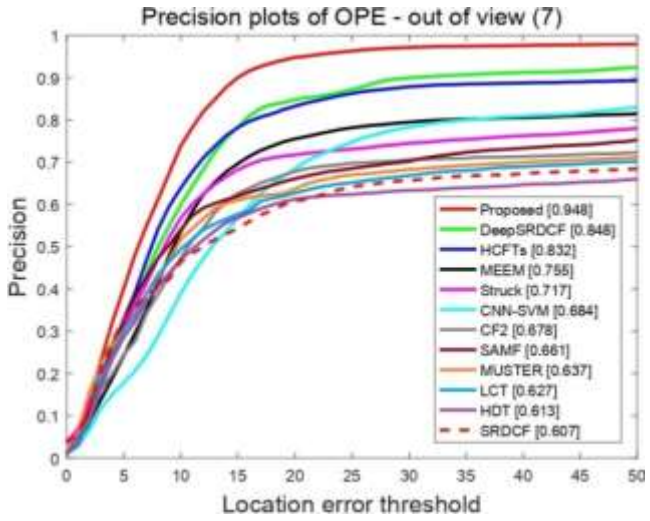
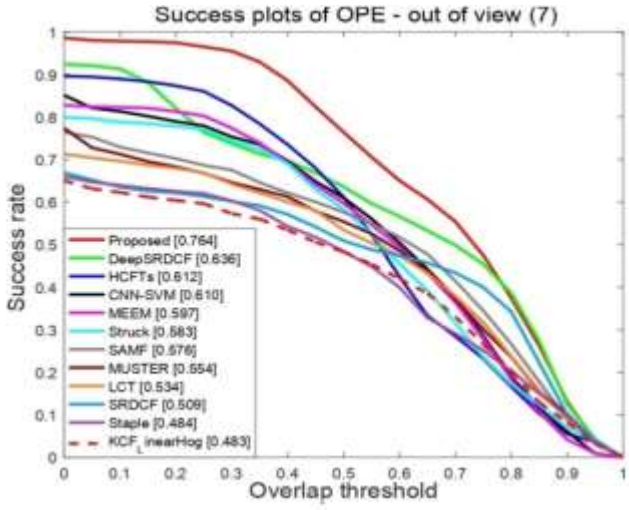
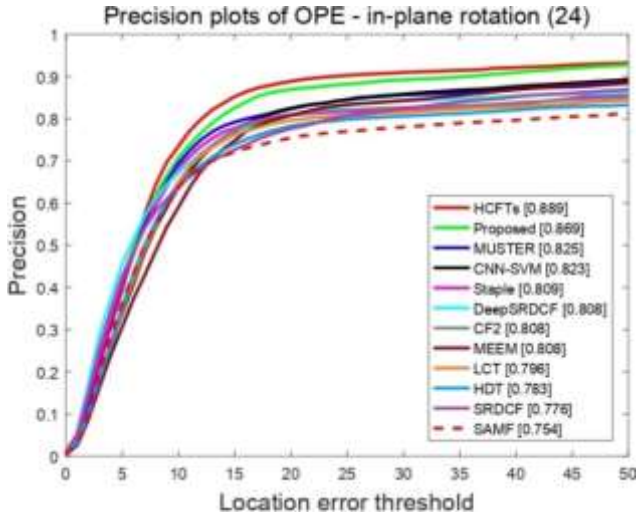


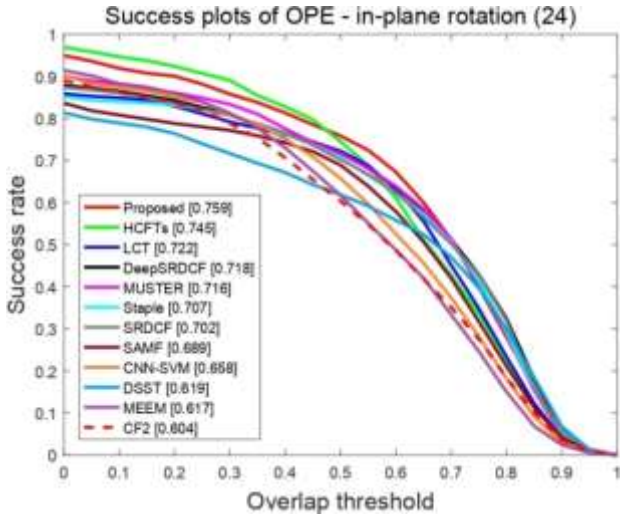
Fig.4. Sets itself apart from eleven cutting-edge trackers by utilizing metrics like distance precision and overlap success on the OTB-50 dataset, showcasing a notable contrast

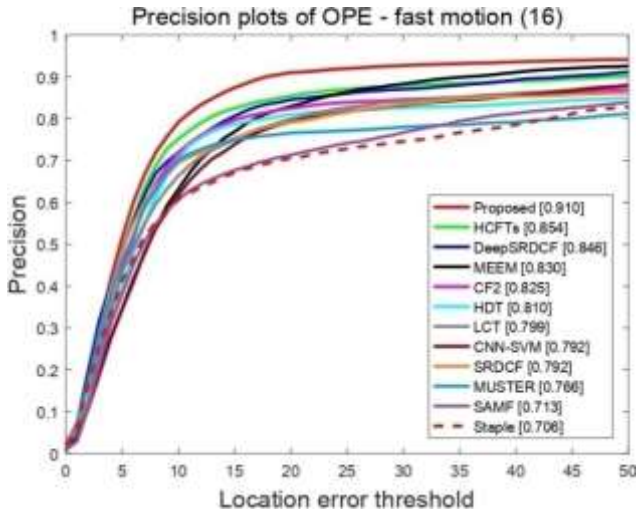


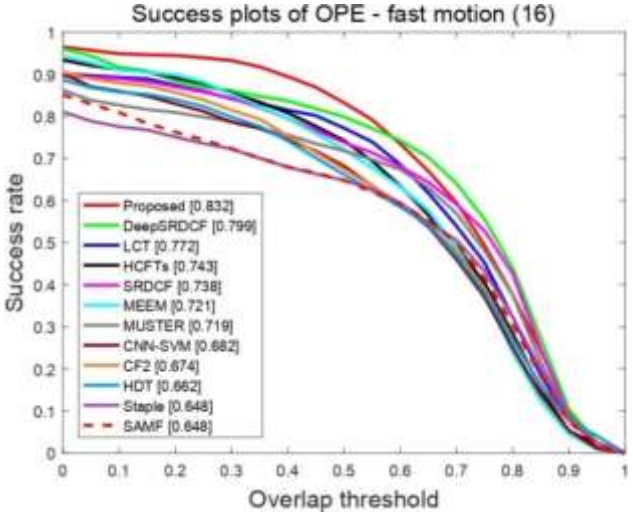

Precision plots of OPE - out of view (7)
Precision
Location error threshold
Proposed [0.948]
DeepSRDCF [0.848]
HCFTs [0.832]
MEEM [0.755]
Struck [0.717]
CNN-SVM [0.684]
CF2 [0.678]
SAMF [0.661]
MUSTER [0.637]
LCT [0.627]
HDT [0.613]
SRDCF [0.607]

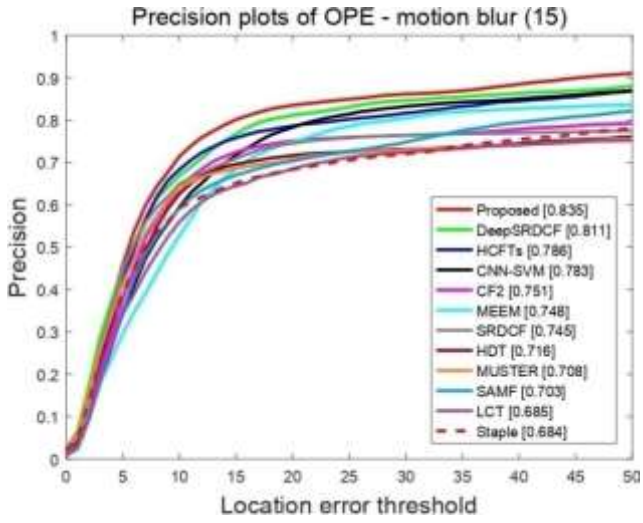

Success plots of OPE - out of view (7)
Success rate
Overlap threshold
Proposed [0.764]
DeepSRDCF [0.636]
HCFTs [0.612]
CNN-SVM [0.610]
MEEM [0.597]
Struck [0.583]
SAMF [0.576]
MUSTER [0.554]
LCT [0.534]
SRDCF [0.509]
Staple [0.484]
KCF_linearHog [0.483]

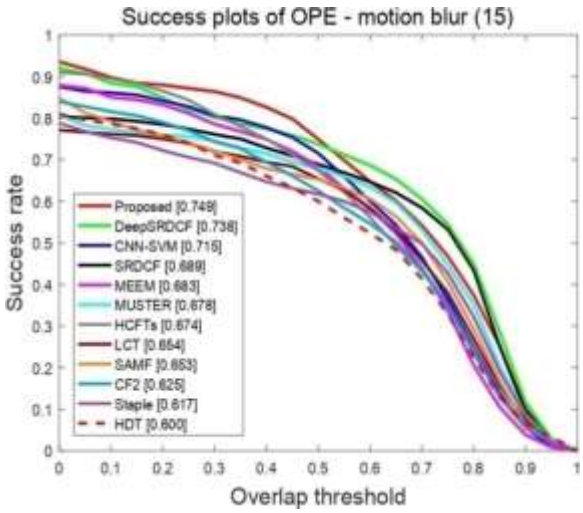

Precision plots of OPE - in-plane rotation (24)
Precision
Location error threshold
HCFTs [0.889]
Proposed [0.869]
MUSTER [0.825]
CNN-SVM [0.823]
Staple [0.809]
DeepSRDCF [0.808]
CF2 [0.808]
MEEM [0.806]
LCT [0.796]
HDT [0.783]
SRDCF [0.776]
SAMF [0.754]

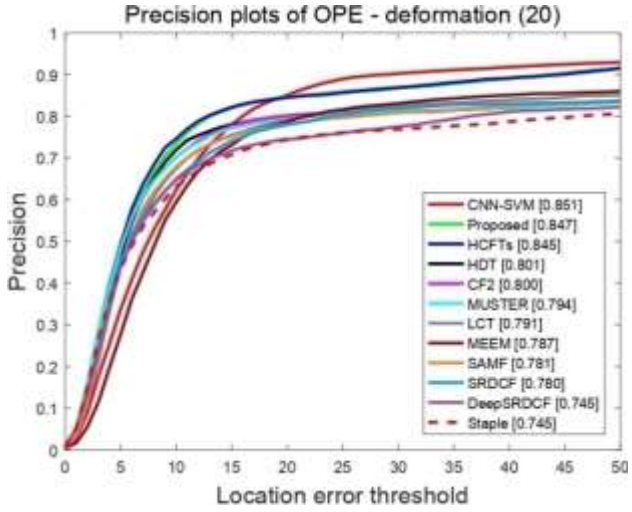

Success plots of OPE - in-plane rotation (24)
Success rate
Overlap threshold
Proposed [0.759]
HCFTs [0.745]
LCT [0.722]
DeepSRDCF [0.718]
MUSTER [0.716]
Staple [0.707]
SRDCF [0.702]
SAMF [0.689]
CNN-SVM [0.658]
DSST [0.619]
MEEM [0.617]
CF2 [0.604]

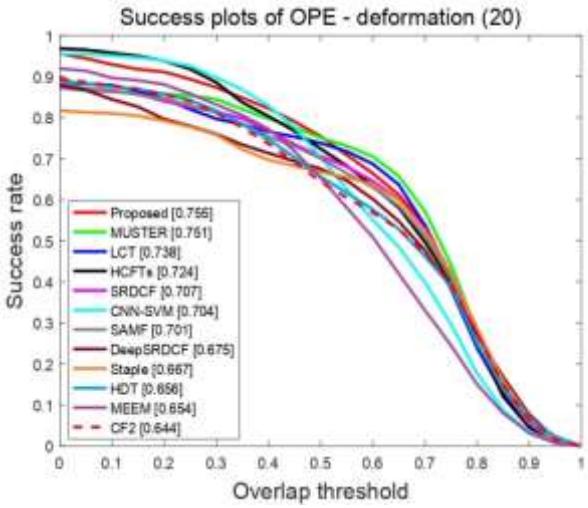

Precision plots of OPE - fast motion (16)
Precision
Location error threshold
Proposed [0.910]
HCFTs [0.854]
DeepSRDCF [0.846]
MEEM [0.830]
CF2 [0.825]
HDT [0.810]
LCT [0.799]
CNN-SVM [0.792]
SRDCF [0.792]
MUSTER [0.766]
SAMF [0.713]
Staple [0.706]

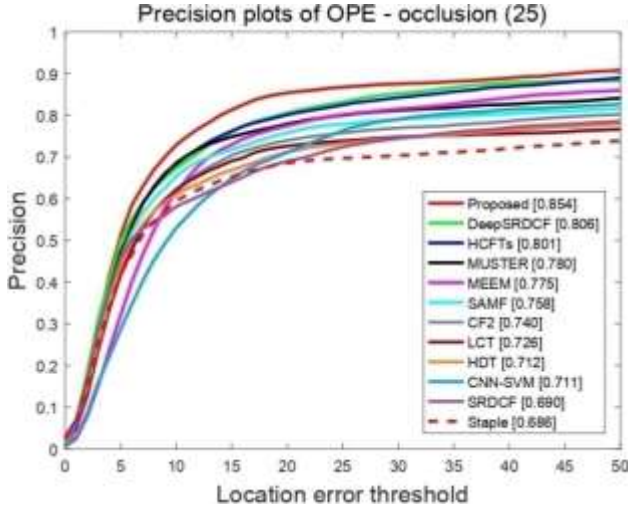

Success plots of OPE - fast motion (16)
Success rate
Overlap threshold
Proposed [0.832]
DeepSRDCF [0.799]
LCT [0.772]
HCFTs [0.743]
SRDCF [0.738]
MEEM [0.721]
MUSTER [0.719]
CNN-SVM [0.682]
CF2 [0.674]
HDT [0.662]
Staple [0.648]
SAMF [0.648]

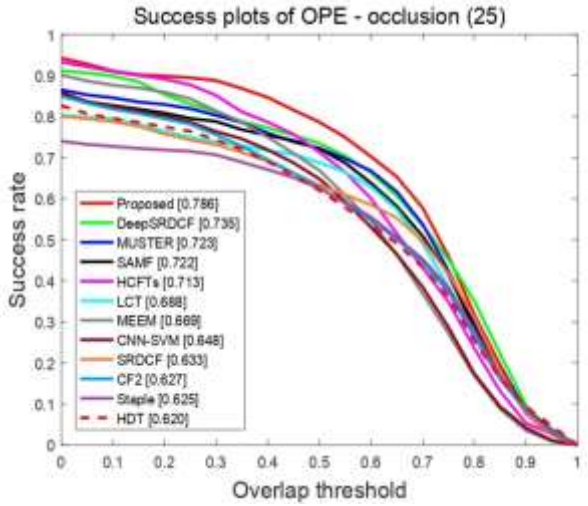

Precision plots of OPE - motion blur (15)
Precision
Location error threshold
Proposed [0.835]
DeepSRDCF [0.811]
HCFTs [0.786]
CNN-SVM [0.783]
CF2 [0.751]
MEEM [0.748]
SRDCF [0.745]
HDT [0.716]
MUSTER [0.708]
SAMF [0.703]
LCT [0.685]
Staple [0.684]


Success plots of OPE - motion blur (15)
Success rate
Overlap threshold
Proposed [0.749]
DeepSRDCF [0.738]
CNN-SVM [0.715]
SRDCF [0.689]
MEEM [0.683]
MUSTER [0.678]
HCFTs [0.674]
LCT [0.654]
SAMF [0.653]
CF2 [0.625]
Staple [0.617]
HDT [0.600]


Precision plots of OPE - deformation (20)
Precision
Location error threshold
CNN-SVM [0.851]
Proposed [0.847]
HCFTs [0.845]
HDT [0.801]
CF2 [0.800]
MUSTER [0.794]
LCT [0.791]
MEEM [0.787]
SAMF [0.781]
SRDCF [0.780]
DeepSRDCF [0.745]
Staple [0.745]


Success plots of OPE - deformation (20)
Success rate
Overlap threshold
Proposed [0.755]
MUSTER [0.751]
LCT [0.738]
HCFTs [0.724]
SRDCF [0.707]
CNN-SVM [0.704]
SAMF [0.701]
DeepSRDCF [0.675]
Staple [0.667]
HDT [0.656]
MEEM [0.654]
CF2 [0.644]


Precision plots of OPE - occlusion (25)
Precision
Location error threshold
Proposed [0.854]
DeepSRDCF [0.806]
HCFTs [0.801]
MUSTER [0.780]
MEEM [0.775]
SAMF [0.758]
CF2 [0.740]
LCT [0.726]
HDT [0.712]
CNN-SVM [0.711]
SRDCF [0.690]
Staple [0.686]


Success plots of OPE - occlusion (25)
Success rate
Overlap threshold
Proposed [0.786]
DeepSRDCF [0.735]
MUSTER [0.723]
SAMF [0.722]
HCFTs [0.713]
LCT [0.688]
MEEM [0.669]
CNN-SVM [0.648]
SRDCF [0.633]
CF2 [0.627]
Staple [0.625]
HDT [0.620]

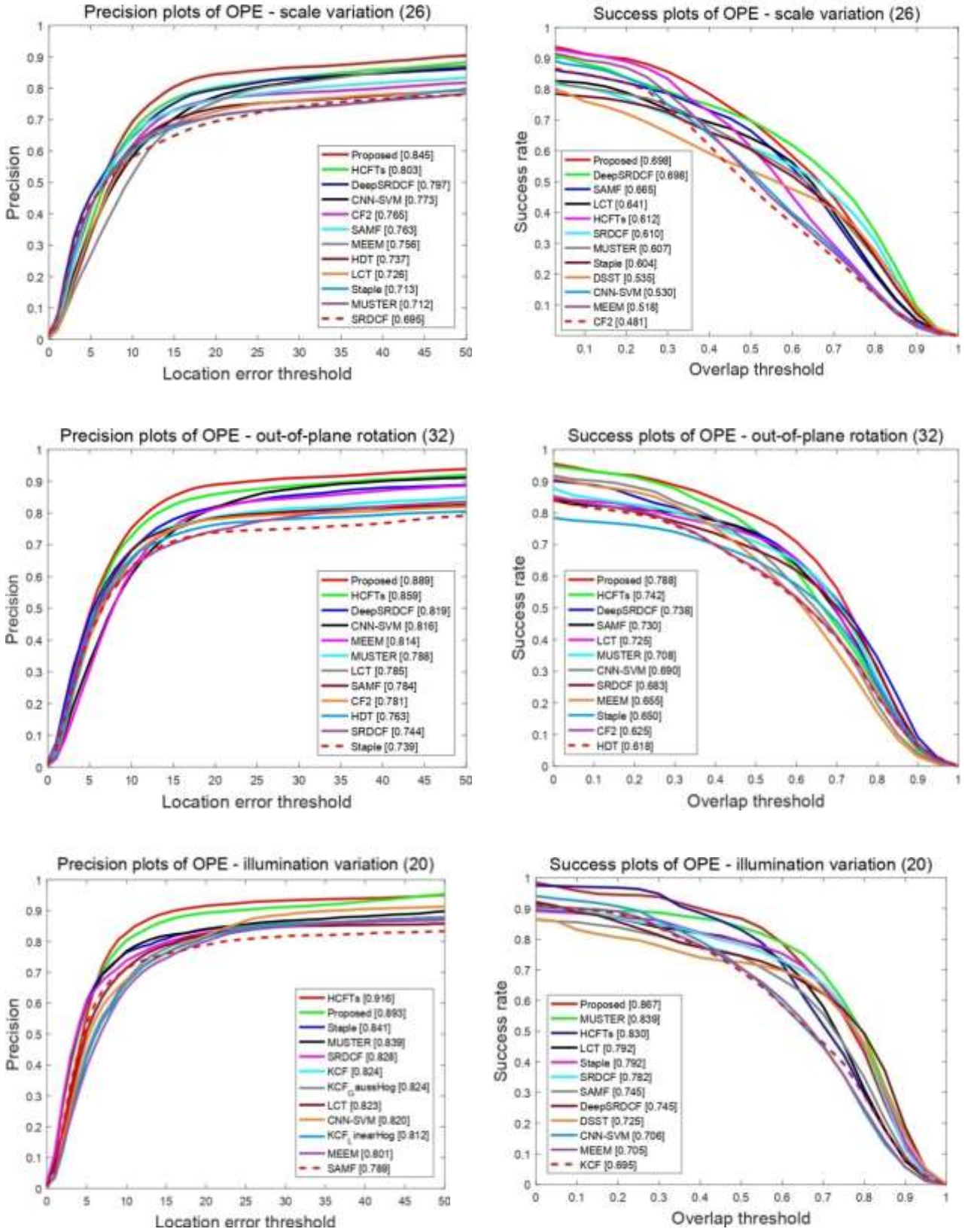


Fig. 5. The charts depict how well tracking performs in 11 different challenging scenarios, measured in terms of both overlap success and distance precision

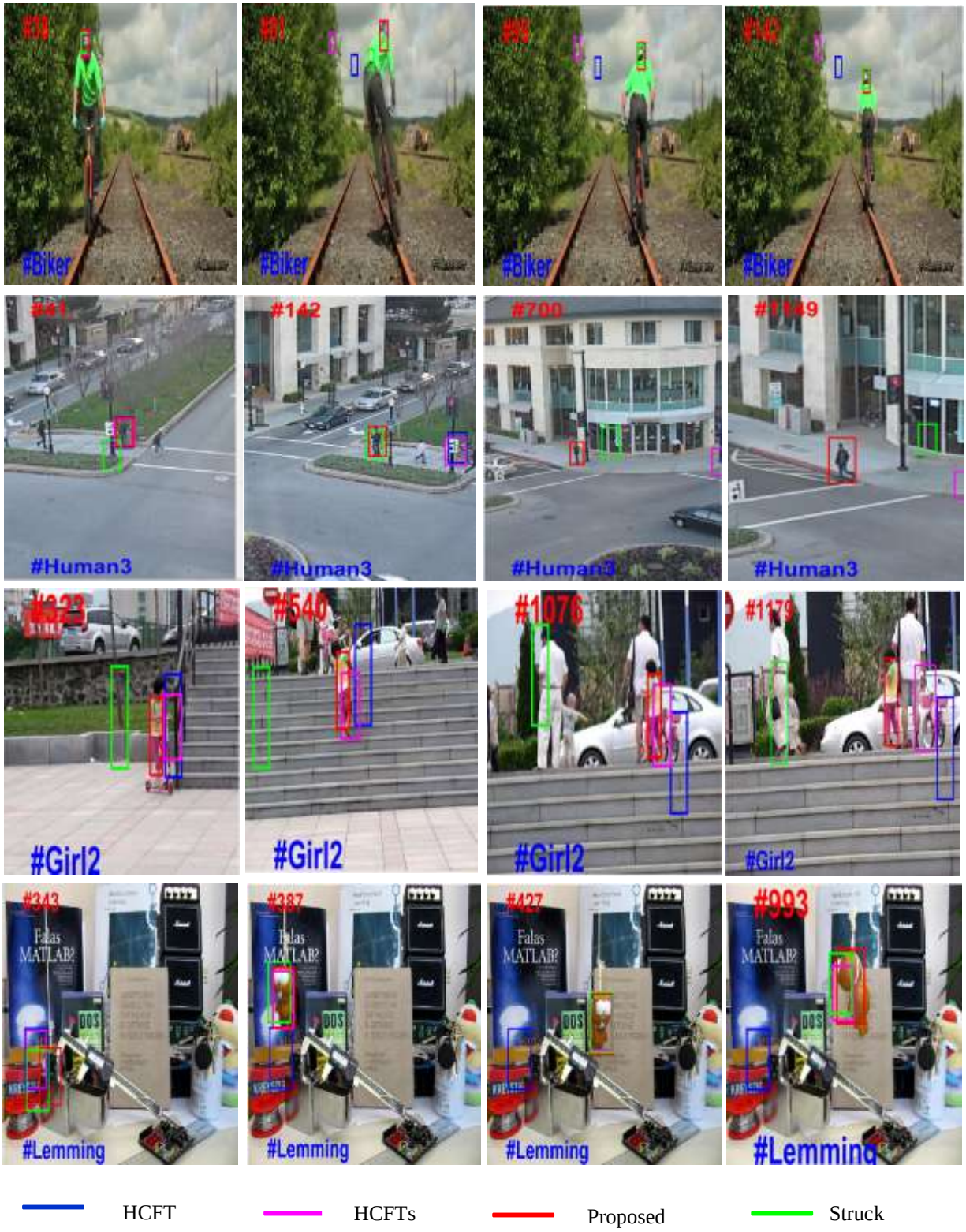


Fig. 6. showcases the qualitative results of our proposed method, along with HCFT [11, 33], HCFTs [32], and Struck [33], on four challenging sequences.

Figure 5 displays the overlap success rate plots and distance precision plots achieved using the OTB50 dataset, specifically focusing on 11 demanding scenarios. These scenarios involve various challenges like scale variation, fast motion, in-plane rotation, deformation, motion blur, occlusion, illumination variation, out-of-plane rotation, background clutter, out-of-view, and low resolution. Upon careful examination of all the sub-figures within Figure 5, it becomes evident that the proposed tracker surpasses its state-of-the-art counterparts in all aspects except for the deformation case.

Figure 6 presents a collection of tracking outcomes from a subset of the OTB-50 benchmark sequences. The aim is to assess the trackers' performance qualitatively, which includes HCFTs [11, 32, 33], Struck [33], and the newly proposed tracker. This evaluation is conducted on four demanding sequences, indicated by blue, magenta, green, and red markers. The sequences, namely Biker, Human3, Girl2, and Lemming, are arranged from the top to the bottom. Each of these sequences poses distinct challenges, encompassing scale variation, occlusion, motion blur, fast motion, out-of-plane rotation, low resolution, deformation, background clutter, and out-of-view scenarios.

The experimental findings indicate that Struck exhibits robust capabilities in effectively managing challenges like scale variation, occlusion, motion blur, and background clutter. This proficiency is particularly evident in sequences like Biker and Lemming. Nevertheless, Struck's efficacy weakens when confronted with deformation, a limitation that becomes apparent in sequences like Human3 and Girl2.

HCFTs excel in scenarios involving scale variation, occlusion, motion blur, and background clutter, as observed in sequences like Lemming. Nevertheless, it is less effective in handling deformation challenges, as witnessed in sequences Biker, Human3, and Girl2.

HCFT shows less effectiveness in dealing with all the mentioned sequences (Biker, Human3, Girl2, and Lemming).

In contrast, our proposed method consistently demonstrates accurate target tracking and outperforms HCFTs, HCFT, and Struck, especially in tracking small targets due to its enhanced robustness.

VI. CONCLUSION

In this paper, we have presented an enhanced visual object tracking algorithm that leverages a synergistic combination of features from CNN layers and Hog features, We have used also the coefficients of the DCT of the image instead of using an image template. The integration of DCT serves had two main goals:

Firstly, to calculate HOG features instead of using RGB. Secondly, to calculate CNN features for images with high saturation, thus improving their performance. We have incorporated Newton's method to improve the tracker's long-term memory and recovery from tracking failures.

Furthermore, we have introduced the HSV energy condition as an efficient approach to switch between RGB and HSV color bases, which improved convolution characteristics.

The proposed algorithm undergoes rigorous validation using the OTB50 datasets. Simulation results demonstrated that the proposed tracker outperforms numerous contemporary trackers, especially in challenging scenarios involving background clutters, motion blur, partial occlusions, and various appearance changes.

As a prospect work, we suggest to incorporate our proposed algorithm in real time applications .

DECLARATIONS

Ethical Approval

This declaration is not applicable.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Hanane NEBBAR. The first draft of the manuscript was written by Hanane NEBBAR and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

No funding was received to assist with the preparation of this manuscript.

Availability of data and materials

The data supporting the findings of this study are available in the [OTB50-100] or can be obtained upon reasonable request from the corresponding author.

REFERENCES

[1] J. Yang, W. Tang and Z. Ding, "Long-Term Target Tracking of UAVs Based on Kernelized Correlation Filte," *International Journal of Innovative Computing, Information and Control*. 2021, vol.9, no. 23, pp. 1--18.

[2] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Robust Visual Tracking via Hierarchical Convolutional Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019, vol. 41, no. 11, pp. 2709--2723.

[3] H. Grabner, C. Leistner, H. Bischof, Semi-supervised on-line boosting for robust tracking, In: Forsyth, D., Torr, P., Zisserman, A. (eds) *Computer Vision ECCV 2008. ECCV 2008. Lecture Notes in Computer Science*. 2008, vol. 5302, Springer, Berlin, Heidelberg.

[4] J. Zhang, J. Sun, J. Wang et al., "Visual object tracking based on residual network and cascaded correlation filters," *Springer, Journal of Ambient Intelligence and Humanized Computing*. 2021, vol. 12, pp. 8427--8440 .

[5] J. Fan, W. Xu, Y. Wu, Y. Gong, "Human tracking using convolutional neural networks," *Transactions on Neural Networks.* 2010, vol. 21, no. 10, pp. 1610--1623.

[6] L. Yang, C. Kong, X. Chang et al, "Correlation filters with adaptive convolution response fusion for object tracking, " *Knowledge-Based Systems*. 2021, vol. 228, 107314.

[7] B. Babenko, M. -H. Yang and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2011, vol. 33, no. 08, pp. 1619--1632.

[8] C. Huang, S. Lucey and D. Ramanan, Learning policies for adaptive tracking with deep feature cascades, *IEEE International Conference on Computer Vision (ICCV).* Venice, Italy, 2017, pp. 105--114.

[9] N. Wang and D.-Y. Yeung, Learning a Deep Compact Image Representation for Visual Tracking, *Advances in Neural Information Processing Systems* 26 (NIPS 2013). 2013.

[10] A. He, C. Luo, X. Tian et al., A twofold siamese network for real-time object tracking, *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Salt Lake City, UT, USA, 2018, pp. 4834--4843.

[11] C. Ma, J. -B. Huang, X. Yang et al., Hierarchical Convolutional Features for Visual Tracking, *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, 2015, pp. 3074--3082.

[12] M. Y. Abbass, K. C. Kwon, N. Kim et al., "Efficient Object Tracking Using Hierarchical Convolutional Features Model and Correlation Filters," *Springer-Verlag, The Visual Computer*, 2021, vol. 37, no. 04, pp. 831--842.

[13] B. Latreche, S. Saadi, M. Kious et al., "A novel hybrid image fusion method based on integer lifting wavelet and discrete cosine transformer for visual sensor networks," *Springer, Multimedia Tools and Applications*, 2019, vol. 78, pp. 10865--10887.

[14] H. Chen, W. Zhang, X. Zhao et al., DCT representations based appearance model for visual tracking, *IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*. Bali, Indonesia, 2014, pp. 1614--1619.

[15] D. He, Z. Gu and N. Cercone, Efficient image retrieval in DCT domain by hypothesis testing, *16th IEEE International Conference on Image Processing (ICIP)*. Cairo, Egypt, 2019, pp. 225--228.

[16] M. Uzair, A. Mahmood, A. Mian, Hyperspectral Face Recognition using 3D-DCT and Partial Least Squares, *In Proceedings of the British Machine Vision Conference*. BMVA, 2013, vol. 78, pp. 10pp.

[17] D. Chen, Q. Liu, M. Sun and J. Yang, "Mining Appearance Models Directly From Compressed Video," *IEEE Transactions on Multimedia*, 2008, vol. 10, no. 02, pp. 268--276.

[18] B. K. Shreyamsha Kumar, M. N. S. Swamy and M. Omair Ahmad, "Visual tracking using structural local DCT sparse appearance model with occlusion detection," *Springer, MMultimedia Tools and Applications*, 2019, vol. 78, pp. 7243--7266.

[19] W. -h. Yun, D. Kim, B. Song et al., Block comparison based face identification using HOG feature, *The 18th IEEE International Symposium on Robot and Human Interactive Communication*, Toyama, Japan, 2009, pp. 484--487.

[20] Y. Wei, Q. Tian and T. Guo, "An Improved Pedestrian Detection Algorithm Integrating Haar-Like Features and HOG Descriptors," *Advances in Mechanical Engineering*, 2013, vol. 05, pp. 484--487.

[21] Y. Li, Y. Zhang, Y. Xu et al., "Robust scale adaptive kernel correlation filter tracker with hierarchical convolutional features," *IEEE Signal Processing Letters*, 2016, vol. 23, no. 08, pp. 1136--1140.

[22] J. T. Mbelwa, Q. Zhao and F. Wang, "Visual tracking tracker via object proposals and co-trained kernelized correlation filters," *The Visual Computer*. 2020, vol. 36, pp. 1173--1187.

[23] M. Y. Abass, KC. Kwon, N. Kim, et al. "A survey on online learning for visual tracking," *The Visual Computer*. 2021, vol. 37, pp. 993--1014.

[24] M. Danelljan, G. Hager, F. S. Khan et al., Learning Spatially Regularized Correlation Filters for Visual Tracking, *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, 2015, pp. 4310--4318.

[25] C. Ma, J.-B. Huang, X. Yang et al., "Adaptive Correlation Filters with Long-Term and Short Term Memory for Object Tracking," *International Journal of Computer Vision*, 2017, vol. 126, pp. 771796.

[26] V. N. Boddeti, T. Kanade and B. V. K. V. Kumar, Correlation filters for object alignment, *IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA, 2013, pp. 2291—2298.

[27] H. K. Galoogahi, T. Sim and S. Lucey, Correlation filters with limited boundaries, *In CVPR2015*. pp. 4630--4638 (2015).

[28] W. Wei, X.-L. Yang, B. Zhou et al., "Combined energy minimization for image reconstruction from few views," *Hindawi, Mathematical*

Problems in Engineering, 2012, vol. 2012, pp. 1--15.

[29] W. Wei, Z. Sun, H. Song et al., "Energy balance-based steerable arguments coverage method in WSNs," *IEEE Access*. 2018, vol. 06, pp. 33766--33773.

[30] A. Abdel-Hadi, Real-time object tracking using color-based Kalman particle filter, *The 2010 International Conference on Computer Engineering & Systems*. Cairo, Egypt 2010, pp. 337--341.

[31] J. Deng, W. Dong, R. Socher et al., ImageNet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA, 2009, pp. 248-255.

[32] C. Ma, J. -B. Huang, X. Yang et al., "Robust Visual Tracking via Hierarchical Convolutional Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, vol. 41, no. 11, pp. 2709--2723.

[33] S. Hare et al., "Struck: Structured Output Tracking with Kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, vol. 38, no. 10, pp. 2096--2109.