Systems and Soft Computing 7 (2025) 200417

Contents lists available at ScienceDirect

Systems and Soft Computing

journal homepage: www.journals.elsevier.com/soft-computing-letters

COA-SSD: A novel chaotic optimization algorithm for enhancement of real-time video surveillance images using single shot detector with MobileNetV2

Vivek Pandiya Raj [*,a], K. Radha [b], S. Harihara Gopalan [c], C. Venkataramanan [d], R. Dhanapal [e], A. Manikandan [*,f]

[a] Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu 600062, India
[b] Vivekanandha College of Engineering for Women, Namakkal, Tamil Nadu 637205, India
[c] Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu 641022, India
[d] Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu 642109, India
[e] Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu 641021, India
[f] Assistant Professor/ECE, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu 603203, India

ARTICLE INFO

Keywords:
Deep learning models
Sobel filter
Local-peak scale-invariant feature transform (LP-SIFT)
Modulated Memory Network (MMN)
MobileNetV2
Coati Optimization Algorithm (COA)

ABSTRACT

Video surveillance involves the deployment of cameras to observe and record activities for security and monitoring purposes. This is vital for crime prevention, public safety, and improving operational efficiency across various environments. The rising demand for real-time security and threat detection has increased interest in sophisticated video surveillance systems. However, challenges persist in achieving an accurate automated analysis under diverse environmental and situational conditions. This paper presents an optimization strategy for video surveillance system deployment using a deep-learning algorithm. This strategy combines the base image with information about the behavioral events depicted in the frame sequence and then uses a Sobel filter to detect boundaries. The present study employs a rapid feature point identification algorithm known as the local-peak scale-invariant feature transform (LP-SIFT), which is based on multi-scale local peak features that are invariant to scale and represent multi-scale fluid turbulence. Traditional segmentation methods in video surveillance often lack contextual understanding and temporal consistency, which Modulated Memory Networks aim to address using adaptive memory-driven feature modulation. Next, we used an (SSD) with MobileNet V2 to classify fragments. SSDs can bypass the region proposal network and reduce latency. SSDs incorporate several optimizations such as multiple cooperative functions and base boxes to compensate for the loss of accuracy. To optimize and enhance the accuracy and efficiency of the video surveillance classification, the model parameters were fine-tuned. The Coati Optimization Algorithm (COA) is a metaheuristic algorithm inspired by coati behaviors: their cooperative strategy when attacking iguanas, and their anti-predator and avoidance behavior strategies. To evaluate the LP-SIFT algorithm, we considered datasets with varying scenes, pixel sizes, and distortion levels, across different lighting conditions and environments.

1. Introduction

The background, motivation, and major challenges involved in video surveillance are outlined in this section to justify the proposed work.

The primary difficulty in executing VI-ReID is the pattern disparity between visible and infrared images [1]. Typically, the two modes share common characteristics. The first method used image-level techniques. Using this approach, generative adversarial neural networks (GANs) and other neural networks can generate intermediate image or imaginer-correspondingg patterns. Although effective in reducing pattern differences, these methods introduce unnatural noise.

There are many difficulties in identifying [2] and detecting behavioral events in video streams owing to a variety of unique reasons. It is usually difficult to distinguish between the end and starting points of an

* Corresponding authors.
E-mail addresses: drvivekpandiyaraj@veltech.edu.in (V.P. Raj), kradhagokul@gmail.com (K. Radha), hs@srec.ac.in (S.H. Gopalan), manikana3@srmist.edu.in (A. Manikandan).

https://doi.org/10.1016/j.sasc.2025.200417
Received 7 June 2025; Received in revised form 23 October 2025; Accepted 17 November 2025
Available online 20 November 2025
2772-9419/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

event. Some examples of these complications are the fact that an event can be performed in various ways at different points in time by one or more bio-objects and that a bio-object can express two different states simultaneously (e.g., jumping). Even though most convolutional networks deal with single frames, knowing the time context still limits their ability to successfully interpret recorded videos.

In [3] recent years, feature-based stitching methods have become more attractive than computationally intensive region-based techniques owing to their low area overlap requirements. Harris et al. pioneered a procedure for stitching images based primarily on feature detection, whereas Lowe et al. [4] created a scale-invariant feature transform (SIFT) utilizing a Gaussian pyramid along with a Gaussian difference pyramid.

In [5], is presented a modified memory network (MMN) that integrates an importance modifier (IM) and a position modifier (PM). These components are specifically designed to address the aforementioned challenges. This [6] module modifies features from previous frames in memory using adaptive weights for spatial and channel dimensions.

To improve the COA, chaotic mapping [7], dynamic learning, cross-section, and Levy flight strategies have been adopted. Initially, chaotic mapping is applied during the population initialization phase by employing chaotic sequences to create higher-quality populations. Subsequently, a lens-based learning approach was used [8].

The main objectives of this study were as follows:

- The Sobel filter is commonly employed for edge detection. It is particularly useful in video surveillance for identifying shifts in pixel intensity that signify edges, such as outlines of objects, motion, or atypical activities.
- LP-SIFT (Local Pattern-SIFT) builds on the standard SIFT algorithm by incorporating local texture patterns, which enhances its ability to withstand variations in lighting and orientation. This adaptive method can improve target recognition and tracking accuracy in complex dynamic video surveillance environments.
- The Modified Memory Network (MMN) consistently achieves high performance compared to state-of-the art methods, demonstrating the effectiveness of both the IM and PM modules, as confirmed by ablation studies.
- The upgraded SSD MobileNetV2 combines the MobileNetV2 architecture with SSD enhancements to enable real-time object detection in video surveillance.
- The enhanced COA was based on four primary strategies: chaotic mapping, dynamic lens imaging learning, Levy flight, and intersection.

2. Literature survey

Currently [9], the two main research directions for monotype reidentification are representation and analogical learning. Representation learning is dedicated to enhancing a network's ability to differentiate pedestrian representations. In contrast, metric learning improves the relative distances between images in the network by placing images with the same identity closer to those with different identities. Deep learning methods [10] improve the performance of single-model person re-identification (Re-ID) performance. However, VI-ReID must address the differences between images from the two modalities. Methods that achieve good results in a single modality cannot be directly applied to multimodality person re-identification.

Single-mode person re-identification [11–13] aims to recognize and retrieve images of individuals captured only under visible-light conditions. This task often involves occlusions, illumination, and viewpoint changes. To address these challenges, methods use CNNs such as ResNet27 to extract unique features. Recently, vision transformers (ViTs) have emerged for image analyses. He et al. [14] applied it to single-modality person re-identification to improve feature dependencies.

Wang et al. [15,16] classified deep learning-based single-shot VOS techniques into three categories based on how they use the initial annotated frames during inference. OSVOS, OSVOS-S, and AVOS utilize image segmentation techniques and refine a comprehensive network to concentrate on objects present in the initial frames [17]. These methods update their segmentation models online to manage objects that change in size over time.

In this study, we combined chaotic mapping, dynamic lens imaging learning, intersection, and Levi's flight strategies to improve the COA [18,19]. A chaotic mapping strategy using chaotic sequences constructs the population to ensure quality during initialization. In addition, we added a lens-based dynamics learning mechanism to enhance population diversity and promote navigation over a larger area [20].

3. Proposed methodology

The framework's design, which includes preprocessing, LP-SIFT, MMN, SSD-MobileNetV2, and COA optimization, is mapped to the proposed method's role. This method uses Sobel filtering to enhance the edges of video frames, offering clearer object boundary definitions for more effective motion detection and segmentation in surveillance videos.The MMN performs adaptive modulation of memory by considering segmentation confidence together with spatiotemporal position information. It uses an enhanced SSD MobileNetV2 architecture for fast and accurate real-time object detection in video surveillance, blending lightweight architectures with advanced feature extraction capabilities. The Coati optimization algorithm optimizes the parameters in video surveillance systems for enhanced target detection and tracking.

3.1. Preprocessing

Considering the temporal context, the model assesses animal behavior based on its location, past migration routes, and predicted future routes.

The two labeled images were resized to 224 × 224 pixels and merged using a Sobel filter. This filter calculates image derivatives along the horizontal and vertical axes and emphasizes brightness changes at object edges and contours [21]. The Sobel filter increased the contrast between the dynamic and static animals, thereby improving the data quality for training the model (Fig. 1).

3.2. Feature extraction

3.2.1. Local-peak scale-invariant feature transform (LP-SIFT)

The LP-SIFT technique illustrated in Fig. 2 comprises five stages: image preprocessing, feature point detection, point description, description, point matching, and image stitching. First, each stage is explained, followed by the application of LP-SIFT for the reconstruction of larger images from randomly created fragments [22].

3.2.2. Image preprocessing

The images in Fig. 3 are linked as a reference image (stored as matrix $M1$) and a registration image (stored as matrix $M2$). In the LP-SIFT algorithm, because the feature points correspond to local peaks, it is numerically beneficial to first apply a small linear background to both images. This preprocessing step improves the detection of feature points, particularly in regions of uniform intensity caused by image saturation.

The modified image matrices $M_{n,k}$ are defined by:

$$M_{n,k}(i,j) = M_k(i,j) + [(i-1) * nc_k + j] * \alpha \quad (1)$$

where $\alpha \ll 1$is the linear noise coefficient. nc_k is the number of columns in image $M_{n,k}$, $k = 1$refers to the reference image, and $k = 2$ refers to the registration image, i and j denote the row and column indices, respectively.

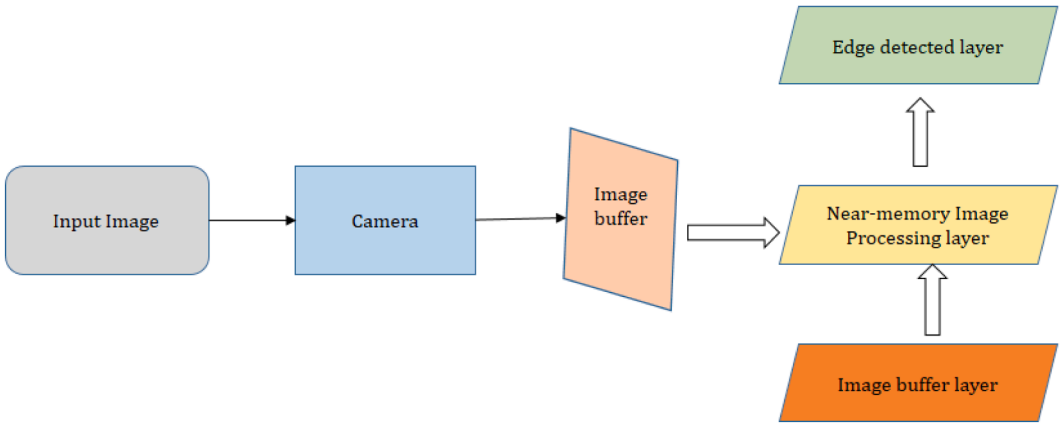


Fig. 1. The Sobel filter architecture is used for edge enhancement during preprocessing.

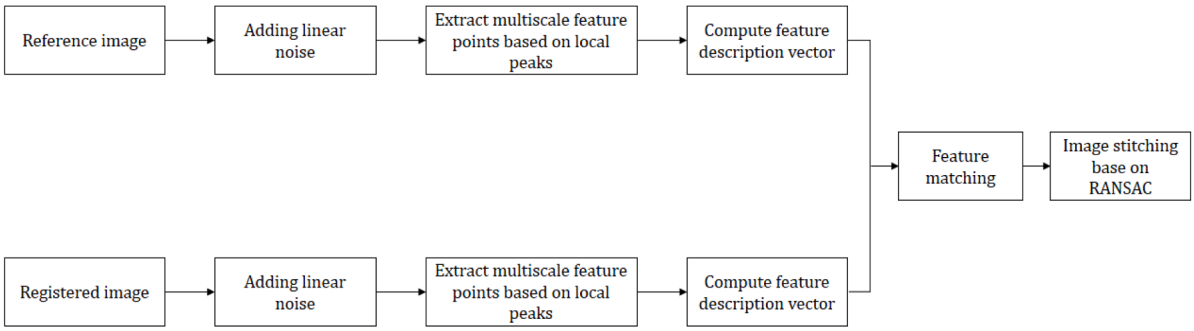


Fig. 2. Framework of the proposed Local Peak Scale-Invariant Feature Transform (LP-SIFT) method.

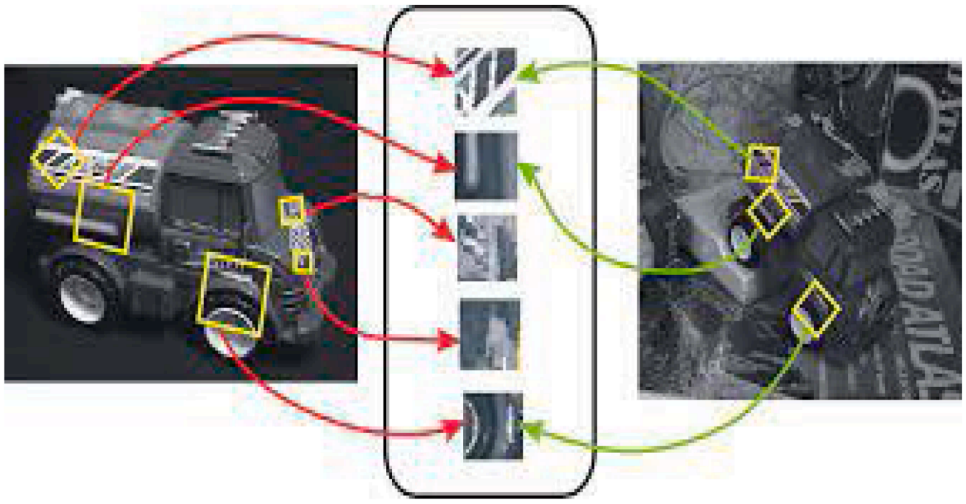


Fig. 3. Illustration of proposed LP-SIFT algorithm showing feature point detection and image registration steps.

3.2.3. Feature point detection

In Step 2 of the LP-SIFT algorithm, local peaks in multiscale images are selected as feature points. The reference and registered images were divided into square query windows of size L, with L ranging from 32 to 128 pixels. For each window, the maximum and minimum pixel intensities were extracted and used as feature points [23].

This operation can be expressed as

$$\begin{cases} M_{n,k,\max(p,L)=\max\left[M_{np,k}\ (i,j)\right].ij\ \in[0,L]} \\ M_{n,k,\max(p,L)=\max\left[M_{np,k}\ (i,j)\right].ij\ \in[0,L]} \end{cases} \quad (2)$$

where $M_{np,k}\ (i,j)$is the p^{th} interrogation window of $L \times L$ pixels,

3.3. Segmentation

3.3.1. Modulated memory network (MMN)

During the memory phase, the prior frame and its associated object mask are utilized to create a modulated memory for future segmentation tasks. These frames are encoded and saved in the memory As shown in Fig. 4 for historical frame I^i and its object mask P^i, a memory encoder *EnM* is first applied, which takes input of both Ii and Pi and output key

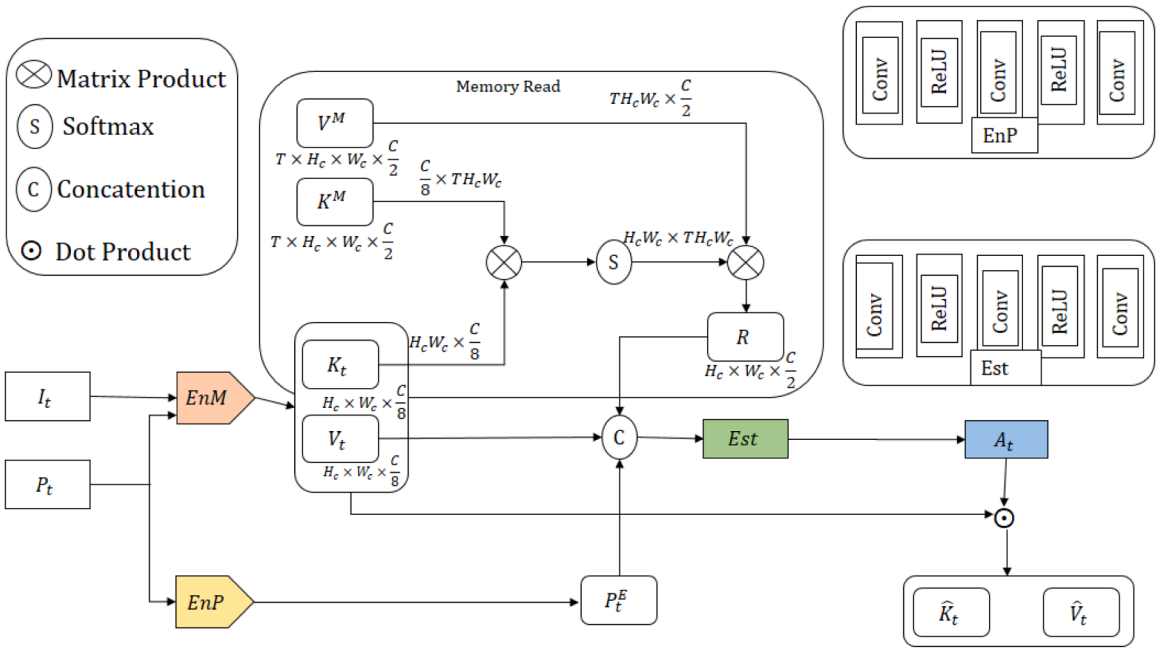


Fig. 4. Architecture of the Modulated Memory Network (MMN) used for temporal consistency in video segmentation.

feature $K_M^i \in R^{HW \times C_k}$ and value feature $V_M^i \in R^{HW \times C_v}$, C_K and C_V are in MMN is:

- C_K =128, the channel dimension of the key feature
- C_V = 512, the channel dimension of the value feature

The spatial dimensions are defined as

Spatial dimensions are $H = sH_{im}$and $W = sW_{im}$, where H_{im} and W_{im} correspond to the spatial dimension of the input frames. $S = 1/16$ is the downsampling stride of the backbone (ResNet-50, following STM).

3.3.1.1. Motivation for modulated memory. Previous solutions store K_M^i and V_M^i to memory directly. The IM module takes as input the value feature V_M^i, a readout feature $R_M^i$and the object mask P^i, and it outputs a weight map A^i.

Affinity matrix calculations

To obtain R_M^i, the features of $\{I_j\}_{j=1}^{i-1}$, i.e., key features $K_M^{i-1} = \left\{k_M^j\right\}_{j=1}^{i-1}$and value features $V_M^{i-1} = \left\{V_M^j\right\}_{j=1}^{i-1}$, are taken as a reference. An affinity matrix *Ai* is first calculated between the key features. For each spatial location p in K_M^i and q in $K_M^{i-1} = \left\{k_M^j\right\}_{j=1}^{i-1}$,

$$A(p,q) = \frac{\exp\left(S_{p,q}\right)}{\sum_q \exp\left(S_{p,q}\right)} \tag{3}$$

where similarity score $s_{p,q}$is computed via:

$$s_{p,q} = c\left(K_M^i(p), K_M^{i-1}(q)\right) \tag{4}$$

Here, c (·,·) denotes the dot product function.

Then, R_M^i is aggregated through the weighted sum of value features $V_M^{i-1} = \left\{V_M^j\right\}_{j=1}^{i-1}$,

$$R_M^i(p) = \sum_q \left(A(p,q) \cdot V_m^{i-1}(q)\right) \tag{5}$$

The IM module can learn to estimate the reliability of frame I^i by comparing R_M^i with V_M^i. Afterward, we adopt *EnP* to encode I^iinto feature P_E^i, of which the number of channels is 32. Then, we concatenate P_E^{i-1}, $R_M^i$and V_M^i together along the channel dimension, and we input it into an estimator *Est* and output the modulation map A^i. Finally, modulation for frame I^i is conducted as follows, for each position p in V_M^i in Eq. (6):

$$V_M(p) = A(p) \cdot V_m^i(p) \tag{6}$$

Modulation occurs in the value features carrying the target object label information. They provided experimental results by comparing modulation modes. After modulation, modulated features of historical frames $\{I_j\}_{j=1}^{i-1}$ stack along the temporal dimension, constructing a modulated memory. $M^i = \left\{K_m \in R^{NHW \times C_K}, V_M \in R^{NHW \times C_v}\right\}$.

In the segmentation process, the MMN uses the modulation memory M^{i-1} to segment the current frame *pIt*. Here, the supporting features are searched by matching. The position modifier (PM) encodes the spatio-temporal position information of the key features into the memory and current frames to improve matching accuracy. This helps to better utilize the first and last keyframes [24]. The coordinates of each dimension (height, width, and time) were encoded into the embedding using sine and cosine functions at different frequencies.

$$PE(p,c) = \begin{cases} sin\left(Norm(p) \bullet w_k\right), & for\ c = 2k \\ cos\left(Norm(p) \bullet w_k\right), & for\ c = 2k+1 \end{cases} \tag{7}$$

where p is the coordinate in dimension, $w_k = \frac{1}{10{,}000^{2k}/\frac{d}{3}}$ and c is channel index. The norm is a normalization function. The spatial coordinates are sequentially numbered, with normalization function defined as Eqn (8)

$$Norm_R(p) = \frac{p}{\max\limits_{p \epsilon\ p\ (p))}} \tag{8}$$

For the temporal dimension, we define temporal coordinates as *Pt*={2,3,4,…,*t*−1,1,1} for each video $s = \{I_1, I_2, I_3, \ldots, I_{t-2}, I_{t-1}, I_t\}$. The normalization function for the temporal coordinates is defined as (9).

$$Norm_R(p) = \frac{1}{p},\ p \epsilon\ p \tag{9}$$

The custom encoding method fixes the position between the current frame and the keyframes, allowing access to key reference information based on the temporal position. Position embeddings for each channel

dimension are combined and added element-wise to key features k_M^i and k_q^t, resulting in modulated key features. k_M^i k_q^t,$k^{NHW\ \times C_t}$ is the key feature of current frame I^i, encoded by *EnQ*.

After obtaining the features, the affinity matrix is determined using (5). The extracted supporting features from the modification memory through weighted pooling are shown in Eq. (8). The decoder Dec generates mask prediction for frame *It* by decoding retrieved support features.

3.4. Classification

3.4.1. Improved SSD Mobilenetv2

A single-shot detector (SSD) enables high-speed real-time object detection without requiring a region proposal. To counter any possible precision drop, the SSD includes supplements, such as several co-features and default boxes. With these improvements, it can be handled easily by lower-resolution images during SSD processing, similar to that of Faster R-CNN. It is aptly termed as a single-shot detector because it captures only its true essence. Unlike other models of the same family that require several passes over an image to realize detection, SSD typically does it in one pass, hence speediness. Despite this speed, it retains high detection capability. It classifies the predictions by view size, generating others at different mAP scales to maintain high detection accuracy. mAP features were extracted from SSD backbone using a pre-trained classification network, and features were later processed in the SSD head [25]. The mAP of the features generally persisted once the last image classification layer was removed from the model. The SSD head implements many more convolution layers built on top of the main model. These layers can draw bounding boxes around the objects in the image. The composite thinning operation is key to differentiating between different objects. Sigma activation in the last layer is reduced from 80 to 5, which constitutes a typical mobile phone terminal. The remaining architecture was realized using 53 artificial stacks along with batch normalization and activation functions; thus, the thinning layers were excluded from the consideration of being nonlinear. Even then, with 53 layers, the model finds it difficult to recognize small objects such as personal items, vehicles, and bicycles. To overcome these problems, we implemented a feature pyramid extractor, as shown in Fig. 5. Given an input image of arbitrary size, it generates specific access points to features of decreasing size through a fully convolutional network. This method is independent of the convolutional architecture used in the backbone, and bolsters the learning ability of the model for detecting small objects. To avoid overfitting, we added additional training samples and reduced the batch size to eight. This improved the accuracy and recall while conserving memory resources.

3.5. Optimization

3.5.1. Coati optimization algorithm

The originality of the Coati Optimization Algorithm (COA) lies in its biologically inspired design, which mimics the cooperative hunting and predator-avoidance behavior of coatis. Unlike conventional metaheuristics such as Particle Swarm Optimization or Grey Wolf Optimization, COA integrates four adaptive strategies—chaotic mapping, dynamic lens-based learning, Lévy flight, and intersection-based exploration—that collectively improve convergence speed and prevent premature stagnation in local minima. This hybrid structure enables COA to dynamically balance exploration and exploitation during parameter tuning, ensuring global optimality in complex, high-dimensional search spaces. Practically, this results in more accurate and stable hyperparameter selection for deep learning-based video surveillance models, leading to superior detection accuracy, faster training convergence, and enhanced real-time performance under diverse environmental conditions. The adaptability and computational efficiency of COA demonstrate its potential as a general-purpose optimizer for other real-time computer vision and intelligent monitoring applications.

The Coati Optimization Algorithm (COA) is a 2023 nature-inspired heuristic method. The governing rules of COA stem from the hunting strategies and predator avoidance behaviors of coatis. In COA, every coati does not have a fixed position, but acts as an independent search agent that moves within the p-dimensional search space.

The algorithm simulates the individual-level decision-making of coatis while searching for prey and escaping predators.

Each coati's position is progressively adjusted to move closer to the globally optimal position, which represents the best possible solution, by trailing the iguanas (leaders) and investigating new areas within the search space.

3.5.2. Initialization process

To begin with, the COA generates m random individuals using Eq. (10), where the maximum boundary for individual values is $X_{\max} = x_j^{max} - x_p^{min}$ and the minimum boundary is $X_{\min} = x_j^{max}, \ldots, x_p^{min}$. Evaluate the initialized random individuals using the objective function.

$$x_{i,j} = x_j^{min} + rand \cdot \left(x_j^{max} - x_j^{min}\right), i = 1, 2, \ldots, m, j = 1, 2, \ldots, p \tag{10}$$

3.5.3. Hunting and attack strategies (Exploration phase)

$$x_{i,j}^{new} = x_{i,j} + b \cdot \left(Igu_j - I \cdot x_{i,j}\right), j = 1, 2, \ldots, \left\lceil \frac{m}{2} \right\rceil \tag{11}$$

When an iguana becomes startled and descends, its landing spot is randomly chosen. In response, the remaining coatis on the ground alter their movements to align with iguana's newly assigned random position. These behaviors are expressed in Eqs. (10) and (11). Schematic showing the updated positions of iguana.

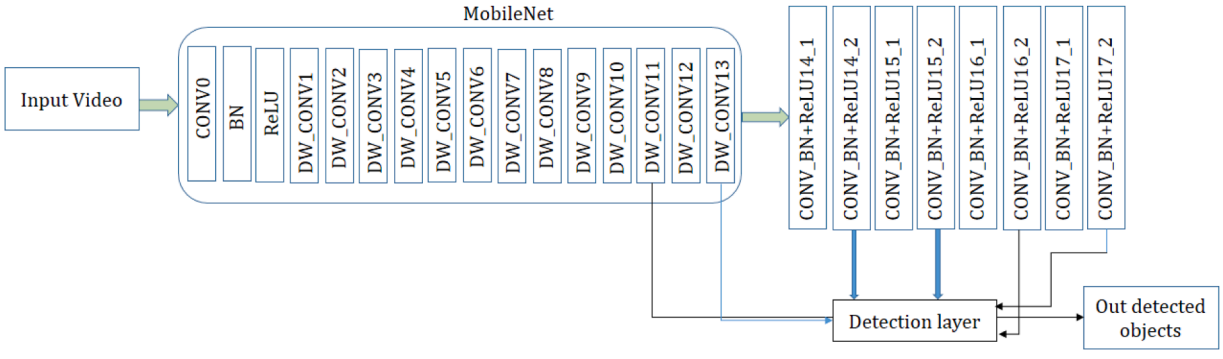


Fig. 5. Improved SSD MobileNetV2 architecture for real-time object detection in video surveillance.

$$Igu_j = x_i^{min} + b \cdot \left(x_j^{max} - x_j^{min}\right), j = 1, 2, \ldots, p \tag{12}$$

$$x_{i,j}^{new} = \begin{cases} x_{i,J} + b \bullet \left(Igu_j - I \cdot x_{i,J}\right), F(Igu) < F_I \\ x_{i,J} + b \bullet \left(x_{i,J} - Igu_j\right), \ otherwise \end{cases}, \ i = \left[\frac{m}{2}\right] + 1, \ldots, m: \ j = 1, \ 2, \ \ldots \ ., \ p \tag{13}$$

where $x_{i,J}$ represents thej^{th} dimension. Igu_j is the j-th dimension of the iguana's position. F_{Igu} is the fitness of the iguana's position. F_I is the fitness of the current coati position. I∈ {1,2}, b∈ [0,1]. Fig. 6 show as in Architecture of Coati Optimization Algorithm

3.6. Evaluation metrics

The confusion matrix uses four metrics: True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR), as Eqs. (14–20).

$$TPR = \frac{Truepositive}{Truepositive + FalseNegative} \tag{14}$$

$$FPR = \frac{Falsepositive}{Falsepositive + TrueNegative} \tag{15}$$

$$TNR = \frac{TrueNegative}{Falsepositive + TrueNegative} \tag{16}$$

$$FNR = \frac{Falsepositive}{Truepositive + FalseNegative} \tag{17}$$

$$Precision = \frac{Truepositive}{Truepositive + Falsepositive} \tag{18}$$

$$Recall = \frac{Truepositive}{Truepositive + Falsepositive} \tag{19}$$

$$Recall = \frac{2}{Precision^{-1} + Recall^{-1}} \tag{20}$$

Therefore, a True Positive cannot be considered simply positive. The misidentifications of positive are the so-called false positives, whereas true negatives are those correctly identified as non-positive.

4. Results and discussion

We present empirical findings, compare the performance with the baselines, and analyze the strengths and limitations of our proposed approach.

The model was constructed using a setup involving two datasets to evaluate the data preprocessing methods. The first dataset relied on traditional image classification and segmentation techniques, whereas the second dataset employed the preprocessing approach implemented in this study.

4.1. Datasets

Datasets with various scenarios, pixel resolutions, and distortion levels were used to examine the efficiency of the LP-SIFT algorithm. Datasets A and B consist of three well-known image pairs from previous studies for Dataset A: (1) mountain scenes, (2) street views, and (3) terrain images; Controller B includes three image pairs taken with a PGKM10 (OnePlus) mobile phone camera with six-megapixel resolution. Dataset B pairs were: (1) building view, (2) campus view with translational movement, and (3) campus view with rotational variation. Figs. 7 and 8 show sample images from Datasets A and B, respectively. Fig. 9 presents the traditionally labeled and preprocessed images generated using the recommended methodology.

This procedure prioritizes stitching the reference images with the most adjacent combinable images. In this initial stage is concatenate the reference and matching images and keep the non-matching images separately in a non-matching set. A new reference image was selected from the unmatched collection using HM criteria. This process is repeated until the unmatched collection is empty, signaling the completion of the mosaic. Fig. 10a shows the original images captured in this experiment. Fig. 10a shows the original image (6400 × 4270). Fig. 10b shows the segmented images extracted from the image. Fig. 10c Result of the stitching process; the time taken for stitching is 158.94 s

Fig. 11 presents the Training and Validation Accuracy and Loss curves of MobileNetV, showing that rapid improvement performance occurred at the beginning epochs followed by stable curves with above 98 % accuracy, strong learning, and effective generalization. The alignment of the training and validation accuracy curves indicated a reduced overfitting. The loss curve showed a sharp downward trend during the first 10 training and validation epochs, after which the loss approached zero, indicating effective convergence.

The mountain dataset, the computation times are recorded as follows: SIFT requires 101.21 s, ORB completes in 0.71 s, BRISK takes 1.30 s, SURF finishes in 0.85 s, and LP-SIFT needs 1.16 s. Among these, ORB was the quickest, whereas SIFT was the slowest. The comparison is illustrated in Fig. 12. Alternatively, the virtual tour

Computation time for the dataset: SIFT 226.62 s, ORB 2.27 s, BRISK 2.22 s, SURF 2.54 s, and LP-SIFT 2.05 s.

The confusion matrix evaluates the performance of classification models in video surveillance applications. True Positives (TP) are events detected by the system, such as the detection of a present person. False Positives (FP) are events in which system claims exist, but they do not, resulting in a false alarm. True Negatives (TN) pertain to unwanted

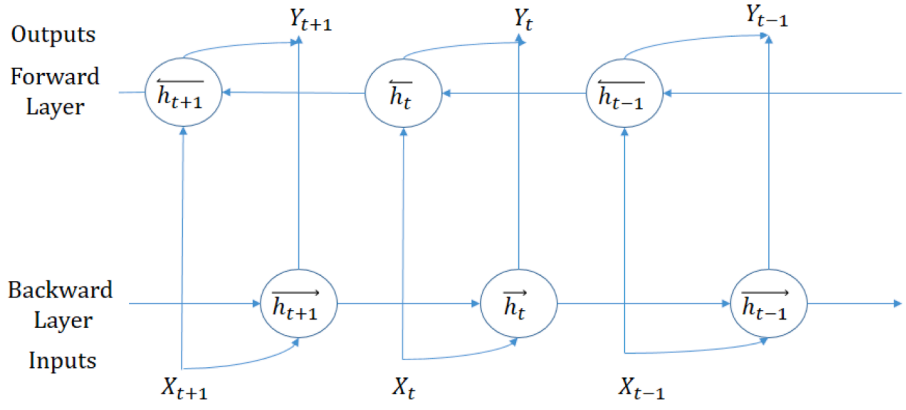


Fig. 6. Architecture of the proposed Coati Optimization Algorithm (COA) illustrating hunting and avoidance strategies.



Fig. 7. Dataset A: (a) mountain scene, (b) street view, (c) terrain image.



Fig. 8. Dataset B: (a) building view, (b) campus view (translation), and (c) campus view (rotation).

events that are ignored, such as when no one is present. False Negatives (FN) occur when an event, such as a present person, fails to be recognized by the system. The confusion matrix for the proposed method is shown in Fig. 13.

In video surveillance applications, TPR refers to the correct identification of events, such as the detection of real intrusions, whereas FPR corresponds to false alarms. The upper-left curve represents the better models. The AUC is a summary measure of model performance, with 1.0 representing perfection and 0.5 representing chance. The ROC curve is crucial for determining the best thresholds to achieve a balance between sensitivity and specificity, which is vital in surveillance to minimize both missed detections and false alarms. Fig. 14 shows the ROC curve for the classification process of the proposed method for video surveillance.

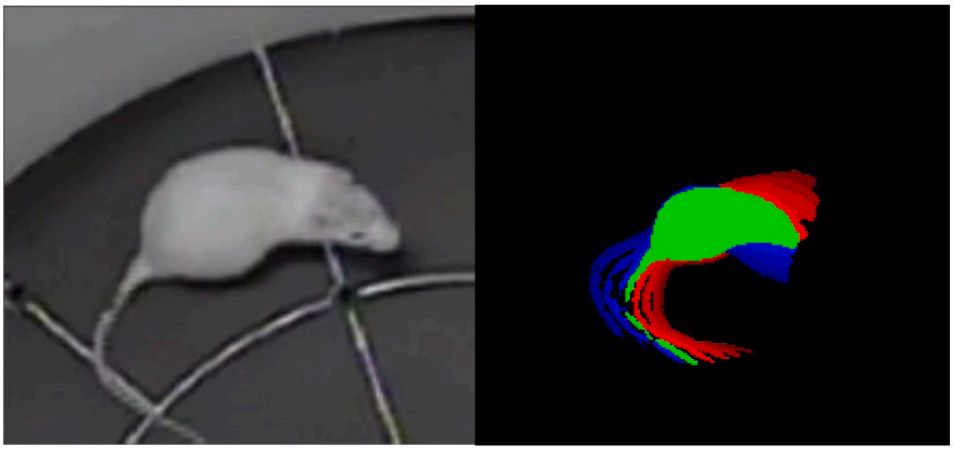


Fig. 9. Color-coded annotations showing temporal variable changes over time.



Fig. 10. Results of the LP-SIFT image stitching process: (a) original images, (b) segmented image sets, and (c) final mosaic (processing time = 158.94 s).

4.2. Ablation study

Validation set ablation tests were performed to determine the contribution of each part of the proposed MMN. Direct training from the data of the training set, with no pre-training, allows for faster validation. Based on the context importance and hash confidence, the IM module modifies the memory frame.

Other explored adjustment techniques include the evaluation of models that apply adjustments to base features only, value features only, or both. As shown in Table 1, the performance improved when the value feature was adjusted. This is because the value feature is primarily used to encode the information related to the prediction mask. This prevented the propagation of silent errors. Conversely, important features encode the appearance information, which is primarily used for similarity assessment. Any alteration to the essential features could lead to improper association of the object and, consequently, reduced accuracy in segmentation.

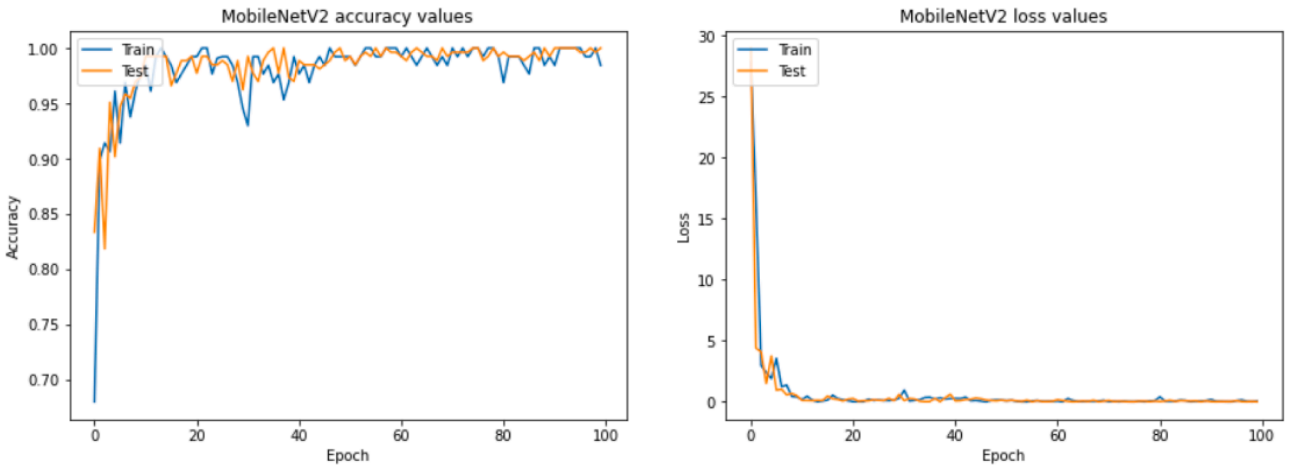


Fig. 11. Training and validation accuracy and loss curves for the MobileNetV2 model.

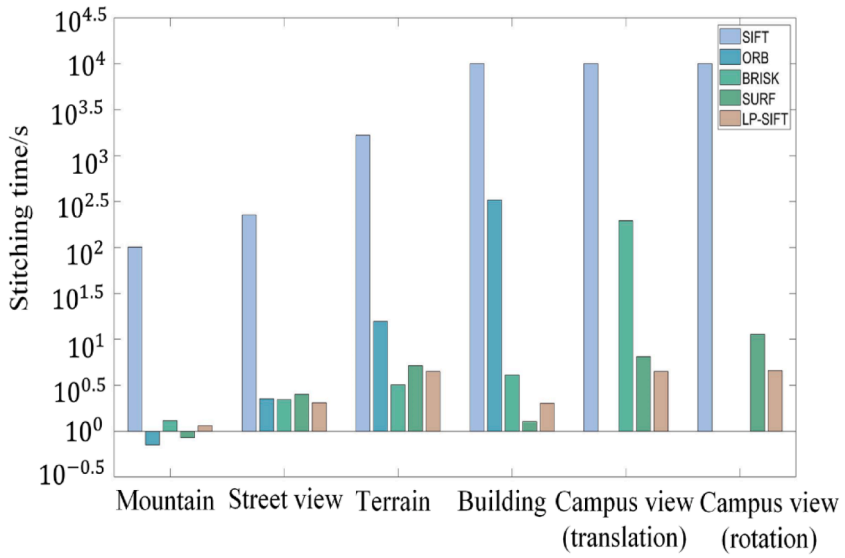


Fig. 12. Comparison of stitching times for the SIFT, ORB, BRISK, SURF, and LP-SIFT algorithms across datasets.

5. Conclusion

This section summarizes the essential contributions of this work, considers the advantages of the method, and points to directions for future work on intelligent video surveillance. The algorithm for feature point detection is discussed as follows: the local peak scale-invariant feature transform (LP-SIFT), which integrates the local maximum scale in SIFT. Using LP-SIFT with RANSAC for image stitching achieved faster processing than conventional SIFT, with applications in topographic mapping, biological and forensic analyses. This study proposes a modified memory network (MMN) for single-shot video object segmentation. The strategy consists of two modules: Importance Rate (IM) and Position Rate (PM). The IM module develops detailed memory and refines the memory embeddings by contextual evaluation of the relevance of segmentation. The five independent machine learning models that we developed to identify crime-scene investigators in video footage included the enhanced SSD MobileNet. This provides a unique method by forming ensemble models for each crime category. The OpenCV library was used to read video streams, detect objects, classify crimes, and generate frame-sharing for playback. The COA was enhanced using four strategies to represent the CMRLCCOA. In the first population generation phase, a raccoon population is generated using a sinusoidal chaotic mapping function to avoid randomness. This method augments the coati population search area and improves population quality. From surveillance to intelligent monitoring, computer vision algorithms have become important for edge detection, object recognition, and real-time tracking. Currently, LP-SIFT, along with algorithms such as Sobel filters and refined SSD MobileNetV2, guarantees superior accuracy, efficiency, and adaptability in watchkeeping, all of which are indispensable tools for security, traffic management, and public safety.

Data and material availability

This article does not involve data sharing, as no datasets were created or examined during the course of this study.

CRediT authorship contribution statement

Vivek Pandiya Raj: Writing – original draft, Methodology, Data curation. **K. Radha:** Conceptualization. **S. Harihara Gopalan:** Formal analysis. **C. Venkataramanan:** Funding acquisition. **R. Dhanapal:** Investigation. **A. Manikandan:** Validation.

Declaration of competing interest

The authors declare that they have no known coompeting financial

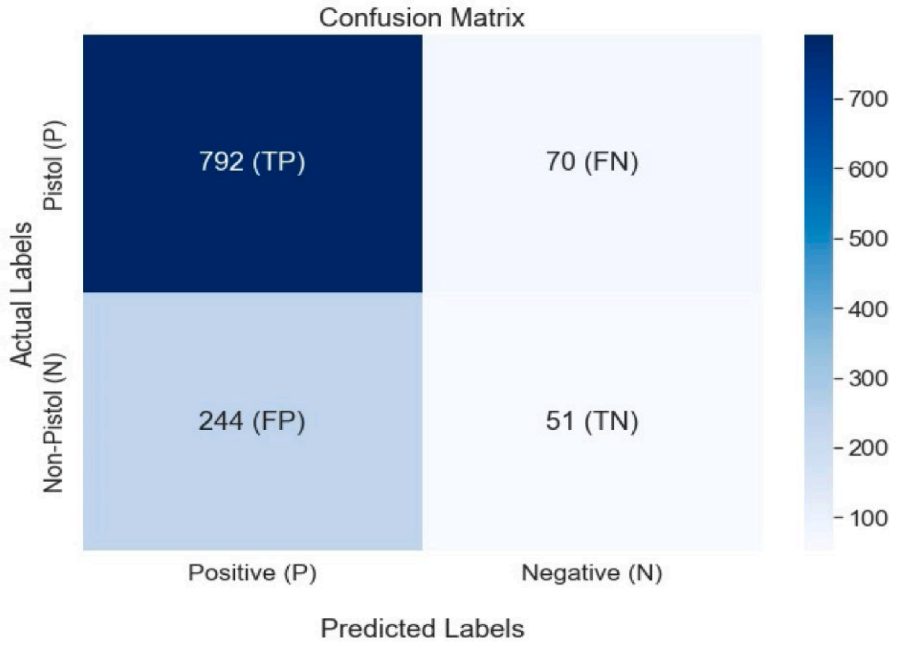


Fig. 13. Confusion matrix Hence, represents the classification performance of the pose video surveillance model.

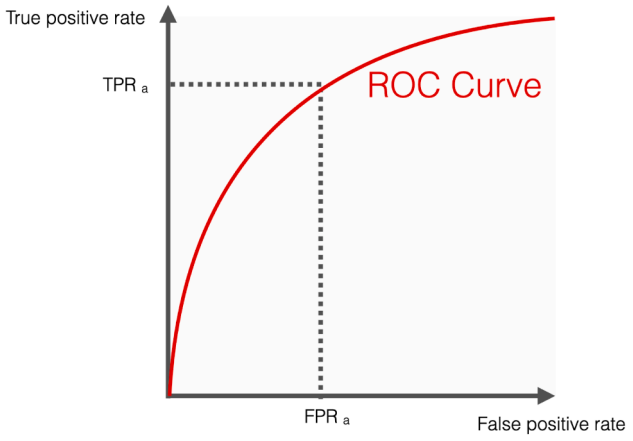


Fig. 14. Receiver operating characteristic (ROC) curve for classification performance evaluation in video surveillance.

Table 1
Results of the ablation studies.

Key	Value	J&F	J	F
×	×	84.9	82.9	86.9
✓	×	73.6	78.3	82.9
×	✓	86.3	82.5	89.3
✓	✓	82.7	84.9	89.5

intercessor personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: No Conflict of interest

Data availability

No data was used for the research described in the article.

References

[1] M. Jayamohan, S. Yuvaraj, Iv3-MGRUA: novel human action recognition feature extraction using Inception v3 and video behavior prediction using modified gated recurrent units with an attention mechanism model, Signal Image Video Process. 19 (1) (2025) 1–12.
[2] P. Kumar, S. Huang, M. Walmer, S.S. Rambhatla, A. Shrivastava, Trokens: semantic-aware relational trajectory tokens for few-shot action recognition, arXiv prepr. (2025) *arXiv:2508.03695.*.
[3] S. Cheng, J. Zhang, Y. Liu, Z. Tu, OwlSight: a robust illumination adaptation framework for dark video human-action recognition, arXiv prepr. (2025) *arXiv:2503.23266.*.
[4] F. Ziaeetabar, F. Wörgötter, Leveraging foundation models for multimodal graph-based action recognition, arXiv prepr. (2025) *arXiv:2505.15192.*.
[5] Z. Liu, S. Chang, A study of digital exhibition visual design led by digital twin and VR technology, Meas.: Sens. 31 (2024) 100970.
[6] Y. Wang, J. Yang, Origin of organic matter pore heterogeneity in oil mature triassic Chang-7 mudstones, Ordos Basin, China, Int. J. Coal. Geol. 283 (2024) 104458.
[7] S. John, Efficient spatial–temporal modeling for real-time video analysis: a unified framework for action recognition and object tracking, arXiv prepr. (2025) *arXiv:2507.22421.*.
[8] Z. He, S. Li, Y. Yu, K. Liu, A ship navigation risk online prediction model based on informer network using multi-source data, Ocean Eng. 298 (2024) 117007.
[9] T.T. Nguyen, Y. Kawanishi, V. John, T. Komamizu, I. Ide, MultiTSF: transformer-based sensor fusion for human-centric multiview and multimodal action recognition, arXiv prepr. (2025) *arXiv:2504.02279.*.
[10] S. Feng, M. Gao, X. Jin, T. Zhao, F. Yang, Fine-grained damage detection of cement concrete pavement based on UAV remote sensing image segmentation and stitching, Measurement 226 (2024) 113844.
[11] M. Zhou, Analysis of Human action recognition using Deep Learning, Appl. Comput. Eng. 160 (2025) 86–92.
[12] N. ul-Huda, H. Ahmad, A. Banjar, A.O. Alzahrani, I. Ahmad, M.S. Naeem, Image synthesis of apparel stitching defects using deep convolutional generative adversarial networks, Heliyon 10 (2024) e26466.
[13] N. Manakitsa, A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision, Technologies 12 (2) (2025) 15.
[14] G. Hu, F. Huang, A. Seyyedabbasi, G. Wei, Enhanced multistrategy bottlenose dolphin optimizer for UAV path planning, Appl. Math. Model. 130 (2024) 243–271.
[15] G. Hu, C. Gong, X. Li, Z. Xu, CGKOA: an enhanced Kepler optimization algorithm for multi-domain optimization problems, Comput. Methods Appl. Mech. Eng. 425 (2024) 116964.
[16] M. Jayamohan, S. Yuvaraj, A novel human actions recognition and classification using semantic segmentation with deep learning techniques, Neural Comput. Appl. (2025) 1–17.
[17] S. Popovych, T. Macrina, N. Kemnitz, M. Castro, B. Nehoran, Z. Jia, J.A. Bae, E. Mitchell, S. Mu, E.T. Trautman, et al., Petascale pipeline for precise alignment of images from serial section electron microscopy, Nat. Commun. 15 (2024) 289.
[18] R. Karpagalakshmi, D. Rani, N. Magendiran, A.. Manikandan, An energy-efficient bio-inspired mobility-aware cluster p-WOA algorithm for intelligent whale

optimization and Fuzzy-Logic-Based zonal clustering algorithm in FANET, Int. J. Comput. Intell. Systems 17 (2024), https://doi.org/10.1007/s44196-024-00651-0.
[19] H. Xing, K.Z. Boey, G. Cheng, Towards open-world human action segmentation using graph convolutional networks, arXiv prepr. (2025) *arXiv:2507.00756*.
[20] M. Jayamohan, S. Yuvaraj, A novel human action recognition using Grad-CAM visualization with gated recurrent units, Neural Comput. Appl. (2025) 1–16.
[21] C. Lv, Z. Lan, T. Ma, J. Chang, D. Yu, Hypersonic vehicle terminal velocity improvement considering ramjet safety boundary constraint, Aerosp. Sci. Technol. 144 (2024) 108804.
[22] S. Ma, Y. Yang, H. Yang, W. Chen, Trajectory optimization of hypersonic vehicle considering the quasi-static assumption of pitch motion, Aerosp. Sci. Technol. 146 (2024) 108969.
[23] C.P. Rosy, S. Yazhinian, M. Therasa, K.R. Surendra, A. Karuppannan, A. Manikandan, A novel channel estimation of MIMO-OFDM using hybrid bionic binary spotted hyena optimization, Int. J. Cogn. Comput. Eng. 7 (2025), https://doi.org/10.1016/j.ijcce.2025.09.003.
[24] J. Manoharan, S. Yuvaraj, A novel human action recognition model by Grad-CAM visualization with multi-level feature extraction using global average pooling with sequence modeling by bidirectional gated recurrent units, Int. J. Comput. Intell. Syst. 18 (1) (2025) 118.
[25] G. Hu, C. Gong, X. Li, Z. Xu, CGKOA: an enhanced Kepler optimization algorithm for multi-domain optimization problems, Comput. Methods Appl. Mech. Eng. 425 (2024) 116964.