International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

Feature Fusion Strategy Based on Concatenation of Histogram of Oriented Gradients and Pretrained CNN Features for Visual Object Tracking

[1]Villari Sreenatha Sarma, [2]P.M. Ashok Kumar,
[3]Vadamala Purandhar Reddy

[1,2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.
[1,3]Audisankara College of Engineering & Technology, Gudur, Andhra Pradesh, India.
[1]villariss@gmail.com, ORCID ID: 0000-0002-7754-2778,
[2]pmashokk@kluniversity.in, ORCID ID: 0000-0002-4134-4163,
[3]vpurandharreddy@gmail.com, ORCID ID: 0000-0003-1755-3317

**Corresponding Author: Villari Sreenatha Sarma*

Abstract: This paper presents a hybrid feature fusion approach for visual object tracking that combines hand-crafted gradient-based descriptors with semantic representations extracted from a pretrained convolutional neural network. The proposed tracker integrates Histogram of Oriented Gradients (HOG) features with ResNet-50-based deep features within a correlation filter framework to improve robustness against appearance variations, occlusion, and illumination changes. A concatenation-based fusion strategy with adaptive confidence-driven weighting is incorporated to dynamically balance the contribution of handcrafted and deep features during tracking. The architecture employs parallel feature extraction branches and multi-scale feature integration to enhance localization performance while maintaining computational efficiency. Experimental evaluation on standard benchmark datasets, including OTB-2015 and related tracking sequences, demonstrates that the proposed fusion strategy provides improved performance compared with individual feature-based tracking approaches and achieves competitive results relative to baseline correlation-filter trackers under challenging conditions. The study also outlines potential directions for further enhancement through online adaptation and attention-based feature fusion strategies.

Keywords: Visual Object Tracking, Feature Fusion, Histogram of Oriented Gradients, Correlation Filter Tracking, Adaptive Feature Weighting.

1 INTRODUCTION

The continuous localization of an object across successive video frames, starting from an initial position, is a fundamental task in computer vision known as visual object tracking. Accurate tracking requires maintaining the temporal consistency of the target's position despite challenges such as occlusion, illumination variation, background clutter, deformation, and rapid motion. These challenges arise frequently in real-world applications, including surveillance systems, autonomous driving, unmanned aerial vehicle (UAV) navigation, and human–computer interaction (HCI) environments. Robust visual tracking therefore plays an important role in enabling reliable scene understanding and supporting intelligent decision-making in dynamic environments [1,2].

Correlation filter (CF)-based trackers are widely adopted because of their computational efficiency and strong performance in real-time tracking scenarios [3]. These methods exploit the convolution theorem to transform spatial-domain convolution into element-wise multiplication in the frequency domain, enabling fast target localization. However, early CF-based trackers suffered from boundary effects and limited robustness to target appearance changes [4]. Subsequent developments introduced spatial regularization and multi-resolution feature representations to improve tracking accuracy and stability under challenging conditions [5,6].

Tracking performance is strongly influenced by the quality of feature representation used to describe the target. Handcrafted descriptors such as Histogram of Oriented Gradients capture local structural information effectively, while deep convolutional neural network (CNN) features extracted from architectures such as VGG or ResNet provide high-level semantic representations [7]. Integrating handcrafted and deep features through feature fusion strategies can improve robustness against occlusion, deformation, and illumination variation by combining complementary spatial and semantic information [8].

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

Correlation filters are widely used in visual object tracking because of their computational efficiency and strong localization capability. These filters reformulate the tracking task as a classification problem that distinguishes the target from the surrounding background in an efficient manner. However, their performance may degrade when the target undergoes occlusion or significant appearance variation. Multi-channel feature representations and adaptive learning strategies have therefore been introduced to improve tracking robustness. For example, incorporation of HOG, CN and saliency features has been shown to enhance tracking against appearance variation [9,10]. Scale-adaptive strategies based on image pyramids and occlusion-aware mechanisms further improve tracking accuracy under challenging conditions [11].

Correlation filters generate response maps whose peak values indicate the estimated target location. Correlation-filter-based tracking methods have demonstrated strong performance across benchmark datasets such as OTB2013, OTB2015, and UAV123. Adaptive learning strategies, for example, have achieved high Distance Precision (DP) scores, a measure of tracking performance [10]. Though correlation filter-based trackers have several advantages, they are challenged by changes in appearance and require frequent updates for optimal performance. Aberrance suppression and feature fusion have been employed to address these problems [12]. Improving robustness against appearance variation remains an important research direction in correlation-filter-based tracking. Robust and adaptable object tracking systems therefore require advanced learning techniques for effective feature extraction. A continuing challenge in correlation-filter-based trackers is updating the target model during online tracking. Simple update strategies that employ a constant learning rate can lead to model contamination and subsequent drift due to occlusions or sudden changes in appearance [13]. Adaptive strategies retain and selectively update representative sample subsets based on tracking confidence metrics such as Average Peak-to-Correlation Energy (APCE) [14].

This work proposes a hybrid visual object tracking framework that integrates handcrafted gradient-based descriptors with pretrained deep semantic features to improve tracking robustness and accuracy. A concatenation-based feature fusion strategy with adaptive weighting is integrated within a correlation-filter tracking framework. The architecture employs parallel feature extraction branches for HOG and CNN features and incorporates multi-scale feature integration with confidence-based adaptive weighting. Experimental evaluation on benchmark tracking datasets demonstrates improved performance compared with individual feature-based tracking approaches while maintaining computational efficiency. Unlike earlier hybrid feature fusion trackers that apply fixed-weight concatenation or static feature integration, the proposed framework introduces confidence-driven adaptive fusion within a correlation-filter tracking architecture combined with multi-scale feature estimation for improved robustness under appearance variation.

The key contributions in the work are as follows

1. Hybrid Semantic-Geometric Framework Integrated high-level ResNet-50 CNN features with low-level HOG descriptors to simultaneously capture semantic context and spatial gradients.
2. Adaptive confidence-based fusion mechanism that dynamically adjusts feature importance based on tracking confidence.
3. Parallel multi-scale architecture that integrates multi-scale features while maintaining computational efficiency.
4. Experimental evaluation on representative OTB-2015 tracking sequences demonstrates measurable performance improvement over individual feature-based tracking approaches using precision and robustness metrics.

2 Literature Survey

The Minimum Output Sum of Squared Error (MOSSE) tracker introduced one of the earliest correlation-filter-based tracking frameworks [3]. The Kernelized Correlation Filter (KCF) tracker further improved performance by incorporating multi-channel features such as Histogram of Oriented Gradients descriptors [11]. The Discriminative Scale Space Tracker (DSST) addressed the issue of scale variation by the use of separate scale filters over multi-scale image pyramids [5]. The Spatially Regularized DCF (SRDCF) reduced boundary effects through spatial regularization of filter coefficients, while Background-Aware Correlation Filters (BACF) improved discrimination performance by effectively utilizing real negative samples [12,13]. The incorporation of deep features into CF trackers led to the development of C-COT and ECO [5] which use deep multi-resolution CNN features.

Deep correlation-filter trackers generally achieve improved accuracy but often introduce increased computational complexity compared with lightweight correlation-filter approaches [15]. Model contamination during target occlusion and sudden appearance changes can occur when simple moving-average update strategies are used [9]. More robust adaptive update strategies improve tracking stability by increasing sample diversity and adjusting filter updates based on confidence scores [15]. Long-term tracking approaches such as MUSTer combine short-term correlation-filter tracking with long-term keypoint-based verification mechanisms [11]. Parallel Tracking and Verifying (PTAV) frameworks enable recovery from tracking failure through real-time verification mechanisms. Also, particle filter based redetection modules assist in the recovery of lost targets [16]. Recent research has focused on effective update mechanisms, attention-based feature fusion strategies, and multi-modal tracking using RGB-D or thermal imagery to improve robustness in cluttered environments [7].

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

With respect to the design, real-time tracking, in particular for UAV tracking, remains a priority. Benchmark datasets such as OTB-2015 [2], Temple Colour 128 [17], UAV123 [18], and UAV20L are widely used for evaluating visual tracking performance. Fusion of handcrafted and deep features remains an effective strategy for balancing tracking precision and computational efficiency. Basic concatenation and weighted summation techniques have proven successful [6]. More sophisticated attention mechanisms dynamically adjust feature importance based on scene characteristics in real time [7]. Colour features such as Colour Names [5] and RGB histograms [8] improve robustness under illumination variation and complex background conditions. Dimensionality reduction using self-adaptive PCA-based fusion while preserving discriminative power has also been investigated in prior work [19].

3 METHODOLOGY

The Histogram of Oriented Gradients technique captures local shape information by performing gradient computation, orientation binning, and block normalization. The resulting descriptors are partially invariant to illumination variation due to gradient normalization. ResNet-50, a pretrained deep convolutional neural network, extracts high-level semantic features from the final convolutional layer, yielding 2048-dimensional feature vectors that encode rich contextual information [20].

3.1. HOG Feature Extraction

For each pixel (x, y) in the image patch $I(x, y)$, gradients in the x and y directions are computed as:

$$G_x(x, y) = I(x + 1, y) - I(x - 1, y) \tag{1}$$

$$G_y(x, y) = I(x, y + 1) - I(x, y - 1) \tag{2}$$

The gradient magnitude $M(x, y)$ and orientation $\theta(x, y)$ are computed as:

$$M(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{3}$$

$$\theta(x, y) = \arctan\frac{G_y(x, y)}{G_x(x, y)} \tag{4}$$

Orientations are quantized into bins (usually 9 bins over 0°–180°). Each pixel votes into orientation bins weighted by its magnitude. Cells are grouped into blocks, and the feature vector B is normalized using the L_2 norm to improve robustness against illumination variation:

$$B_{norm} = \frac{B}{\sqrt{\|B\|_2^2 + \epsilon^2}} \tag{5}$$

where ϵ is a small constant and B denotes the unnormalized HOG block feature vector.

3.2. ResNet-50 Feature Extraction

Let x be the input image patch. The ResNet-50 feature extraction function is represented as:

$$F_{ResNet}(x) = \phi(x; \theta) \tag{6}$$

where ϕ denotes the mapping of input through the ResNet network with pretrained weights θ. The output feature vector extracted from the final convolutional layer has dimension 2048.

3.3. Feature Fusion by Concatenation

The HOG feature vector $F_{HOG} \in \mathbb{R}^{d_{HOG}}$ and ResNet features $F_{ResNet} \in \mathbb{R}^{d_{ResNet}}$ are fused by concatenation:

$$F_{fused} = [F_{HOG}; F_{ResNet}] \in \mathbb{R}^{d_{HOG}+d_{ResNet}} \tag{7}$$

This fused representation preserves complementary information from both feature types. The fused features are integrated into a correlation filter tracking framework for robust target localisation [21].

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

3.4. Adaptive Fusion Module

The adaptive weighting module dynamically adjusts the contribution of handcrafted and deep features during tracking. In visual tracking, HOG features effectively capture structural boundary information, whereas ResNet features provide higher-level semantic representations of the target. However, their usefulness fluctuates; for example, if the lighting changes, HOG might fail, and if the object undergoes a significant deformation, the high-level ResNet features might become more reliable.

3.4.1. The Adaptive Gating Mechanism

The adaptive fusion module employs a sigmoid-based gating function to estimate feature reliability weights. Instead of using a fixed ratio (e.g., 50% HOG, 50% CNN), the system learns a dynamic weight w based on the current appearance of the target.

$$\omega = \sigma(w_h f_{HOG} + w_r f_{ResNet} + b) \tag{8}$$

f_{HOG} and f_{ResNet} are the raw feature vectors extracted from the current image patch. W_h and W_r are the Learned weight matrices (linear layers) that project the high-dimensional features into a common latent space to evaluate their reliability, 'b' denotes a bias term that enables flexible adjustment of the fusion weighting during tracking. $\sigma(\cdot)$ denotes the sigmoid activation function that constrains the output weighting factor to the interval [0,1]. The gating mechanism is given below.

Step A: Dimensionality Alignment

HOG and ResNet features have different scales and dimensions (ResNet is 2048-D, while HOG depends on cell/block size). The module first passes both through a small bottleneck layer (fully connected) to ensure they are comparable before the addition operation in the sigmoid function.

Step B: Weight Generation

The module estimates feature reliability using the projected feature representations. The learned projection parameters evaluate feature consistency to determine their relative importance during fusion. For example, degraded gradient structure may reduce the weighting assigned to HOG features. Conversely, reliable semantic representations increase the weighting assigned to CNN-based features.

Step C: Channel-Wise Feature Scaling

Once the scalar w is generated, it is applied to the features. There are two common ways to apply this:

1. Convex Combination: $f_{fused} = \omega \cdot f_{HOG} + (1 - \omega) \,.\, f_{ResNet}$ (9)

2. Concatenation Scaling: $f_{fused} = [\omega \cdot f_{HOG,} + (1 - \omega) \,.\, f_{ResNet}]$ (10)

This mechanism ensures that the feature representation with higher estimated reliability contributes more strongly to the final fused descriptor.

3.4.2. Impact on Correlation Filter Tracking

The weighted fused feature representation is transformed into the Fourier domain for correlation filter computation. The importance of adaptive weighting becomes clear in the Response Map (R) calculation:

$$R = F^{-1}\left(\sum_{d=1}^{D} \hat{H}^d \odot (W \odot \hat{z}^d)\right) \tag{11}$$

where $\hat{H}^d$ denotes the Fourier transform of the fused feature channel and W represents the adaptive fusion weighting coefficient. Without adaptive weighting, degraded feature quality may produce noisy response maps with multiple peaks, potentially leading to tracking drift. With adaptive weighting, the tracker reduces the contribution of unreliable features and improves response-map sharpness for more accurate target localization.

3.5. The Translation Filter $(\hat{H}_{trans})$

This is a 2D filter responsible for locating the target's center (x, y). The learning objective is to produce a high response at the target center and near-zero responses in surrounding background regions. The translation filter for each channel d is computed as:

$$\hat{H}^d = \frac{\hat{y} \odot \hat{\boldsymbol{f}}_{\boldsymbol{d}}}{\sum_{k=1}^{D}(\hat{\boldsymbol{f}}_{\boldsymbol{k}} \odot \hat{\boldsymbol{f}}_{\boldsymbol{k}}^{*}) + \lambda} \tag{12}$$

- $\hat{y}$: The Fourier transform of the Gaussian label.
- $\hat{f}_d$: Fourier transform of the fused HOG–ResNet feature channel.
- λ: regularization parameter used to prevent overfitting and numerical instability.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

3.6. The scale filter (H_{scale})

A one-dimensional scale filter H_{scale} is learned by minimizing the squared error between the correlation response and a desired Gaussian scale label y:

$$\min_{H_{Scale}} \sum_{d=1}^{D} \|H^d_{\text{scale}} * F^d - y\|^2 + \lambda \sum_{d=1}^{D} \| H^d_{\text{scale}}\|^2 \quad (13)$$

- D: Total number of fused feature channels (HOG + ResNet).
- y: A 1D Gaussian function with its peak at the current scale (center of the K samples).
- λ: regularization parameter

Just like the translation filter, we solve this in the Fourier domain for efficiency. Using the Sherman-Morrison formula, the solution for each channel d is:

$$\hat{H}^d_{scale} = \frac{\hat{y} \cdot \hat{F}^{\,d,*}}{\sum_{k=1}^{D} \hat{F}^k \cdot \hat{F}^{\,k,*} + \lambda} \quad (14)$$

- $\hat{F}^{d,*}$: The complex conjugate of the Fourier-transformed feature.
- $\odot$: Element-wise multiplication.

3.7. Object Tracking

The proposed object tracking framework, shown in Fig. 1, outlines a hybrid tracking pipeline that integrates handcrafted features with deep convolutional features to localize objects in video sequences. The process initiates by extracting a search patch from the current frame, which is then processed through a dual feature extraction stage. By utilizing HOG for structural representation and ResNet-50 for semantic feature extraction, the system captures a robust representation of the target. These distinct feature sets are integrated in the Adaptive Fusion block, which dynamically weights the inputs to ensure the tracker remains resilient against challenges like lighting changes or complex backgrounds.

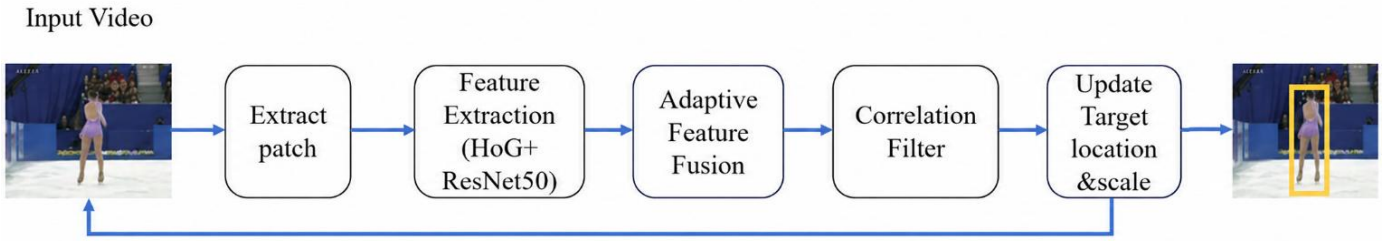

Fig.1. HOG and Pretrained CNN Feature Fusion Architecture for Visual Object Tracking

Following feature fusion, the fused representation is processed using a correlation filter that generates a response map whose maximum value indicates the estimated target location. Once the location and scale are updated, the system outputs the final tracked object coordinates. Crucially, a feedback loop returns this updated spatial information to the input stage, allowing the system to center its search patch for the subsequent frame. This recursive update mechanism maintains temporal consistency across frames and improves tracking stability as the object appearance evolves over time. The detailed tracking procedure is summarized in Algorithm 1.

3.8. Tracking Parameter Settings

The tracker uses a search window size equal to 2.5× the target bounding-box size. The translation filter learning rate was set to 0.02, and the scale filter learning rate was set to 0.025. The regularization parameter λ was fixed at 10^{-4}. A scale pyramid with 33 levels and scale step factor 1.02 was used. ResNet-50 features were extracted from the final convolutional activation layer prior to global pooling. The adaptive fusion module used a 128-dimensional bottleneck projection layer for feature alignment.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

Algorithm 1: HOG-CNN Fusion Tracker with Scale Estimation

Inputs : Frame I_t, previous state $(x, y, S)_{t-1}$
Learned filters $\widehat{H}_{trans}$and $\widehat{H}_{scale}$

Phase-1: **Translation Estimation (2D Search)**

1. **Extract search patch:** crop a patch from frame $I_t$centered at position $(x, y)_{t-1}$with window size proportional to scale S_{t-1}.
2. **Feature Extraction & Adaptive Fusion:**
   - Extract HOG features $f_{HOG}$and ResNet-50 features f_{ResNet}.
   - Compute dynamic weights: $\omega = \sigma(w_h f_{HOG} + w_r f_{ResNet} + b)$.
   - Form the fused feature representation: $f_{fused} = [\omega f_{HOG}, (1-\omega) f_{ResNet}]$.
3. **Correlation response:** compute the response map R_{trans}in the Fourier domain:

$$R_{trans} = F^{-1}\left(\frac{\sum_{d=1}^{D} \hat{A}_{trans}^{d} \odot \hat{z}_{trans}^{d}}{\widehat{\boldsymbol{B}}_{trans} + \lambda}\right) \quad (15)$$

where $\hat{A}_{trans}$ and $\hat{B}_{trans}$ denote numerator and denominator terms of the translation filter in the Fourier domain.

4. **Target localization:** Update the position to the peak response location:

$$(\hat{x}, \hat{y})_t = \arg\max(R_{trans}) \quad (16)$$

Phase-2: **Scale Estimation (1D Search)**

5. **Construct Scale Pyramid:** Extract K patches centered at $(\hat{x}, \hat{y})_t$ with varying sizes:

$$size_k = a^k \cdot S_t - 1, k\epsilon\left\{\left[\frac{1-k}{2}\right], \cdots \left[\frac{k-1}{2}\right]\right\} \quad (17)$$

where a is the scale step, typically 1.02.)

6. **Feature Normalization:** Resize all K patches to a fixed reference size and extract fused features.
7. **1D Correlation:** Form a scale feature descriptor by concatenating the features of each patch level and correlate with the 1D scale filter:

$$\hat{s}_t = s_{t-1}.\arg\max\left(F^{-1}\left(\widehat{H}_{scale} \odot \hat{z}_{scale}\right)\right) \quad (18)$$

4 Results and Discussion

The proposed tracking strategy is evaluated using the OTB dataset along with additional tracking sequences from TrackerNet and GOT-10K datasets. The first dataset, the Object Tracking Benchmark (OTB), is an integral part of the visual tracking domain and contains 100 video sequences for the purpose of evaluating different tracking methodologies. The dataset is widely used as a benchmarking platform for comparing the performance of different tracking techniques. OTB-2015 comprises 100 video sequences, which were carefully assembled to cover a wide range of tracking difficulties, such as occlusion and changes in scale and illumination [11]. The benchmark is designed with a clear set of definitions for evaluation to enable researchers to measure the impact of the tracking methodologies in a rigorous manner.

Although OTB-2015 remains a foundational benchmark dataset, newer datasets such as DTTO and NT-VOT provide additional evaluation scenarios for extended tracking analysis. In this study, a representative subset of twenty-five sequences from the OTB-2015 dataset was evaluated to analyse tracker performance across major tracking challenges. The selected sequences include representative tracking challenges such as occlusion, scale variation, illumination variation, fast motion, and in-plane rotation. The proposed tracking framework was implemented using MATLAB 2024. The experiments were performed on a desktop system with an Intel Xeon 3.10 GHz CPU, 8 GB RAM, and an NVIDIA RTX 6000 GPU. Evaluation followed the standard OTB-2015 protocol proposed by Wu et al. [2]. The OTB-2015 evaluation framework includes One-Pass Evaluation (OPE), Temporal Robustness Evaluation (TRE), and Spatial Robustness Evaluation (SRE) metrics for analysing tracker reliability [2]. Wu et al. added that the sequences in the OTB dataset have 11 attributes, which pose a challenge for trackers. Tracker performance is analysed based on its ability to handle these eleven challenging attributes. Comparative evaluation results across the selected twenty-five sequences are presented against several baseline tracking methods. Fig. 2 illustrates the OTB dataset.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

Fig. 2. OTB Dataset

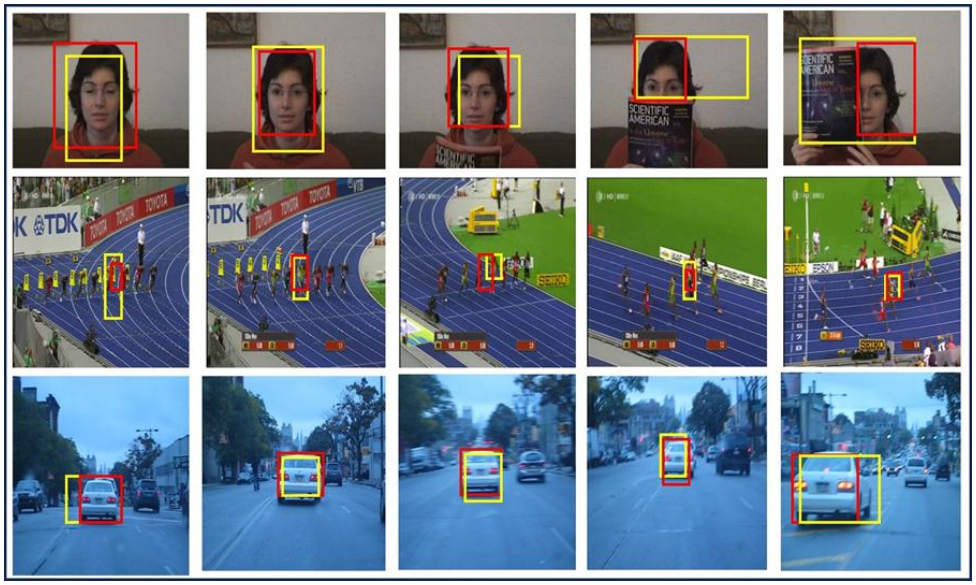

Fig. 3. Quantitative Results of the Proposed Tracker.

The proposed tracker's quantitative results are shown in Fig. 3. Ground truth is shown in a red colour bounding box, whereas the tracked result is shown in a yellow bounding box. In the first human sequence, the proposed technique effectively tracks the object under challenges of occlusion and scaling. In the second bolt sequence, the target is successfully tracked under illumination variation, in-plane rotation, and fast-motion conditions. Similar robustness is observed in the car sequence under multiple tracking challenges. The HOG-CNN fusion tracking approach ensures computational efficiency by leveraging simple yet complementary feature extraction methods. HOG extraction operates linearly with image size, while CNN feature extraction requires a fixed forward pass. Feature fusion through concatenation introduces only linear computational overhead with respect to feature dimensionality, and tracking is performed using correlation filters implemented via FFT operations.

HOG descriptors usually utilize 3,780 dimensions, CNN features vary based on the architecture, like VGG or ResNet, between 512 and 2,048 dimensions, and the fused representation gets the combined dimensionality ranging from 4,292 to 5,828. The tracker adopts an adaptive template update mechanism in which tracking confidence controls the learning rate to improve robustness against abrupt appearance changes. Multi-scale tracking involves response evaluation across a range of scale levels, while the response map informs the system when tracking has failed. The experimental results demonstrate improved adaptability to appearance variation and enhanced robustness under challenging tracking conditions.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

In the OTB-2015 evaluation subset, the fused representation achieved a precision score of 88.2% on the selected evaluation subset of OTB-2015 sequences, compared with 82.1% for CNN-only features and 71.3% for HOG-only features. The method attained an accuracy of 0.56, robustness of 0.89, and EAO of 0.31, indicating stable tracking behaviour across multiple appearance-change scenarios. Qualitative results of the OPE on the precision plot can be found in Figs. 4-9. In the case of the fast motion challenge, the proposed tracker achieves a precision of 67.7% compared with 76% achieved by the MUSTer tracker, as depicted in Fig. 4. The proposed tracker performs well in illumination with a precision of 76.9% when compared with the GradNet of 84.3%. Under occlusion conditions, the proposed tracker achieved a precision score of 74.2% compared with 80% for the MUSTer tracker. Under scale variation conditions, the proposed tracker achieved a precision score of 75.2% compared with 81.2% for the MUSTer tracker.

Under in-plane rotation conditions, the proposed tracker achieved a precision score of 69.9% compared with 77.7% for the MUSTer tracker. The overall OPE precision score of the proposed tracker is 76.1%, compared with 85.1% achieved by the top-ranked tracker. The proposed tracking framework provides a flexible basis for further improvement through the integration of more discriminative feature representations. The success rate evaluated using the area-under-curve (AUC) metric further confirmed consistent tracking performance across the evaluated sequences. Table 1 summarises the comparative precision performance of the proposed tracker against representative handcrafted, deep-feature, and hybrid tracking baselines including HOG-based tracking [11], CNN-based tracking [15], MUSTer [11], and GradNet [7].

Table 1. Precision Comparison with Baseline Trackers on Selected OTB-2015 Sequences

Tracker	Precision (%)
HOG only [11]	71.3
CNN only [15]	82.1
MUSTer [11]	76.0
GradNet [7]	84.3
Proposed	88.2

The proposed tracker achieves competitive precision performance among representative handcrafted, deep-feature, and hybrid baseline trackers. The precision plots in Figs. 4–9 compare the proposed tracker with representative baseline tracking methods including KCF [4], SRDCF [5], MUSTer [15], BACF [17], SAMF [24], CT [25], GradNet [26], and the Variance Ratio Feature Shift (VR-V) tracker [27]. In the precision plots shown in Figs. 4–9, the numerical values reported in parentheses in the legend indicate the precision score computed at a location error threshold of 20 pixels using the One-Pass Evaluation (OPE) protocol.

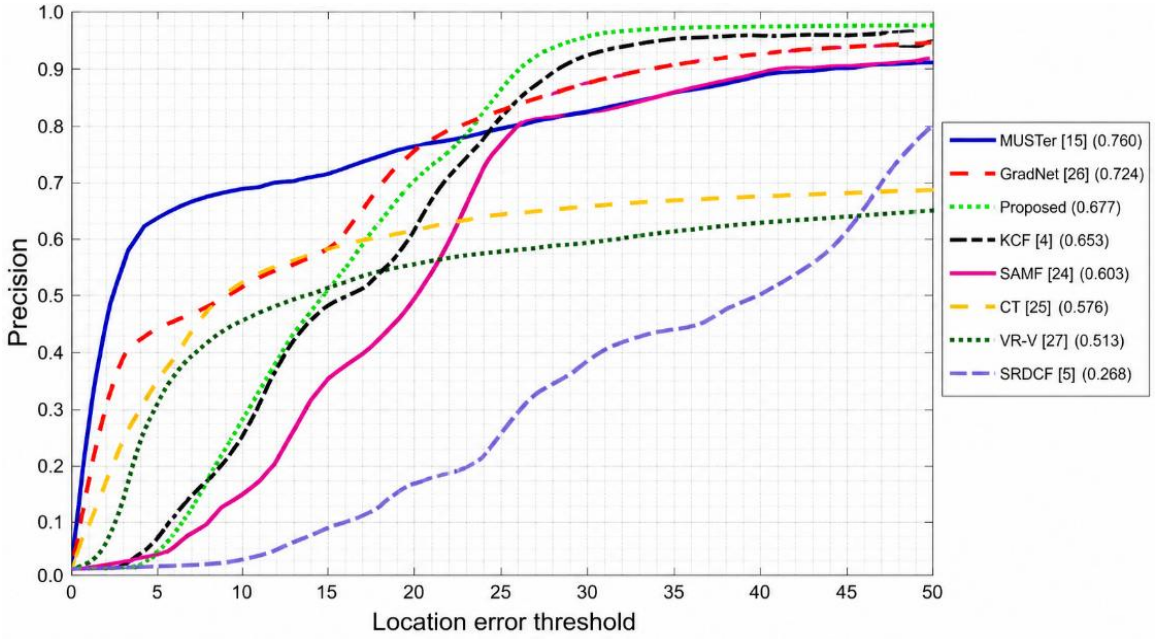

Fig. 4. OPE Fast Motion Precision Results

The proposed HOG–CNN fusion framework provides a flexible basis for further improvements through attention-based feature weighting, online model adaptation, and multi-modal feature integration using RGB-D or thermal imagery. Future work may also investigate transformer-based feature representations and lightweight backbone architectures to improve tracking robustness and computational efficiency under real-world deployment constraints.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

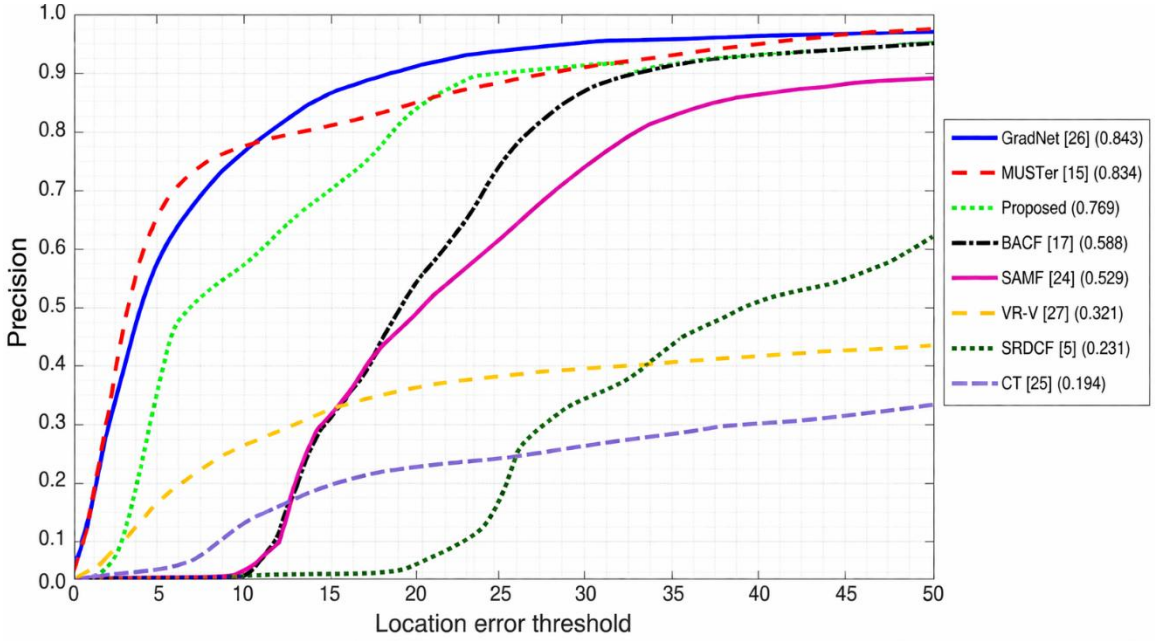

Fig. 5. OPE Illumination Precision Results

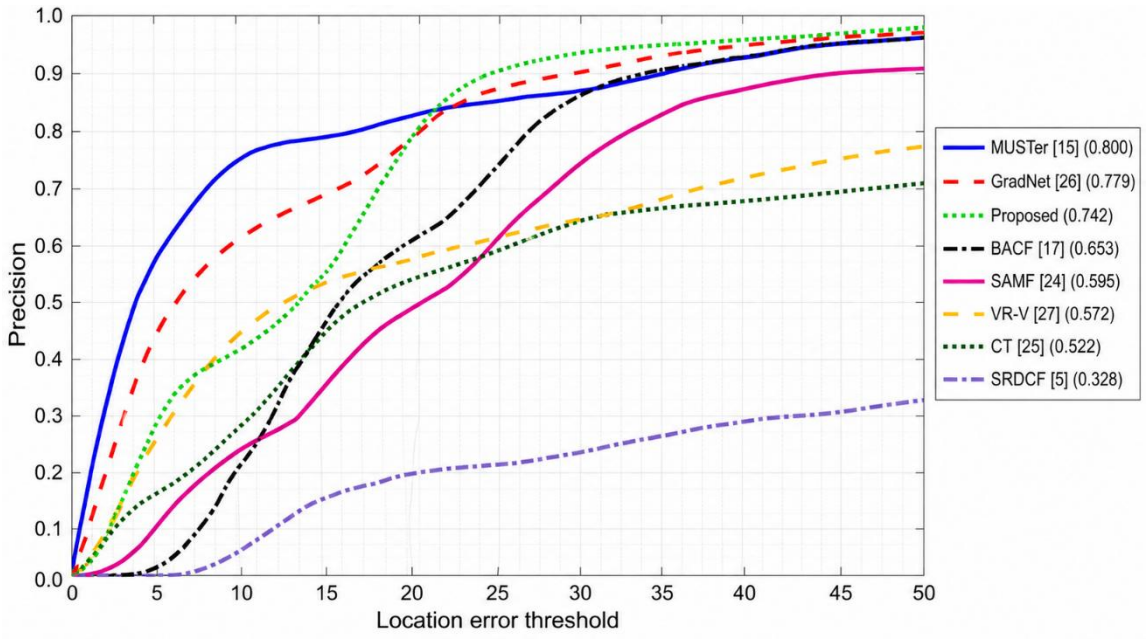

Fig. 6. OPE Precision Plot – Occlusion

Fig. 5 presents the precision performance of the proposed tracker under illumination variation conditions using the OPE evaluation protocol. The results indicate that the proposed tracker achieves competitive localization accuracy compared with representative correlation-filter-based trackers such as MUSTer, BACF, and SAMF. Although GradNet achieves slightly higher precision performance under illumination variation, the proposed feature fusion strategy demonstrates stable tracking behaviour by effectively combining gradient-based structural features with deep semantic representations. This confirms the robustness of the adaptive fusion framework against brightness fluctuations and appearance inconsistencies. Fig. 6 illustrates tracker performance under occlusion conditions. The proposed tracker maintains consistent localization accuracy even when partial target visibility is reduced. Compared with baseline trackers such as BACF, SAMF, and CT, the proposed fusion-based representation preserves discriminative feature information during temporary visibility loss. Although MUSTer achieves slightly higher precision values, the proposed method demonstrates reliable recovery capability through adaptive feature weighting during occlusion scenarios.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

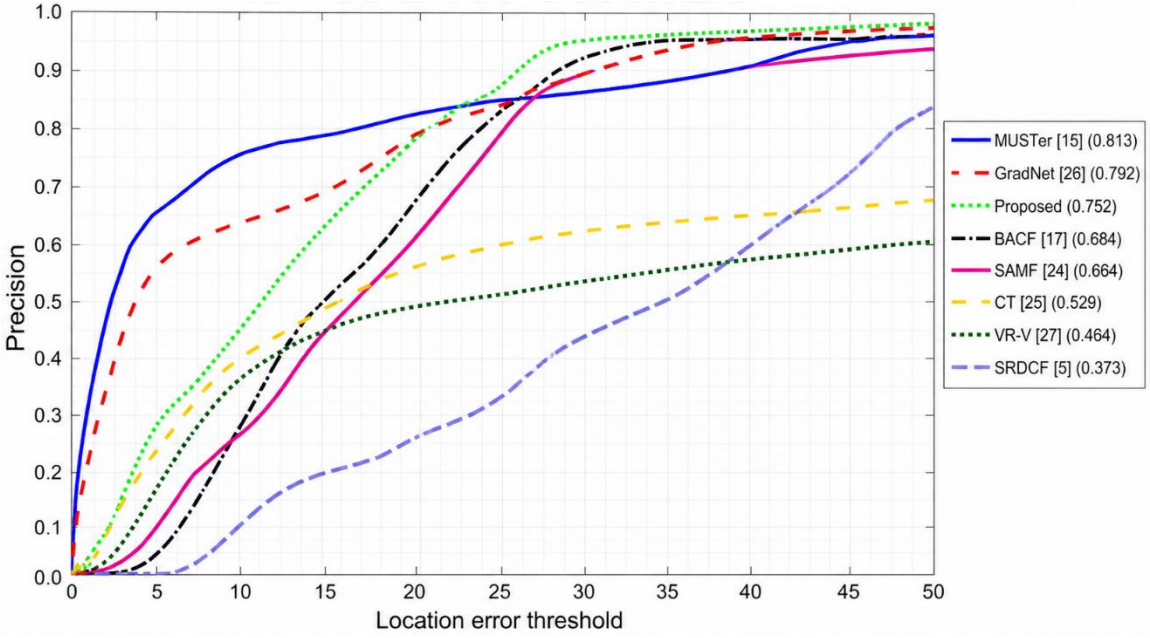

Fig. 7. OPE Scale Variation Precision Results

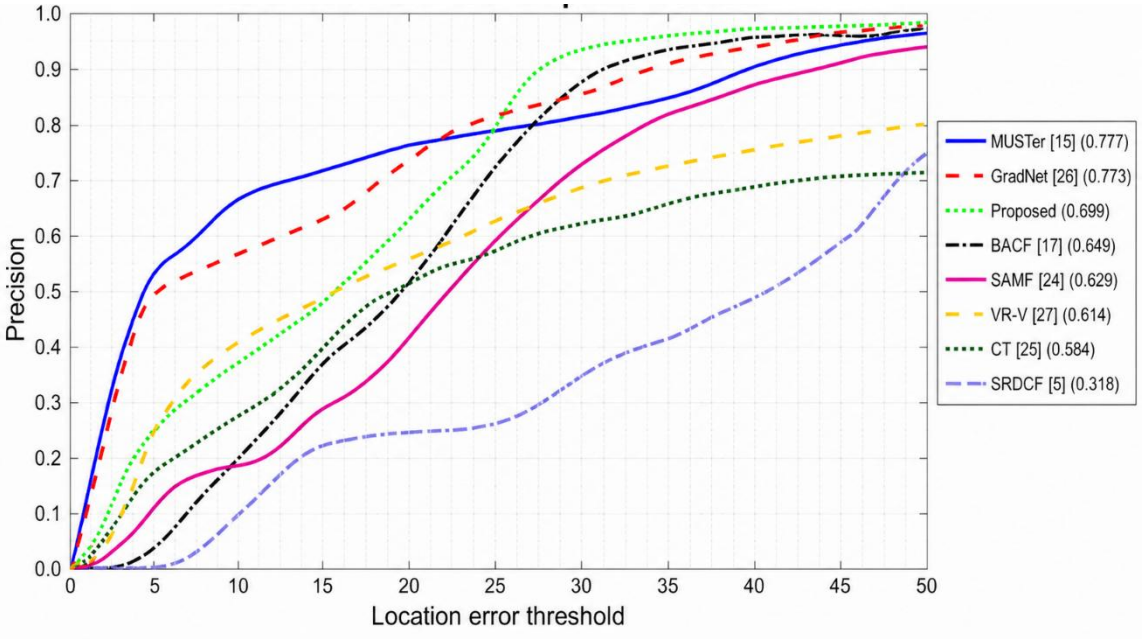

Fig. 8. OPE In-Plane Rotation Precision Results

Fig. 7 shows precision evaluation results under scale variation conditions. The proposed tracker effectively adapts to target size changes using the integrated scale estimation mechanism within the correlation-filter framework. The adaptive fusion strategy enables improved robustness compared with conventional handcrafted-feature trackers such as CT and SAMF. While MUSTer achieves marginally higher precision performance, the proposed tracker maintains competitive localization accuracy across varying scale levels. Fig. 8 presents tracker performance under in-plane rotation conditions. The proposed fusion-based tracking framework demonstrates improved robustness against rotational appearance changes through the complementary integration of HOG and ResNet features. Compared with baseline correlation-filter trackers such as BACF, CT, and SRDCF, the proposed method achieves improved precision stability across different rotation levels, confirming the effectiveness of adaptive feature fusion in handling geometric transformations.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

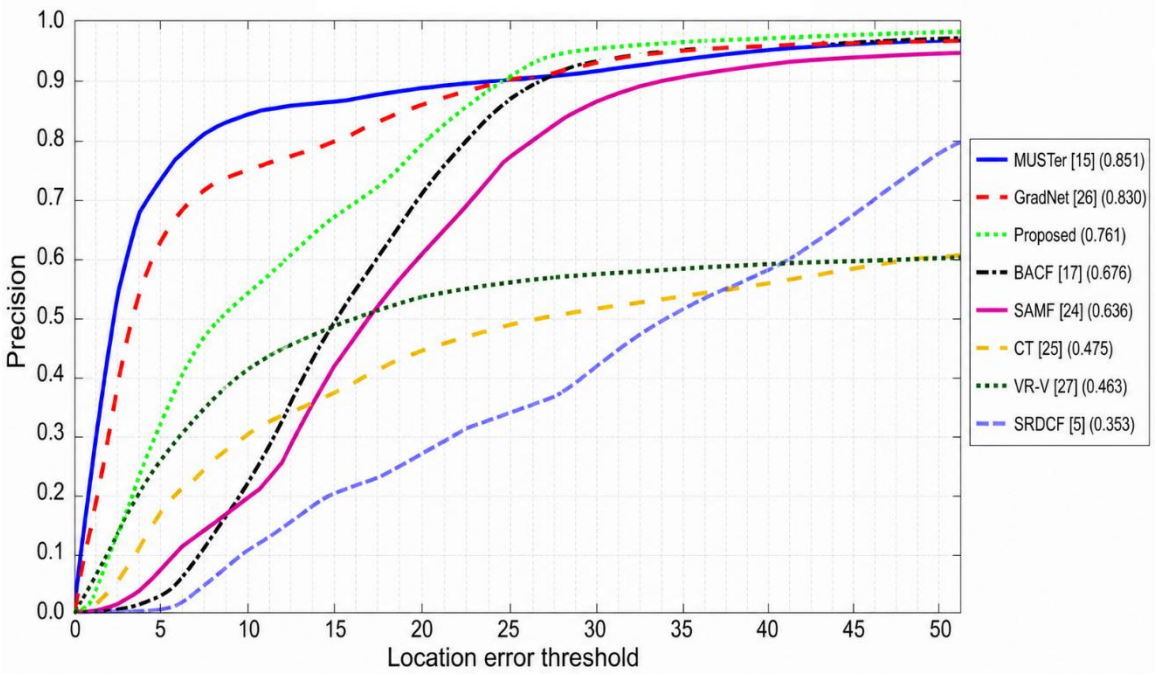

Fig. 9. OPE Overall Precision Results

Fig. 9 presents the overall precision comparison across the evaluated OTB-2015 tracking sequences using the OPE protocol. The proposed tracker achieves competitive overall precision performance relative to representative baseline trackers including BACF, SAMF, CT, and SRDCF. Although MUSTer and GradNet achieve slightly higher precision values, the proposed adaptive fusion framework demonstrates balanced performance across multiple tracking challenges, confirming its effectiveness as a hybrid semantic–geometric tracking strategy.

5 Conclusions

The HOG-CNN fusion tracking framework optimally balances unique handcrafted and deep learning features for precise and efficient tracking. HOG descriptors and CNN features provide complementary structural and semantic representations that improve discriminative tracking performance. The system's stability across appearance variation and scale changes is supported by the APCE-based template update mechanism and confidence-driven adaptive model updating strategy. Experimental evaluation on benchmark tracking sequences from OTB-2015 demonstrates that the fusion-based tracker achieves improved performance compared with individual feature-based tracking approaches. The proposed HOG–CNN tracking framework provides a balanced solution for visual tracking applications requiring accuracy, adaptability, and computational efficiency.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS

The authors declare no conflicts of interest related to this study.

References

[1] M. Kristan *et al*., "The Visual Object Tracking VOT2015 Challenge Results," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, Chile, 2015, pp. 564-586, doi: 10.1109/ICCVW.2015.79.

[2] Y. Wu, J. Lim and M. -H. Yang, "Object Tracking Benchmark," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834-1848, 1 Sept. 2015, doi: 10.1109/TPAMI.2014.2388226.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

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

[4] J. F. Henriques, R. Caseiro, P. Martins and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583-596, 1 March 2015, doi: 10.1109/TPAMI.2014.2345390.

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

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

[7] S. Hu, L. Sun and H. Yu, "Accurate Visual Tracking with Attention Feature Fusion," *2021 26th International Conference on Automation and Computing (ICAC)*, Portsmouth, United Kingdom, 2021, pp. 1-6, doi: 10.23919/ICAC50006.2021.9594244.

[8] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, Nov. 1991, doi: 10.1007/bf00130487.

[9] D. Wang, W. Fang, W. Chen, T. Sun, and A. T. Chen, "Model Update Strategies about Object Tracking: A State of the Art Review," *Electronics*, vol. 8, no. 11, p. 1207, Oct. 2019, doi: 10.3390/electronics8111207.

[10] M. Masood and G. Raja, "An adaptive learning based aberrance repressed multi-feature integrated correlation filter for Visual Object Tracking (VOT)," *Mehran University Research Journal of Engineering and Technology*, vol. 43, no. 4, p. 14, Oct. 2024, doi: 10.22581/muet1982.2832.

[11] W. Xing, *et al*., "Correlation filter based visual object tracking," in *Visual Object Tracking from Correlation Filter to Deep Learning*, Singapore: Springer, 2021, ch. 3. doi: 10.1007/978-981-16-6242-3_3.

[12] S. Du and S. Wang, "An Overview of Correlation-Filter-Based Object Tracking," in *IEEE Transactions on Computational Social Systems*, vol. 9, no. 1, pp. 18-31, Feb. 2022, doi: 10.1109/TCSS.2021.3093298.

[13] S. Jiang, S. Li, C. Zhu and N. Yan, "Efficient correlation filter tracking with adaptive training sample update scheme," *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 549-555, doi: 10.1109/ICPR48806.2021.9413005.

[14] B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8971-8980, doi: 10.1109/CVPR.2018.00935.

[15] 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," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 749-758, doi: 10.1109/CVPR.2015.7298675.

[16] H. K. Galoogahi, T. Sim and S. Lucey, "Correlation filters with limited boundaries," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 4630-4638, doi: 10.1109/CVPR.2015.7299094.

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

[18] C. Zhu, S. Jiang, S. Li, and X. Lan, "Efficient and practical correlation filter tracking," *Sensors*, vol. 21, no. 3, p. 790, Jan. 2021, doi: 10.3390/s21030790.

[19] Y. Xiao, Y. Wu, and F. Xu, "An adaptive correlation filtering tracker with dual feature channels," *Journal of Visual Communication and Image Representation*, vol. 111, p. 104556, Aug. 2025, doi: 10.1016/j.jvcir.2025.104556.

[20] H. Fan and H. Ling, "Parallel Tracking and Verifying: A Framework for Real-Time and High Accuracy Visual Tracking," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 5487-5495, doi: 10.1109/ICCV.2017.585.

[21] P. Liang, E. Blasch and H. Ling, "Encoding Color Information for Visual Tracking: Algorithms and Benchmark," in *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5630-5644, Dec. 2015, doi: 10.1109/TIP.2015.2482905.

[22] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Lecture notes in computer science*, 2016, pp. 445–461. doi: 10.1007/978-3-319-46448-0_27.

[23] S. Hu, Y. Ge, J. Han, and X. Zhang, "Object Tracking Algorithm Based on Dual Color Feature Fusion with Dimension Reduction," *Sensors*, vol. 19, no. 1, p. 73, Dec. 2018, doi: 10.3390/s19010073.

[24] Y. Li and J. Zhu, "A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration," in *Lecture notes in computer science*, 2015, pp. 254–265. doi: 10.1007/978-3-319-16181-5_18.

[25] K. Zhang, L. Zhang, and M.-H. Yang, "Real-Time Compressive Tracking," in *Lecture notes in computer science*, 2012, pp. 864–877. doi: 10.1007/978-3-642-33712-3_62.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 23-35 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 31 Jan 2026 | Revised: 10 Apr 2026 | Accepted: 21 Apr 2026 | Published: 28 Apr 2026

[26] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang and H. Lu, "GradNet: Gradient-Guided Network for Visual Object Tracking," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 6161-6170, doi: 10.1109/ICCV.2019.00626.

[27] R. T. Collins, Yanxi Liu and M. Leordeanu, "Online selection of discriminative tracking features," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1631-1643, Oct. 2005, doi: 10.1109/TPAMI.2005.205.