An Enhanced Visual Object Tracking Approach based on Combined Features of Neural Networks, Wavelet Transforms, and Histogram of Oriented Gradients

Mohammed Bourennane
Department of Electrical Engineering, University of Biskra, Biskra, Algeria
bourennane@gmail.com

Nadjiba Terki
LESIA Laboratory of Research, Department of Electrical Engineering, University of Biskra, Biskra, Algeria
n.terki@univ-biskra.dz

Madina Hamiane
College of Engineering, Royal University for Women, Bahrain
mhamiane@ruw.edu.bh

Abdallah Kouzou
LAADI Laboratory, Faculty of Science and Technology, University of Djelfa, Djelfa, Algeria and
Electrical and Electronics Engineering Departement, Nisantasi University, Istanbul, Turkey
a.kouzou@uni-djelfa.dz

Received: 28 April 2022 | Revised: 7 May 2022 | Accepted: 12 May 2022

Abstract- In this paper, a new Visual Object Tracking (VOT) approach is proposed to overcome the main problem the existing approaches encounter, i.e. the significant appearance changes which are mainly caused by heavy occlusion and illumination variation. The proposed approach is based on a combination of Deep Convolutional Neural Networks (DCNNs), Histogram of Oriented Gradient (HOG) features, and discrete wavelet packet transforms. The problem of illumination variation is solved by incorporating the coefficients of the image discrete wavelet packet transform instead of the image template to handle the case of images with high saturation in the input of the used CNN, whereas the inverse discrete wavelet packet transforms are used at the output for extracting the CNN features. By combining four learned correlation filters with the convolutional features, the target location is deduced using multichannel correlation maps at the CNN output. On the other side, the maximum value of the resulting maps from the correlation filters with convolutional features produced by the previously obtained HOG feature of the image template are calculated and are used as an updating parameter of the correlation filters extracted from CNN and from HOG. The major aim is to ensure long-term memory of the target appearance so that the target item may be recovered if tracking fails. In order to increase the performance of HOG, the coefficients of the discrete packet wavelet transform are employed instead of the image template. The obtained results demonstrate the superiority of the proposed approach.

Keywords-visual tracking; deep convolution neural networks; wavelet transform; HOG features

I. INTRODUCTION

Visual Object Tracking (VOT) is becoming a very active area of research, attracting much attention due to its importance within numerous applications such as unmanned control systems, motion analysis, and video processing [1-5]. VOT is basically used to estimate an unknown visual target trajectory based on a known initial starting state of the considered target. However, the visual tracking remains a difficult problem to be solved accurately despite the efforts in this area during the last decade, due to the new challenges that have been induced by new technology evolution and which make the target objects often experience important changes in their appearance, such as the scale variation, fast motion, in-plane rotation, deformation, motion blur, occlusion, illumination variation, out-of-plane rotation, background clutter, etc.

Convolutional Neural Network (CNN) features have been recently put to use in a variety of computer vision applications [6-8], e.g. object identification, image segmentation, and image classification [9]. The effective use of the rich hierarchical features of CNNs in visual tracking, has brought significant improvement. It has been proved that the convolutional layers have the ability of ensuring the presentation of the invariant features against the variation of the target appearance which can be very useful in visual object tracking applications. Unfortunately, it has been found that the CNNs have a major limitation resulting from the fact that they are built based on the principles of other visual classification tasks [10]. Authors in [11] proposed the exploitation of the rich hierarchical features of the Deep Convolutional Neural Networks (DCNNs) to ensure enhanced accuracy and robustness of visual object tracking. It has been proposed that, in order to ensure invariant feature representation with respect to significant variations in the target appearance, the outputs of the last convolutional layers should be used to encode the semantic information of the target. This technique faced the problem of losing the precise

Corresponding author: Mohammed Bourennane

localization of the target due to large spatial resolution, it has even been resistant to target significant appearance changes. In the same time, it has been observed that the features of the earlier convolutional layers can ensure precise localization of targets but in the same time they are less sensitive to target appearance changes. Based on the concept that various layers in a CNN model give varying levels of information in describing an object [2-4], some authors have attempted to address this issue by combining feature representations provided by different CNN layers with correlation filters to realize efficient tracking performance [2-5]. Despite the achieved advantages of these techniques, compared to those based on CNN, this proposal presented some limitations. It has been based essentially on the learning and updating of the correlation filters in the frequency to overcome the problem of appearance variations. This led to unwanted boundary effects and important degradation of the tracking model quality. Furthermore, it has been found that it cannot be effective for long time tracking and it cannot ensure the detection of the target position failures. In the same time, it has faced a major problem against changes of illumination within specific color sequences [13]. Authors in [14] proposed an efficient hybrid image fusion method based on the Integer Lifting Wavelet Transform (ILWT) and the Discrete Cosine Transform (DCT) to generate fused images with high visual quality which can be used to reduce some visual tracking problems.

Traditional signal processing methodologies, such as multi-resolution analysis utilizing wavelets, have been thoroughly investigated, allowing them to be more interpretable than CNNs. In fact, there have been several prior works, which have incorporated wavelet representations into CNNs [15]. Authors in [16] proposed Wavelet CNNs (WCNNs) and demonstrated how to generalize filtering and downsampling by reformulating convolution and pooling layers. Authors in [17] presented a CNN that is similar to the dense convolutional network (DenseNet). Haar wavelets were employed as convolution and pooling layers, which are often used in multi-resolution analysis. In this approach, the feature maps generated by the subsequent convolutional blocks have been concatenated with these wavelet layers. Authors in [18] used the Dual-Tree Complex Wavelet Transformation (DTCWT) in addition to WCNNs to solve the organ tissue image segmentation problem, whereas, by moving activation layers into wavelet space, authors in [19] employed a new concept of learning filters based on activations in the domain of wavelet. Authors in [15] describe the Deep Adaptive Wavelet Network (DAWN) architecture, which employs a combination of the lifting technique and CNNs to learn features via multi-resolution analysis. The DAWN algorithm is designed to obtain a wavelet representation of the input at each decomposition level. This contrasts with the black-box nature of CNNs. The DAWN architecture, unlike standard wavelets, is data-driven and adapts to the input pictures. Additionally, it is trainable from start to finish and achieves cutting-edge texture classification with a small set of trainable parameters.

During the recent years, many detection and classification problems have been solved with the Histogram of Oriented Gradient (HOG) features [20-22]. Local forms of the designated object are captured exactly by the HOG. Two sets of feature descriptors are commonly used [20]. This combination enhances detection efficiency but increases feature dimensions and computational complexity. Authors in [23] proposed a Haar-HOG-based approach which has shown better performance in terms of speed and efficiency than the algorithms based only on separate use of the Haar-like feature or the HOG descriptor. The proposed Haar-HOG algorithm has been found to be more accurate than the Haar-like features-based algorithm. On the other side, compared with algorithms using HOG descriptor only, this algorithm has a higher detection rate and a reduced false positive rate. The main contributions of this article can be summarized as:

- The wavelet decomposition based on different frequency sub-bands such as Low-Low (LL), Low-High (LH), High-Low (HL), and High-High (HH), have been used instead of RGB (Red-Green-Blue) image to resolve the problem of illumination variation in such cases when the saturation exceeds $\frac{2}{3} \times 100\%$.
- Based on the importance of combining feature representations from different CNN layers [3, 4, 24], a model of Hierarchical Convolutional Filters (HCF) is proposed. The proposed model is composed of different convolutional layers (conv1-4, conv3-4, conv4-4, and conv5-4).
- Wavelet decomposition, made up of four layers [LL, LH, HL, HH] instead of using the original image, has been used to improve the performance of the extracted HOG features.
- In addition, an update control approach has been designed to allow the appearance changes identification while preventing model drift. This has been carried out by calculating the maximum value of the resulting maps from correlation filters with convolutional feature products of HOG features for the image template that has been previously obtained, and which has been used as a parameter to the updating of the correlation filters.
- For the evaluation of the proposed approach, the large-scale benchmark datasets OTB50 with 50 challenging image sequences and OTB100 with 100 challenging image sequences have been used.

II. THE PROPOSED ALGORITHM

The main aim of the algorithm proposed in this paper is to present a new contribution that can overcome the main difficulties encountered in visual tracking most of the previous proposed approaches face under target appearance changes such as severe occlusion and illumination variation. In this section, the proposed algorithm is described in detail. The different stages of the proposed tracking algorithm are shown in Figure 1. Firstly, based on [3, 4], the target location is estimated by learning 4 two-dimensional correlation filters with CNN features. Secondly, according to the properties of the input image, the selection of the use of RGB or GRAY with the wavelet decomposition is carried out.

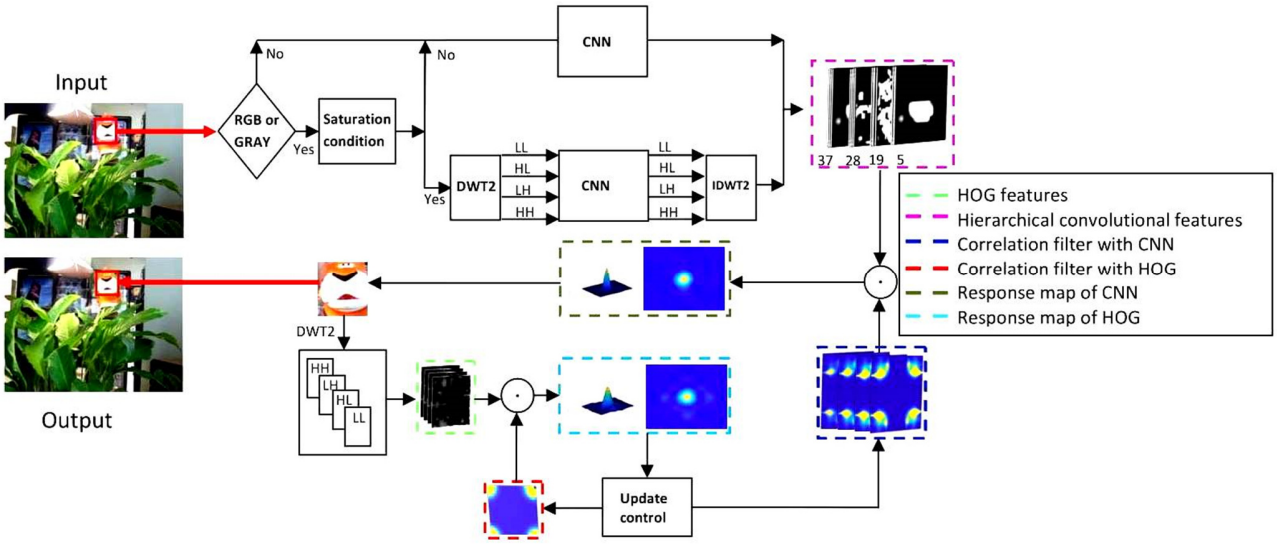

Fig. 1. Main stages of the proposed algorithm.

Thirdly, the maximum value of the resulting maps is calculated from the correlation filters with convolutional feature products of HOG features based on the previously obtained image template. This calculated value is used as a parameter to update the correlation filters.

A. Convolution Features

Due to the very interesting properties of CNNs in ensuring accurate separation between the object and its background, they have been used largely to improve many aspects of computer vision. We present the translation estimation with the creation of a translation model by form extraction using a CNN model. VGGNet-19 [25] feature extractor trained on the ImageNet dataset [26] is used to encode the target appearance. While the characteristics propagate to deeper layers, the spatial resolution progressively decreases, but semantic discrimination between objects belonging to various categories is enhanced. The determination of the exact target item position in visual object tracking is more relevant than semantic category. Bilinear interpolation [3] is used to resize each input frame to 224×224 size.

Firstly, the fully connected layers are removed and the outputs of the convolution layers conv1-4, conv3-4, conv4-4, and conv5-4 are used as deep features. In addition, a cosine window to weight each feature channel is used to eliminate the boundary discontinuities [16, 27]. When the CNN depth increases, the spatial resolution of a target object decreases progressively because of the pooling processes. By using the bilinear interpolation given in (1), each feature map is also rescaled to size $\frac{M}{4}\times\frac{N}{4}$, where M and N are the dimensions of the feature vector x, to correct the spatial resolution across the pooling layers.

$$x_i = \sum_k \alpha_{ik} h_k \quad (1)$$

B. Correlation Filters

Usually a correlation tracker search for the maximum value on the response in a discriminative classifier to locate target objects is used [28-30]. In this research, all convolutional layer outputs are used as multi-channel features [31, 32]. We assume that x is the l^{th} layer of a feature vector of size $M \times N \times D$, where M, N, and D are the image width, image height, and the number of channels respectively. We ignore the dependence of M, N, and D on the layer index l and note $x^{(l)}$directly as x. All the circular shifts of x along the M and N dimensions are taken as training samples. A Gaussian function label: $y(m,n) = e^{-\frac{(m-M/2)^2+(n-N/2)^2}{2\sigma^2}}$, where σ is the kernel width, is attributed to each shifted sample $x_{(m,n)}(m,n) \in \{0,1,\dots,M-1\} \times \{0,1,\dots,N-1\}$. By solving the minimization problem (2), a correlation filter W with the same size of x is trained.

$$W^* = \text{argmin}_W \sum_{m,n} \left\| W \cdot x_{m,n} - y(m,n) \right\|^2 + \lambda \|W\|^2 \quad (2)$$

where λ is a regularization parameter ($\lambda \geq 0$).

Linear kernel in a Hilbert space is used to induce the inner product in (2), i.e. $W \cdot x_{m,n} = \sum_{d=1}^{D} W_{m,n,d}^T x_{m,n,d}$. As the label $y(m,n)$ is soft (not binary), so no hard-threshold sample is required. The minimization problem in (2) could be solved in each individual feature channel using Fast Fourier Transformation (FFT), since it's similar to training the vector correlation filters [33]. The capital letters denote the Fourier transformed signals. In the frequency domain, the learned filter on the d^{th} ($d \in \{1,\dots,D\}$)channel is given in (3):

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

where Y is the Fourier transformation form of $y = \{y(m,n) | (m,n)\{0,1,\dots,M-1\} \times \{0,1,\dots,N-1\}\}$ and the bar refers to the complex conjugation. The operator $\odot$ is the Hadamard product. For each image patch in the next frame the l^{th} layer feature vector is noted as z with size $M \times N \times D$. The l^{th} correlation response map is given by:

$$f_l = \mathcal{F}^{-1}(\sum_{d=1}^{D} W^d \odot \bar{Z}^d) \quad (4)$$

The operator $\mathcal{F}^{-1}$ refers to the inverse FFT transform, the position of maximum value of the correlation response map f_l of size $M \times N$ refers to the target location on the l^{th} convolution layer.

C. Estimation of Coarse-to-Fine Translation

For each set of correlation response maps $\{f_l\}$, we can deduce the target translation of each layer, i.e. to search for the maximum value of the earlier $(l\text{-}1)^{\text{th}}$ layer, the location of maximum value in the last, l, layer is taken as regularization. If $(\hat{m}, \hat{n}) = \underset{m,n}{\operatorname{argmax}} f_l(m,n)$ refers to the location of the maximum values on the l^{th} layer, then the optimal location of the target in the $(l\text{-}1)^{\text{th}}$ layer is given by:

$$\underset{m,n}{\operatorname{argmax}} f_{l-1}(m,n) + \gamma f_l(m,n) \quad (5)$$

where $|m - \hat{m}| + |n - \hat{n}| \leq r$.

The constraints imposed on m and n limit the searched area in the $(l\text{-}1)^{\text{th}}$ correlation response map to the $r \times r$ neighboring regions of $(\hat{m}, \hat{n})$. From the last to the inner layers, each response value is weighted by a regularization term γ and propagated back to the response maps of the earlier layers [3, 4]. Finally, by maximizing the result of (5) on the layer with the best spatial resolution, the target location is estimated. On the other hand, using (2), (3), and (5), we can calculate the maximum response of the correlation filter of HOG considering l =1, and γ =1.

D. Model Update

In this work, the correlations filters are updated as proposed in [33]. Initially, we update the numerator A^d and denominator B^d of the correlation filter W^d in (3), separately using a moving average:

$$A_t^d = (1-\eta)A_{t-1}^d + \eta\, Y \odot \bar{X}_t^d \quad (6a)$$

$$B_t^d = (1-\eta)B_{t-1}^d + \eta \sum_{i=1}^{D} X_t^i \odot \bar{X}_t^i \quad (6b)$$

$$W_t^d = \frac{A_t^d}{B_t^d + \lambda} \quad (6c)$$

We update the correlation filters extracted from CNN and HOG features conservatively, because the conservatively learned filter is robust to noisy updates and succeeds in estimating the confidence of every tracked result. To see if tracking failures occur, we establish a T_0 threshold. If the maximum filter response of the correlation filter of HOG is greater than T_0, the tracked result z has a very high degree of confidence. In that case, we update the correlation filters.

III. DISCRET WAVELET PACKET TRANSFORMS

Two-Dimensional Discrete Wavelet Transform (2D-DWT) can be used to decompose an image into sub-signals, which present the original image components (I) under different frequency ranges. 2D-DWT is used in the input side to ensure the process of splitting the original image (I) into 4 sub images (ILL, IHL, ILH, and IHH). At first, two down-sampling filters (noted as ↓2) of low (L) and high (H) bands are used yielding to two rows (IL and IH). Then each obtained images in both rows passes through two filters of low (L) and high (H) down-sampling bands which means 4 filters are used in this phase to obtain 4 sub-images as 2 columns: the first column is (ILL, IHL) and the second column is (ILH, IHH). ILL presents the approximation coefficient matrix which is obtained from the passage through 2 simultaneous low-pass filters and the other 3 present the detail coefficients matrices IHL (horizontal features), ILH (vertical features), and IHH (diagonal features), as shown in Figure 2. Moreover, the 2-D DWT has a separable characteristic with the scaling function $\Phi_{\text{LL}}(x,y)$ and three 2D-wavelets, $\Psi_{\text{HL}}(x,y)$, $\Psi_{\text{LH}}(x,y)$, and $\Psi_{\text{HH}}(x,y)$, which can be expressed as follows [34]:

$$\Phi_{\text{LL}}(x,y) = \Phi(x)\Phi(y) \quad (7)$$

$$\Psi_{\text{HL}}(x,y) = \Psi(x)\Phi(y) \quad (8)$$

$$\Psi_{\text{LH}}(x,y) = \Phi(x)\Psi(y) \quad (9)$$

$$\Psi_{\text{HH}}(x,y) = \Psi(x)\Psi(y) \quad (10)$$

where $\Phi(x)$ and $\Phi(y)$ are the wavelet functions following the x-axis (horizontal) and the y-axis (vertical), $\Psi(x)$ and $\Psi(y)$ are the horizontal and vertical 1D scaling functions.

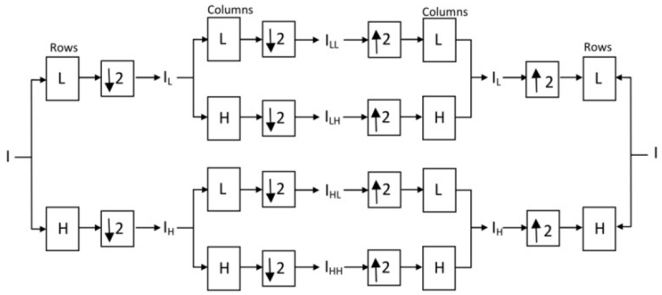

Fig. 2. Downsample and upsample comparison of DWT and IDWT.

In contrast, the inverse DWT (IDWT) is used in the output for adding 4 sub-images to the original one using up-sampling filters (noted as ↑2) with the same concept as the DWT but with the inverse operation as shown in Figure 2. It is clear that the inputs are the 4 sub-images (ILL, IHL, ILH and IHH) and the output is the filtered original image (I).

IV. SATURATION CONDITION

In the proposed approach, the illumination variation has been handled in a reliable way based on a new proposed concept. The main idea of the proposed concept is based on integrating the wavelet decomposition obtained from the DWT [ILL, ILH, IHL, IHH] when the saturation of the image is very high instead of using the image components (RGB) directly in the network. The saturation power state can be computed for each input frame according to the following steps:

- First step: the conversion of the red, green, and blue values of an RGB image to hue (H), saturation (S), and value (V) values of an HSV image.
- Second step: the calculation of the saturation energy according to [13]:

$$E_s = 100 \times \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(S_{ij})^2}{E_T} \quad (11)$$

$$E_T = \sum_{i=1}^{m}\sum_{j=1}^{n}(H_{ij})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(S_{ij})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(V_{ij})^2 \quad (12)$$

where E_S refers the saturation energy and E_T refers to the total energy of the input frame.

- Third step: if $E_s > \frac{2}{3}\times 100$, then the illumination is very weak. In this case, the wavelet decomposition [LL, LH, HL, HH] is used in the input of CNN. In the opposite case, the RGB decomposition of the input image is used. This step is carried out following the proposed approach.

The combined DWT and CNN method is found to be robust and therefore it is capable of alleviating the problem of illumination variation. As an example, Table I presents the percentage of energy saturation of the Singer2 sequence calculated along 6 frames following (11).

TABLE I. THE PERCENTAGE OF ENERGY SATURATION IN THE SINGER2 SEQUENCE

Frame	**6**	8	10	**12**	14	16	**18**
Energy saturation (%)	**76.58**	76.30	76.82	**25.16**	71.62	73.39	**72.03**

It is clear from Table I that the energy saturation varies from a frame to another between the minimum value of 25.16% corresponding to frame 12 and the maximum value of 76.82% corresponding to frame 10. Based on the condition mentioned in the third step, it can be observed that energy saturation is low only in the case of frame 12 and the required condition is not satisfied. In this case the RGB approach is used. It is obvious that for the other frames this required condition is satisfied, hence the wavelet decomposition [LL, LH, HL, HH] is used in the input of CNNs in these frames. It can be concluded that under the application of the proposed approach, the case of illumination variation can be handled more accurately based on the beneficial features of 2D-DWD. Figure 3 illustrates the accurate placement of the target within the 6 chosen frames of the Singer2 sequence. The blue frame corresponds to the initial position of the tracked object and the red frame to the used tracker based on the proposed approach combining the wavelet and the CNNs. It is obvious that the proposed approach allows robust tracking of the moving object under illumination variations and in the same time it maintains the long-term memory of target appearance which ensures a high degree of accuracy in locating the target in the majority of the frames of the sequence. On the other hand, for the validation of the proposed approach based on the proposed saturation condition, two tests have been carried out based on the calculation of the error tracking in both cases, with (red) and without (blue) the saturation condition. It is clearly observed in Figure 3 that the tracking error obtained under the proposed approach is minimized to a very low value compared to the standard case in which the saturation condition variation is not taken into account.

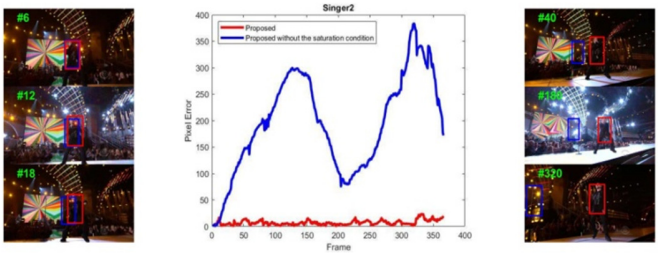

Fig. 3. A frame-by-frame display of the results of the Singer2 sequence tracking, with and without the saturation condition (in pixels).

V. EXPERIMENTAL RESULTS

The proposed algorithm has been validated and evaluated on two benchmark datasets, OTB50 [35], which includes 50 videos and OTB100 [35], which includes 100 videos. The tracking algorithm has been implemented in MATLAB on an Intel I7-8750H 2.20GHz CPU with 16GB RAM and the MatConvNet toolbox [36], while the feature extraction using CNN forward propagation has been carried out on a GeForce GTX1060 GPU. The CNN-based VGG-Net-19, consisting of 19 layers (16 convolution layers, 3 fully connected layers, 5 MaxPool layers, and 1 SoftMax layer) [25], has been trained on the free large-scale hierarchical image database ImageNet [26] and was adopted for feature extraction. The features are used only from the outputs of pool 1, pool 3, pool 4, and pool 5. The size of the search window is fixed to 1.8 times the target size. The regularization parameter of (2) is set to $\lambda = 10^{-4}$, the kernel width is taken as 0.1 for the generation of the Gaussian function labels, the learning rate η in (6) is set to 0.01, and T_0 is taken as 0.3. The value of γ is set as 1, 0.5, 0.25, and 0.15 for the conv5-4, conv4-4, conv3-4, conv1-4 layers respectively.

The performance of the proposed tracker has been evaluated based on two performance metrics, the Area-Under-the-Curve (AUC) and the Distance Precision (DP). For the validation of the proposed tracker's performance, a comparison has been carried out based on trackers presented in previous works: ASLA [37], CSK [38], DSST [30], MEEM [39], MUSTER [40], SAMF [41], SRDCF [27], Struck [42], siamfc3s [43], HCFTs [4], HDT [11], Staple [44], CNN-SVM [45], CF2 [3], LCT [46], KCF [32], TLD [47], KCF_GaussHog [32],KCF_LinearHog [32], BACF [48], DeepSRDCF [49], DRVT [50], MemDTC [51], MemTrack [51], SRDCFdecon [52]. The obtained results, corresponding to the two considered performance metrics, have been shown by two curves for One-Pass Evaluation (OPE), such as the DP rate vs. the location error threshold which measures the proportion of frames with distance between the tracking results and the ground truth less than a certain number of pixels, and the success rate vs. the overlap threshold which describes the percentage of successful frames as shown in Figures 4 and 5 for the OBT50 and OBT100 datasets respectively. The location error threshold variation is taken within the interval of [0, 50] and the overlap threshold variation within the interval of [0, 1]. The legend of the precision plots of OPE shows the ranking of the different trackers compared to the proposed tracker based on the DP score sat a threshold of 20 pixels, whereas the legend of success plots of OPE shows the ranking of the same trackers as in the previous Figure based on the AUC score. The obtained results for both datasets prove that the proposed tracker outperforms the other state-of-the-art trackers.

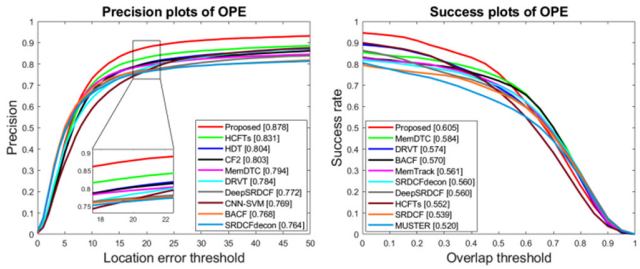

Fig. 4. OPE curves on the OTB50 dataset. Left: Overlap Precision (OP), right: Center Localization Error (CLE).

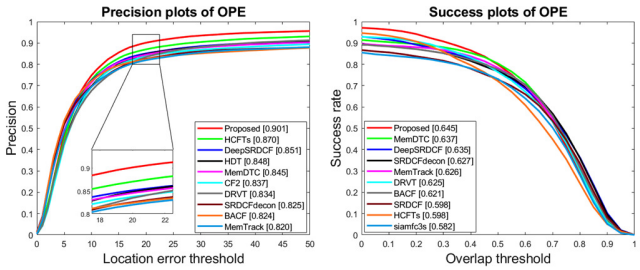

Fig. 5. OPE curves on the OTB100 dataset. Left: OP, right: CLE.

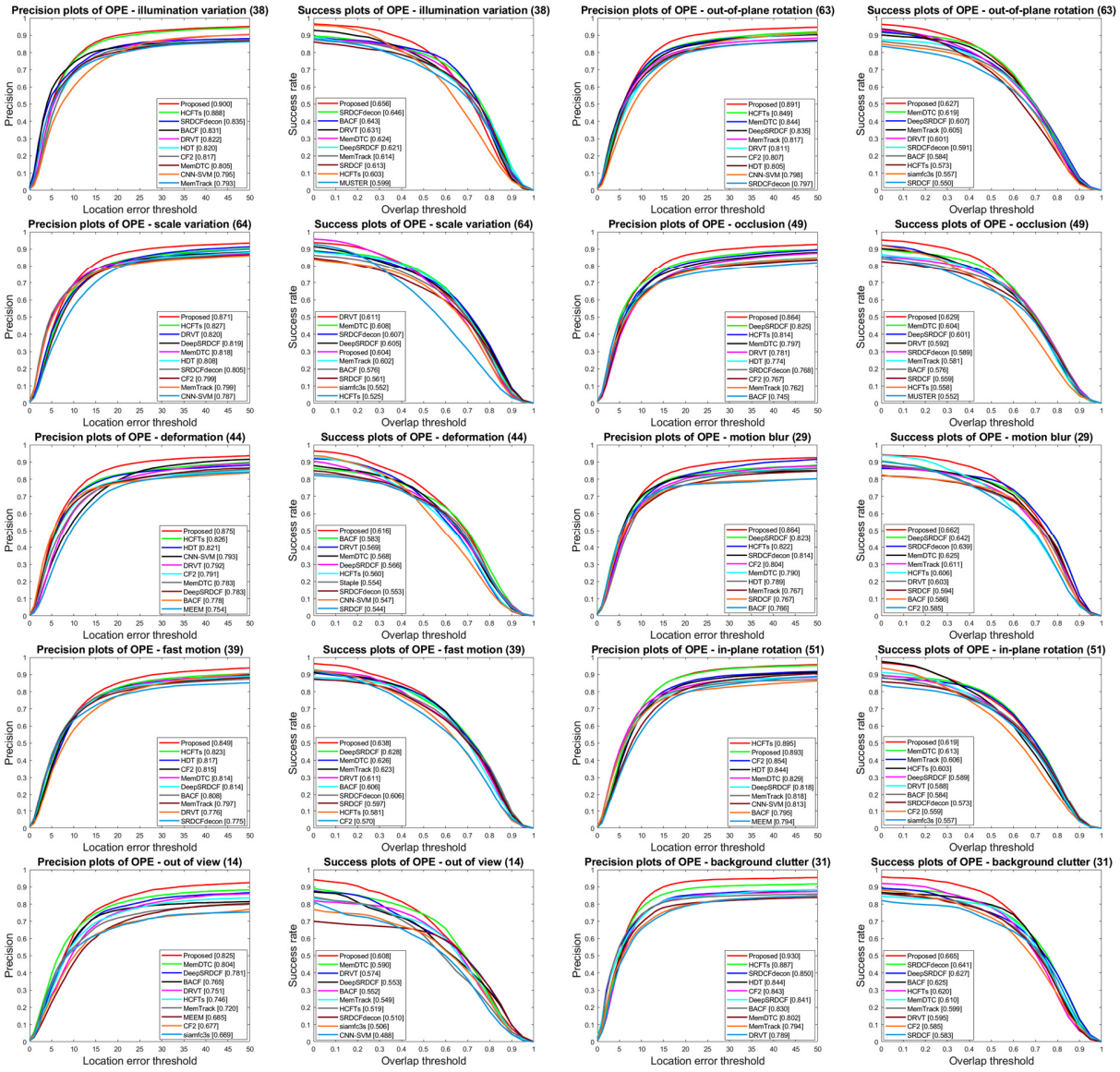

Fig. 6. Overlap success plots and distance precision plots in 10 tracking challenge situations.

For the clarity of result presentation, only the top 10 ranked trackers are taken into account as shown in Figures 5-6. It is clear that the proposed tracker performs favorably with an AUC of 64.5% and 60.5% and a DP of 90.1% and 87.8% on OTB-100 and OTB-50 respectively.

The obtained results prove clearly that the proposed tracker outperforms the second best tracker (HCFT), with a gain of 3.1% in the average DP, and with by 0.8% in the average overlap precision the second best tracker (MemDTC) in the case of OTB-100. For the OTB-50, the proposed tracker has an

improved gain of 4.7% and 2.1% for the average DP and average overlap precision compared to the second best tracker which is the same as the case of OTB-100. Figure 6 illustrates the overlap success rate plots and the DP plots obtained on OTB100 dataset for 10 challenging instances such as scale variation, fast motion, in-plane rotation, deformation, motion blur, occlusion, illumination variation, out-of-plane rotation, background clutter, and out-of-view. It can be clearly noted within all the sub-figures of Figure 6 that the proposed tracker outperforms the aforementioned state-of-the-art counterparts.

Fig. 7. Qualitative results of the proposed method, along with trackers from [3, 4, 32, 33, 45] challenge sequences.

Figure 7 shows the tracking performance of several representative trackers such as HCFTs [4], CNN-SVM [45], CF2 [3], DSST [30], KCF [32], and the proposed tracker on 6 challenging sequences. From top to bottom, the sequences are Human3, Girl2, Human5, Box, Car1, and Lemming respectively. Human3 has the challenge of scale variation occlusion, deformation, background clutters, and out-of-plane rotation. Girl2 contains scale variation, occlusion, deformation, out-of-plane rotation, and motion blur. Human5 includes scale variation, occlusion, and deformation. Car1 compromises illumination variation, motion blur, scale variation, fast motion and background clutters. Box holds illumination variation, motion blur, occlusion, out-of-plane rotation, in-plane rotation, background clutters, scale variation, and out-of-view. Lemming gathers illumination variation, fast motion, occlusion, scale variation, out-of-view, and out-of-plane rotation. It is obvious that the proposed tracker handles all these complicated scenarios better than the other trackers.

CNN-SVM with deep features performs well when scale variation, out-of-plane rotation, and occlusion are present (Human3, and Girl2), but it is less effective in handling drastic variations (Human5, Box, Lemming). CF, DSST and KCF are less effective in dealing with occlusion and deformation (Human3, Girl2, Human5, Box, and Lemming). HCFTs performs well in presence of scale variation, occlusion, motion blur and background clutter (Box, Lemming) but it is less effective in dealing with deformation (Human3,Girl2 and Human5). Overall, the proposed tracker operates adaptively and robustly when confronted with a variety of challenging factors. It is worth noting a minor drawback which has been faced during the application of the proposed tracker. The significant change in the aspect ratio in the case of Jump sequence has caused the missing of the target, which implies that a more robust design is required for scale variation, and aspect ratio adjustment strategy for the proposed tracker to overcome completely these kinds of deficiencies.

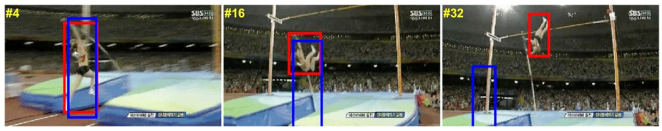

Fig. 8. The proposed tracker failed on the sequence Jump from OTB-100. The red and blue bounding boxes indicate the ground truth and the proposed tracker results respectively.

To check the effectiveness of the proposed tracker, it has been implemented based on two different proposed methods such as the HOG with wavelet (HOG-DWT) and HOG without wavelet, which have been combined with Hierarchical Convolutional Features (HCFs), with and without wavelets and it has been evaluated with the OTB-100. The obtained results of the tracker based on the two proposed methods, are shown in Figure 9, taking into account different combinations such as:

- Proposed: the tracker is based on HOG with DWT and HCFs with DWT.
- Proposed, No DWT in CNN: the tracker is based on HOG with DWT and HCFs without DWT.
- Proposed, No DWT in HOG: the tracker is based on HOG without DWT and HCFs with DWT.
- Proposed, No DWT in HOG and CNN: the tracker is based on HOG without DWT and HCFs without DWT.
- Proposed, No HOG: the tracker is based only on HCFs with DWT.
- Proposed, No HOG, No DWT in CNN: the tracker is based on HCFs without DWT.

It is clear that combining HCF, HOG, and wavelets ensured optimal results, although HOG affects the proposed method in object tracking under the occurrence of out-of-view, occlusion, out-of-plane rotation, motion blur, scale variation, deformation, and illumination variation. From the obtained results, it can be said that the exploitation of wavelets in HOG has improved the proposed tracker in handling occlusion, scale variation, and out-of-plane rotation. Furthermore, the use of wavelet in HCFs

improves the proposed tracker's behavior in handling only the illumination variation occurrence. Through the above analysis, it can be concluded that the combination of HCFs, HOG, and DWT can greatly improve the robustness and accuracy of the proposed tracker.

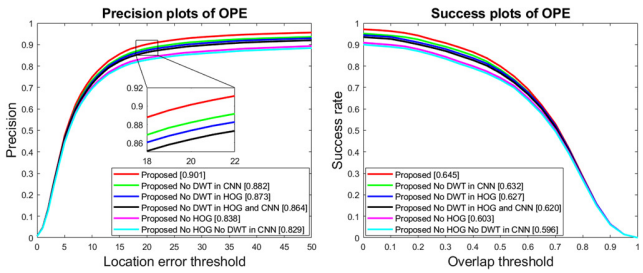

Fig. 9. Performance evaluation of the proposed tracker at each stage using the OTB-100 dataset.

In the present analysis, the use of different types of wavelets has been investigated on the tracker's performance. Figure 10 shows the calculated precision for a series of distance thresholds (percentage of frames where the distance to the ground truth is within the threshold) of Singer2 sequence. It can be clearly noticed that the wavelet type 'bior2.4' leads to obtaining the best result of 98.4%, which further justifies the validity of the proposed approach.

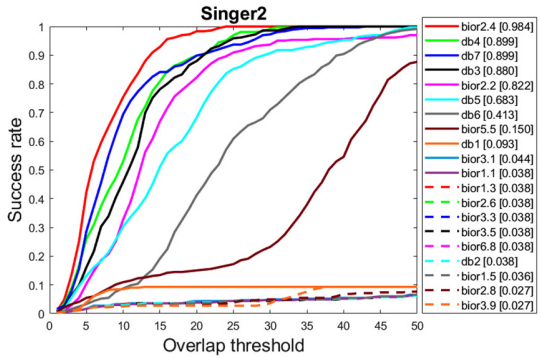

Fig. 10. Precision of sequence Singer2 with different types of wavelets.

VI. CONCLUSION

In this paper, an effective combination among CNN layer features, HOG features, and DWT image wavelet transforms, based on the exploitation of the hierarchical CNN features which have been trained on a large scale data base, has been proposed for the improvement of the visual object tracking algorithm. The output layers of the CNNs are used to preserve the semantics of the target objects, which are robust to significant appearance changes. The input layers of the CNNs are exploited for encoding more precise spatial details, which are useful for precise localization. Both features with the precise details are used at the same time for visual object tracking, while a linear correlation filter has been trained on each CNN layer for the deduction of the targeted location based on hierarchical correlation maps in a coarse-to-fine manner. In the same time, to enhance the accuracy of the proposed tracker and to overcome the problem of drifting encountered during the update process of the correlation filter, an approach for ensuring such process in real time along each step has been proposed. This approach is based on training of the correlation filter on HOG features in order to make it a tool to update the filters produced by CNN and HOG features. Furthermore, to improve the performance of the proposed tracker, the DWT has been utilized to achieve two main goals: the calculation of the HOG features instead of using RGB and the calculation of CNN features in the case of images with high saturation. The obtained results from the extensive carried out simulations, show that the proposed tracker outperforms the state of the art trackers. However, despite the proven effectiveness of the proposed tracker, there is a need to further improve its robustness in the future.

REFERENCES

[1] F. A. Dharejo *et al.*, "A deep hybrid neural network for single image dehazing via wavelet transform," *Optik*, vol. 231, Apr. 2021, Art. no. 166462, https://doi.org/10.1016/j.ijleo.2021.166462.

[2] M. Y. Abbass, K.-C. Kwon, N. Kim, S. A. Abdelwahab, F. E. A. El-Samie, and A. A. M. Khalaf, "Efficient object tracking using hierarchical convolutional features model and correlation filters," *The Visual Computer*, vol. 37, no. 4, pp. 831–842, Apr. 2021, https://doi.org/10.1007/s00371-020-01833-5.

[3] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical Convolutional Features for Visual Tracking," in *IEEE International Conference on Computer Vision*, Santiago, Chile, Dec. 2015, pp. 3074–3082, https://doi.org/10.1109/ICCV.2015.352.

[4] 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*, vol. 41, no. 11, pp. 2709–2723, Aug. 2019, https://doi.org/10.1109/TPAMI.2018.2865311.

[5] A. Zgaren, W. Bouachir, and R. Ksantini, "Coarse-to-Fine Object Tracking Using Deep Features and Correlation Filters," in *15th International Symposium on Visual Computing*, San Diego, CA, USA, Nov. 2020, pp. 517–529, https://doi.org/10.1007/978-3-030-64556-4_40.

[6] Y. Said, M. Barr, and H. E. Ahmed, "Design of a Face Recognition System based on Convolutional Neural Network (CNN)," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5608–5612, Jun. 2020, https://doi.org/10.48084/etasr.3490.

[7] P. Chakraborty and C. Tharini, "Pneumonia and Eye Disease Detection using Convolutional Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5769–5774, Jun. 2020, https://doi.org/10.48084/etasr.3503.

[8] S. Alqethami, B. Almtanni, W. Alzhrani, and M. Alghamdi, "Disease Detection in Apple Leaves Using Image Processing Techniques," *Engineering, Technology & Applied Science Research*, vol. 12, no. 2, pp. 8335–8341, Apr. 2022, https://doi.org/10.48084/etasr.4721.

[9] J. Zhang, J. Sun, J. Wang, and X.-G. Yue, "Visual object tracking based on residual network and cascaded correlation filters," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 8, pp. 8427–8440, Aug. 2021, https://doi.org/10.1007/s12652-020-02572-0.

[10] Y. Bai, T. Xu, B. Huang, and R. Yang, "Deep Deblurring Correlation Filter for Object Tracking," *IEEE Access*, vol. 8, pp. 68623–68637, 2020, https://doi.org/10.1109/ACCESS.2020.2986311.

[11] Y. Qi *et al.*, "Hedged Deep Tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Jun. 2016, pp. 4303–4311, https://doi.org/10.1109/CVPR.2016.466.

[12] C. Ma, Y. Xu, B. Ni, and X. Yang, "When Correlation Filters Meet Convolutional Neural Networks for Visual Tracking," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1454–1458, Jul. 2016, https://doi.org/10.1109/LSP.2016.2601691.

[13] D. E. Touil, N. Terki, and S. Medouakh, "Hierarchical convolutional features for visual tracking via two combined color spaces with SVM

classifier," *Signal, Image and Video Processing*, vol. 13, no. 2, pp. 359–368, Mar. 2019, https://doi.org/10.1007/s11760-018-1364-z.

[14] B. Latreche, S. Saadi, M. Kious, and A. Benziane, "A novel hybrid image fusion method based on integer lifting wavelet and discrete cosine transformer for visual sensor networks," *Multimedia Tools and Applications*, vol. 78, no. 8, pp. 10865–10887, Apr. 2019, https://doi.org/10.1007/s11042-018-6676-z.

[15] M. X. Bastidas Rodriguez *et al.*, "Deep Adaptive Wavelet Network," in *IEEE Winter Conference on Applications of Computer Vision*, Snowmass, CO, USA, Mar. 2020, pp. 3100–3108, https://doi.org/10.1109/WACV45572.2020.9093580.

[16] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet Convolutional Neural Networks," arXiv, arXiv:1805.08620, May 2018. https://doi.org/10.48550/arXiv.1805.08620.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243.

[18] H. Lu, H. Wang, Q. Zhang, D. Won, and S. W. Yoon, "A Dual-Tree Complex Wavelet Transform Based Convolutional Neural Network for Human Thyroid Medical Image Segmentation," in *IEEE International Conference on Healthcare Informatics*, New York, NY, USA, Jun. 2018, pp. 191–198, https://doi.org/10.1109/ICHI.2018.00029.

[19] F. Cotter and N. Kingsbury, "Deep Learning in the Wavelet Domain," arXiv, arXiv:1811.06115, Nov. 2018. https://doi.org/10.48550/arXiv.1811.06115.

[20] W. Yun, D. Kim, B. Song, and H. Yoon, "Block comparison based face identification using HOG feature," in *18th IEEE International Symposium on Robot and Human Interactive Communication*, Toyama, Japan, Oct. 2009, pp. 484–487, https://doi.org/10.1109/ROMAN.2009.5326203.

[21] W. Zhang, G. Zelinsky, and D. Samaras, "Real-time Accurate Object Detection using Multiple Resolutions," in *IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8, https://doi.org/10.1109/ICCV.2007.4409057.

[22] M. Villamizar, F. Moreno-Noguer, J. Andrade-Cetto, and A. Sanfeliu, "Efficient rotation invariant object detection using boosted Random Ferns," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun. 2010, pp. 1038–1045, https://doi.org/10.1109/CVPR.2010.5540104.

[23] Y. Wei, Q. Tian, and T. Guo, "An Improved Pedestrian Detection Algorithm Integrating Haar-Like Features and HOG Descriptors," *Advances in Mechanical Engineering*, vol. 5, Jan. 2013, Art. no. 546206, https://doi.org/10.1155/2013/546206.

[24] D. E. Touil, N. Terki, and S. Medouakh, "Learning spatially correlation filters based on convolutional features via PSO algorithm and two combined color spaces for visual tracking," *Applied Intelligence*, vol. 48, no. 9, pp. 2837–2846, Sep. 2018, https://doi.org/10.1007/s10489-017-1120-z.

[25] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, arXiv:1409.1556, Apr. 2015. https://doi.org/10.48550/arXiv.1409.1556.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, Jun. 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[27] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning Spatially Regularized Correlation Filters for Visual Tracking," in *IEEE International Conference on Computer Vision*, Santiago, Chile, Dec. 2015, pp. 4310–4318, https://doi.org/10.1109/ICCV.2015.490.

[28] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun. 2010, pp. 2544–2550, https://doi.org/10.1109/CVPR.2010.5539960.

[29] K. Zhang, L. Zhang, and M.-H. Yang, "Real-Time Compressive Tracking," in *12th European Conference on Computer Vision*, Florence, Italy, Oct. 2012, pp. 864–877, https://doi.org/10.1007/978-3-642-33712-3_62.

[30] M. Danelljan, G. Hager, F. Khan, and M. Felsberg, "Accurate Scale Estimation for Robust Visual Tracking," in *British Machine Vision Conference*, Nottingham, UK, Sep. 2014, https://doi.org/10.5244/C.28.65.

[31] H. K. Galoogahi, T. Sim, and S. Lucey, "Multi-channel Correlation Filters," in *IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, Dec. 2013, pp. 3072–3079, https://doi.org/10.1109/ICCV.2013.381.

[32] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, Mar. 2015, https://doi.org/10.1109/TPAMI.2014.2345390.

[33] V. N. Boddeti, T. Kanade, and B. V. K. V. Kumar, "Correlation Filters for Object Alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, Jun. 2013, pp. 2291–2298, https://doi.org/10.1109/CVPR.2013.297.

[34] F. A. Dharejo *et al.*, "A deep hybrid neural network for single image dehazing via wavelet transform," *Optik*, vol. 231, Apr. 2021, Art. no. 166462, https://doi.org/10.1016/j.ijleo.2021.166462.

[35] Y. Wu, J. Lim, and M.-H. Yang, "Object Tracking Benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015, https://doi.org/10.1109/TPAMI.2014.2388226.

[36] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional Neural Networks for MATLAB," in *23rd ACM international conference on Multimedia*, Brisbane, Australia, Oct. 2015, pp. 689–692, https://doi.org/10.1145/2733373.2807412.

[37] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, Jun. 2012, pp. 1822–1829, https://doi.org/10.1109/CVPR.2012.6247880.

[38] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the Circulant Structure of Tracking-by-Detection with Kernels," in *12th European Conference on Computer Vision*, Florence, Italy, Oct. 2012, pp. 702–715, https://doi.org/10.1007/978-3-642-33765-9_50.

[39] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust Tracking via Multiple Experts Using Entropy Minimization," in *13th European Conference*, Zurich, Switzerland, Sep. 2014, pp. 188–203, https://doi.org/10.1007/978-3-319-10599-4_13.

[40] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "MUlti-Store Tracker (MUSTer): A cognitive psychology inspired approach to object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, Jun. 2015, pp. 749–758, https://doi.org/10.1109/CVPR.2015.7298675.

[41] Y. Li and J. Zhu, "A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration," in *Computer Vision - ECCV 2014 Workshops*, Zurich, Switzerland, Sep. 2014, pp. 254–265, https://doi.org/10.1007/978-3-319-16181-5_18.

[42] S. Hare *et al.*, "Struck: Structured Output Tracking with Kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2096–2109, Jul. 2016, https://doi.org/10.1109/TPAMI.2015.2509974.

[43] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Computer Vision – ECCV 2016 Workshops*, Amsterdam, Netherlands, Oct. 2016, pp. 850–865, https://doi.org/10.1007/978-3-319-48881-3_56.

[44] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary Learners for Real-Time Tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Jun. 2016, pp. 1401–1409, https://doi.org/10.1109/CVPR.2016.156.

[45] S. Hong, T. You, S. Kwak, and B. Han, "Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network," in *32nd International Conference on Machine Learning*, Lille, France, Jul. 2015, pp. 597–606, https://doi.org/10.48550/arXiv.1502.06796.

[46] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, Jun. 2015, pp. 5388–5396, https://doi.org/10.1109/CVPR.2015.7299177.

[47] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012, https://doi.org/10.1109/TPAMI.2011.239.

[48] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning Background-Aware Correlation Filters for Visual Tracking," in *IEEE International Conference on Computer Vision*, Venice, Italy, Oct. 2017, pp. 1144–1152, https://doi.org/10.1109/ICCV.2017.129.

[49] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional Features for Correlation Filter Based Visual Tracking," in *IEEE International Conference on Computer Vision Workshop*, Santiago, Chile, Dec. 2015, pp. 621–629, https://doi.org/10.1109/ICCVW.2015.84.

[50] X. Li, Q. Liu, N. Fan, Z. Zhou, Z. He, and X. Jing, "Dual-regression model for visual tracking," *Neural Networks*, vol. 132, pp. 364–374, Dec. 2020, https://doi.org/10.1016/j.neunet.2020.09.01.

[51] T. Yang and A. B. Chan, "Visual Tracking via Dynamic Memory Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 360–374, Jan. 2021, https://doi.org/10.1109/TPAMI.2019.2929034.

[52] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Jun. 2016, pp. 1430–1438, https://doi.org/10.1109/CVPR.2016.159.