Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8665891, 16 pages
https://doi.org/10.1155/2021/8665891

Research Article

Object Detection and Movement Tracking Using Tubelets and Faster RCNN Algorithm with Anchor Generation

Prabu Mohandas,[1] **Jerline Sheebha Anni**,[2] **Rajkumar Thanasekaran**,[1] **Khairunnisa Hasikin**,[3] **and Muhammad Mokhzaini Azizan**[4]

[1]*-Department of Computer Science and Engineering, National Institute of Technology Calicut, 673601, Kozhikode, India*
[2]*Department of Computer Science and Engineering, MEA Engineering College, Malappuram, –679325 Kerala, India*
[3]*Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, 50603, Lembah Pantai, Kuala Lumpur, Malaysia*
[4]*Department of Electrical and Electronics Engineering, Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia*

Correspondence should be addressed to Prabu Mohandas; prabu_pdas@yahoo.co.in

Received 9 April 2021; Accepted 20 July 2021; Published 10 August 2021

Academic Editor: Yuanpeng Zhang

Object detection in images and videos has become an important task in computer vision. It has been a challenging task due to misclassification and localization errors. The proposed approach explored the feasibility of automated detection and tracking of elephant intrusion along forest border areas. Due to an alarming increase in crop damages resulted from movements of elephant herds, combined with high risk of elephant extinction due to human activities, this paper looked into an efficient solution through elephant's tracking. The convolutional neural network with transfer learning is used as the model for object classification and feature extraction. A new tracking system using automated tubelet generation and anchor generation methods in combination with faster RCNN was developed and tested on 5,482 video sequences. Real-time video taken for analysis consisted of heavily occluded objects such as trees and animals. Tubelet generated from each video sequence with intersection over union (IoU) thresholds have been effective in tracking the elephant object movement in the forest areas. The proposed work has been compared with other state-of-the-art techniques, namely, faster RCNN, YOLO v3, and HyperNet. Experimental results on the real-time dataset show that the proposed work achieves an improved performance of 73.9% in detecting and tracking of objects, which outperformed the existing approaches.

1. Introduction

Elephants are pachyderms that live in the forest and move as groups in the search of food and water. Due to deforestation and climatic factors, elephant's movement in and around the forest areas has been increasing. These movements of elephants have led to problems such as elephants moving into human residing areas, elephants crossing roads nearby forest border areas, and crop raiding. As a result, the danger risk of human encountering the herds of elephants has become significantly dangerous, which may cause fatalities and destructions of human habitat. Therefore, there is an urgent need for a technological approach in detecting and tracking the movement of the elephant herds. This paper looked at solving this challenge and proposed a methodology for elephant movement tracking and tried to find an optimal solution in detecting and tracking movements of the elephant. There have been several measures proposed such as electric fencing, elephant proof trench, acoustic detection, and image detection methods. However, these methods have certain disadvantages in tracking the elephant movement [1]. Through the video object detection methodology, the movement of elephant herds can be observed effectively. These herds, moving in between different groups, also become a factor in choosing

the object detection method [2]. On this approach, the analysis of the movements may also produce significant findings in terms of knowing their behavior and pattern of movement.

Video object detection is a technique involving object detection using video data compared to conventional object detection using static images [3]. Application areas of video object detection methods that have greater impact are autonomous driving and video surveillance. Video object detection approaches in the earlier stages have relied on manually analyzed features [4]. With the advancement in deep learning and convolutional neural networks, deep learning methods have been more effective than conventional approaches for various tasks in computer vision, speech processing, and multimodality signal processing. Specialized algorithms have been developed that can detect, locate, and recognize objects in images and videos, some of which include RCNN, RetinaNet, and YOLO.

In the proposed work, object localization and tracking are achieved using faster RCNN along with the tubelet generation. But using faster RCNN alone has the drawback of extracting similar features from the images when RPN is trained with minibatch size, and also, the network may need lot of time in the object detection process. However, object detection performance in faster RCNN requires further improved performance due to the problems in object detection such as occlusion and deformation. The proposed work overcomes the drawback of faster RCNN, through the framework faster RCNN with the tubelet generation method.

The elephant object in the images has been taken for analysis, which comprises of different patterns of object presence. Different patterns of objects in the images have been analyzed using the faster RCNN approach with anchor generation. Existing tubelet generation methods consider bounding box detection as object proposals and generate their own tubelet to track the objects. In the object tracking process, there has been a drifting problem leading to imprecise object location. To achieve precise localization, a tubelet generation method based on object detection has been proposed. Bounding boxes based on tubelet detection results in image object detection. During the frame detection in video, objects may be missed because of blur, artifacts, group object movement, etc. Hence, object detection and tracking had been made using tubelets for achieving the precise localization of the objects.

The primary contribution of the work includes the framework for (i) combining object detection in video frames and object tracking, (ii) object region proposals generated using faster RCNN for object classification while the tubelet generation method applied for object tracking, (iii) anchor generation method used with the faster RCNN process to predict the objects and its locations, (iv) real-time datasets collected from various forest areas had been used in the analysis, and (v) the elephant in the video frames considered for object movement and tracking.

2. Related Works

Significant research work has been made in the past for the object detection discussed in this section. Object detection and localization that had been made in recent years used the still image. Video object detection methods have been effective in tracking the object localized. Hence, previous related works are analyzed for the object location and tracking.

Object detection in video has made great progress with neural networks and object detection algorithms. Still image object detections have been used effectively in refining object location in images, but it does not guarantee complete object instances in the image will be detected [5, 6]. The video object detection method has been applied in vast areas such as surveillance, transportation, and animal movement in forest areas [7, 8]. Selective search is an existing method for object detection, which generates box proposals for possible object locations by merging adjacent pixels in images [9, 10]. Object detection methods using only still images to detect objects lack accuracy because they cannot handle temporal and contextual information.

With the use of temporal in videos, object localization methods were proposed, which improves the object detection process. The object localization process is merely based on video frames similar to image object detection [11, 12]. Temporal consistency in the videos will have an impact in ensuring detection results for the video frames analyzed. Video analysis can vary significantly, such as human actions to object movement events [13]. In the existing approaches for video event, detection requires detecting and tracking objects initially, such as people, animals, and vehicles, then recognizing the actions of the objects.

Recognizing an object in the video is a developing area of research because of many fine-grained spatiotemporal variations [14, 15]. The objective of object classification is to find the object which appears in the video. In the proposed work, the problem of object localization and tracking is considered. The localized action detects changes in the spatiotemporal variations in a video.

2.1. Object Detection. Object detection is the process of detecting the bounding box which has the maximum score of detection for the given input image. Object detection in the video has been challenging due to varied image quality leading to unstable object classification in comparison to the object detection in static images. Using the tubelet generation method, the challenges in the video object detection method can be overcome by linking similar objects in the video to form tubelets [16]. The branch-bound method had been used for effective detection of bounding boxes [17]. Object detection methods such as still image detection, spatiotemporal, and contextual information in video were not explored completely [18, 19]. Hence, object detection methods combining still image and video will be effective.

Object detection performance has improved significantly with the deep neural networks. Neural network structures such as GoogLeNet, VGG, and ResNet were used to develop the learning capabilities on computer vision datasets for object detection, segmentation, and tracking [20, 21]. Neural network data such as images had been compressed during the transmission over the network and restored whenever required. It will help improve the detection accuracy [22]. Convolutional neural networks have shown improved performance in image analysis, especially in the areas of object

recognition and tracking [23, 24]. Bounding box proposals were generated from the image based on each location containing an object of interest [25]. Features extracted from each box proposal to classify it as one of the object classes. Feature extraction along with classification techniques will achieve low error rate in object detection [26]. Multiple networks had to be trained based on the different features extracted and bitrate compressions of the images taken for analysis [27]. Frameworks such as fast RCNN and faster RCNN formulate the object detection problem by training it on neural networks [28].

2.2. Object Tracking Methodologies. Object tracking is an important aspect in the process of locating the moving object in the video sequence. It is achieved by locating the target objects in consecutive frames and image pixels [29, 30]. To track the objects, object detection has to be made which has been attained using bounding box proposals in the proposed work. Machine learning approaches by extracting the features from video frames are used in tracking objects by locating the objects in the frames [31]. There have been different tracking algorithms such as Bayes, Euclidean distance, and intersection over union (IoU), for object tracking. The IoU algorithm has been efficient which involves finding the IoU between all combinations of objects of the current and previous frames. The IoU tracker can operate at thousands of frames per second, which outperforms other methods [32]. Accuracy and speed factor of the proposed work depend on the object tracker performance; hence, by using the object tracker like IoU, efficient results can be achieved.

Object detection in the video has been given increased attention due to the introduction of large datasets. Object detection in the video depends on the temporal information in the video. An efficient way to overcome this problem in the video is to analyze the temporal context of objects by linking the objects in the video to form tubelets. A tubelet used in object detection was defined as a series of bounding boxes associated with an object in image. In the proposed work, object proposals were formed by the region proposal network (RPN). The object proposals adopted in the video sequences will select the proposals between their neighboring frames with the scenario of the IoU overlap.

2.3. Challenges in Existing Approaches. Existing approaches have great success on detecting objects in static images, while detecting objects in videos remains a great challenge yet to be solved with great distinction. The challenges include factors such as drastic appearance, location change of the same object with the change in time, object occlusion, and motion blur [33]. In short, object detection approaches need to classify the object and also should be able to localize the objects in the video sequence. A previous method such as template-based action matching was used in object localization and classification [34]. Table 1 comprises of the notations used in the proposed approach.

Table 1: Notation with its description used in the proposed approach.

Symbol	Description
S_t	Tracking confidence of object
S_d	Detection score based on bounding box
B	Bounding box
O	Intersection over Union
β	Bounding boxes detected
Tb_i, Tb_j, and Tb_k	Tubelet generated from video sequences
a_1^i, a_2^i,$a_3^i \cdots a_w^i$	Visual features
Δx_i^t, Δy_i^t,Δw_i^t,Δh_i^t	Relative movement of objects
W_w and b_w	Learning parameters
b_t^i	Bounding box locations
m_1^i, $m_{2,}^i m_3^i$	Object movement
$\{\tilde{M}\}$, $\{M\}$	Normalized movements of object
$I_1, I_{2,}, I_3$	Video sequence
$O_{t,}$	Object track
D_n	Object detections
T	Represents number of frames

3. Method for Object Detection and Tracking

To achieve object localization and classification, tubelet-based object detection with the faster RCNN was proposed. The objective is to predict the high recall regions by detecting the objects in the image among the proposed regions. The detected region will be a background and objects from the given video set. Then, the model refines the localization and tracking of the object.

There is a need to detect and track objects such as elephants, due to its pattern of movement in different scenarios. Hence, the faster RCNN approach along with anchor generation has been proposed to detect the elephant's presence in different scenarios. Then, the elephant's movement had been tracked through the tubelet generation method.

The overall process of the proposed approach has been represented in Figure 1. Elephant detection and movement tracking has been made with the feature extraction and tubelet generation methods. Elephants must be initially detected to track the movement of the elephant in the video frames. Internal functions based on the feature extraction and tubelet generation methods were described in Figure 2.

The proposed approach for object detection and tracking is described in Figure 2. It includes an object classification process for classifying the objects detected in the video frames and a tubelet generation method for tracking the objects. For the given input video sequences, object proposals were generated. Based on the object proposal in the video frames, bounding boxes were determined in the object detection process. In the tracking process, objects have been classified and tubelets are generated.

3.1. Object Localization and Classification Using Faster RCNN. Object localization is to predict the object in the given video set. Similarities between the object locations were determined by the selective search approach, based on similarity criteria

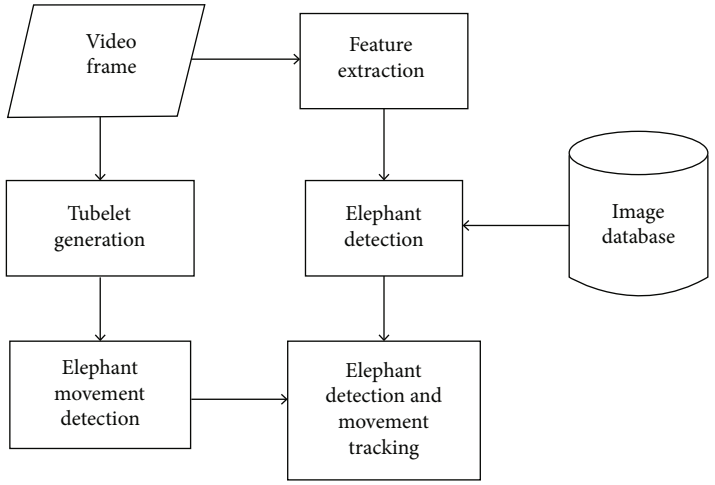

Figure 1: A proposed approach in a diagrammatic representation. After capturing the images of the elephant, feature extraction and tubelet generation methods were applied to track the elephant movement.

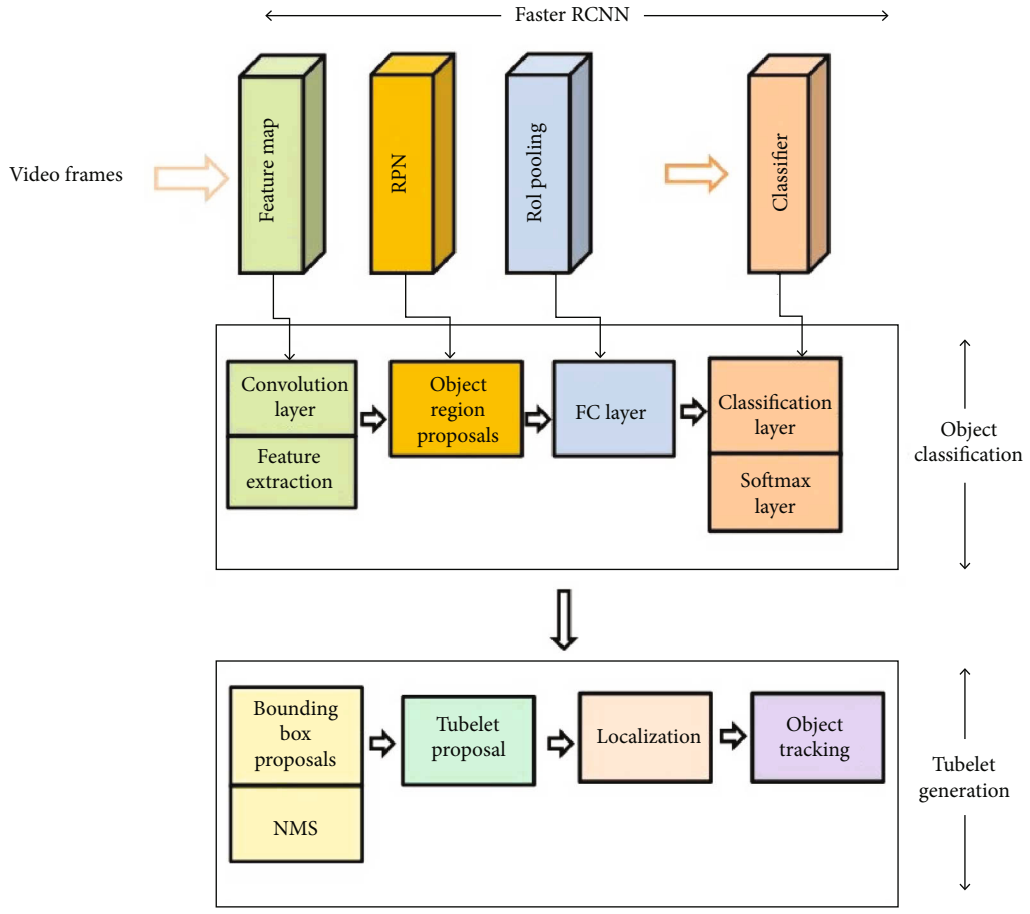

Figure 2: Block diagram for object detection and tracking. Object classification has been achieved through the layers such as the convolution layer, FC layer, and softmax layer. The tubelet generation method localizes the objects and tracks their movements.

for color, texture, and size. A selective search approach has limitations such as objects having different spatial locations within the image and varied aspect ratios. Hence, a large number of regions has to be selected which could result in a time-consuming process. To overcome these problems, the faster RCNN algorithm has been used to localize the objects and

improve the classification and detection process. The faster RCNN algorithm has significant improvement in the object detection process due to the region proposal network (RPN) which generates the object proposals [35]. RPN generates proposals for the objects in the images where the object exists.

Faster RCNN was used in the object detection task, which has two major functions such as generating region proposals and using these proposals in the network to detect the objects described in Figure 3. An input image is given to the convolutional layer which extracts the image pixels from the input image. Image pixels demonstrate the effectiveness of the image representation due to challenges such as corrupted input data [36]. Input consists of images arranged in the dimensions such as width and depth. Input holds the raw pixels of images with 3 color channels, consisting the feature map of an image [37]. Image prediction variations for the input images were based on the image pixel representations [38]. The convolution layer filters the image pixels, and the convolution operation is performed to attain a feature map. This is due to the fact that different objects were present in the images which had to be localized.

Elephant images captured in the forest areas of Theppakadu had been taken for analysis. The image containing an elephant object is passed to the convolution neural network in which faster RCNN generates object regions of interest (RoI). The next step has to pretrain CNN on image classification tasks, for defining the model. CNN in object classification takes input in the form of an image and provides the output as a category of the given images. CNN learns the feature along with the input data and uses two-dimensional convolution layer which is ideal for processing two-dimensional pictures. The region proposal network (RPN) was trained using the bounding box representation. RPN had to be fine-tuned for the regional proposal task which will be initialized by pretrain image classifiers. In the form of pixel coordinates, feature detection returns the region of interest. RoI will be a sequence of bounding boxes which is to be likely object positions. RPN is used in generating proposals, for the regions where the object is present. Feature maps are passed through a RPN for returning the object region proposals, which are classified further for object prediction and classification. Features are extracted from the images, are classified into different object classes, and return the bounding box. From the features extracted, the model was trained using the proposals generated by RPN. Then, faster RCNN had been used to initialize the RPN training in specific layers such as the convolution layer for object detection and classification.

3.2. Object Classification and Localization. The region proposal network in faster RCNN is given an input image and generates a set of object proposals for the corresponding feature map. The feature vector generated from the object proposals was fed into the output branches for object classification. In the last layer, object classification and localization were achieved. In the object detection task, each of the proposals will be of different shapes. Object proposals are detected to be different shapes based on the region of interest. Region of interest pooling converts the image proposals into a fixed shape. Fixed sized feature maps are produced from nonuniform inputs by max pooling on the inputs.

For a given image, RoI pooling of each RoI depends on different parameters. It takes two inputs such as a feature map obtained from a convolutional network with convolution and max pooling layers and a matrix representing the regions of interest. The first column in the matrix denotes the image index, and the remaining column represents the coordinates of the object region. A fully connected (FC) layer has a softmax layer and a linear regression in which region proposals were passed for classifying and bounding box proposals for objects.

For the given input, based on the region of interest, a section of the input feature map is taken and scaled it to a fixed size. The scaling is done on the basis of dividing the region proposal into fixed size sections and finding the largest value of each section. Object localization in images was made using the similarity grouping of the nearby pixels. Similarities between the nearby pixels are acquired and merged with them. By repeated merging of the image pixels, the object location in an image was obtained.

The input image consists of an elephant as an object which was passed into a convolution layer to obtain the feature map as described in Figure 4. Then, the image filtered was passed through the RPN to obtain the localization of the objects which consists of elephant objects localized in the given input image. Object region proposals are of varied shapes and it has been normalized by the RoI pooling function. Hence, the objects will be of the same size located in an image. Then, the objects localized in had been classified. Based on the input image, the object has been localized and classified as elephant. The input image is of $32 \times 32 \times 3$ describing the resolution and size of the image. Considering the given image in the JPG form with the dimension of 320×320, the representative array is of $320 \times 320 \times 3$. Numbers describe the pixel intensity which is of the value 0 to 255.

For the given input image, the feature map produced by applying the filter over the object locations of the image and an array of $28 \times 28 \times 1$ was obtained as a feature map. 784 different object locations were obtained which can fit on to a 32×32 input image. 784 object locations are mapped to a 28×28 feature map. Filters perform a feature identifier function which includes things like edges and curves.

CNN has three main types of layers such as the convolutional layer, pooling layer, and fully connected layer. The convolution layer consists of filters which are small spatially that extends to the given input image. For the given input image, a 2-dimensional activation map that gives the responses of the filter for each position obtained, which means the network will learn the feature for the given input. There will be a number of filters in the convolution layer and an activation map will be produced by each of them. The output of the convolution layer will be combining the activation maps.

The pooling layer is to reduce the spatial dimension of the representation. It reduces the computation in the network and also controls overfitting. The pooling layer operates independently on the given input images and, by using the MAX operation, resizes the given input images. In the fully

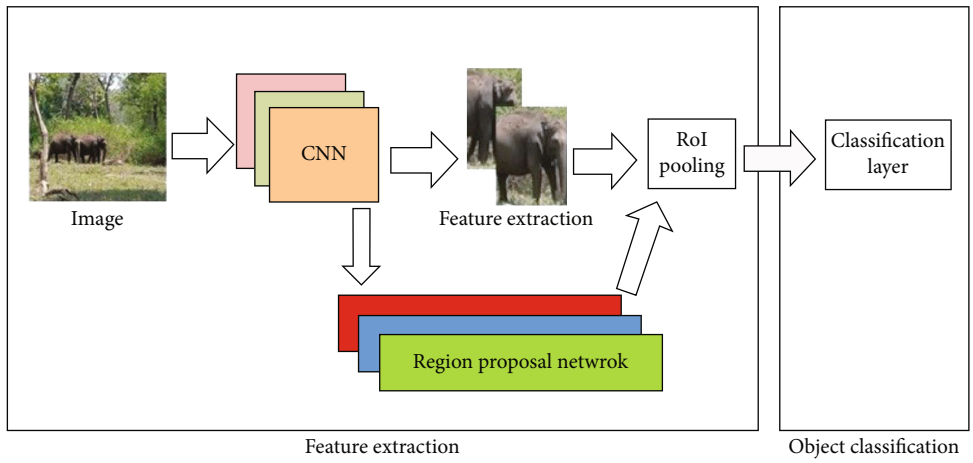

Figure 3: Faster RCNN for object detection. In the input images, CNN was used to extract the features, forming region of interest (RoI) pooling to classify the objects.

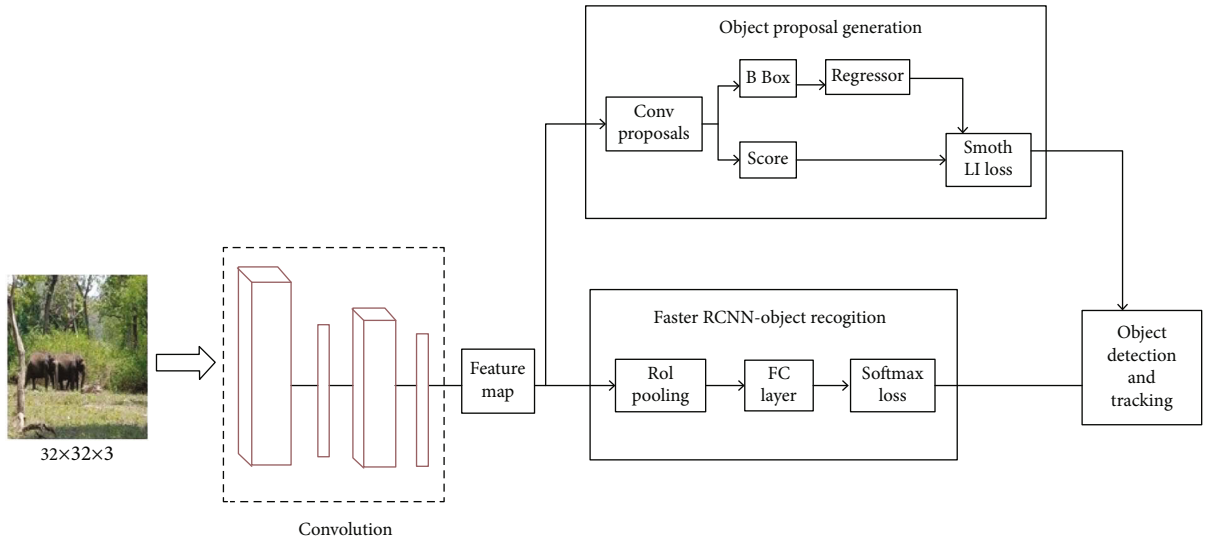

Figure 4: Proposed approach for object detection and recognition. Bounding box representation applied in object proposal generation for the elephant images acquired in the Theppakadu forest area. Objects were recognized using the faster RCNN approach.

connected layer, neurons have full connections to all activations in the previous layer.

The softmax loss layer determines the logistic loss of the softmax for the given input images. The function of the layer is to compute the logistic loss operation for different layers. The softmax layer has two inputs such as the predicted label and a fully connected layer. The smooth loss function does the classification performed with deep neural networks trained with the cross-entropy loss. Results suggest that cross-entropy is the learning objective for infinite data.

RPN generate the proposal for the objects. These object proposals are called bounding boxes. Bounding boxes were to determine the object locations in the domain by considering the ground truth box and IoU. RoI pooling in the object detection network resolves the problem of fixed size image requirement. The fully convoluted layer produces an output of N dimensional object locations that has to be detected from the given input. Each of the objects represents the probability of an object location. Thus, the objects detected from the input image including trees and elephants are classified using the softmax layer.

3.3. Object Generation Module. In faster RCNN, RPN is used instead of the selective search module, which greatly improves the time and accuracy of object detection. Algorithms can also directly extract features to predict the object and its location. Methods such as OverFeat, YOLO, and SSD directly predict objects and return its location.

The proposed approach uses RPN to generate anchors used in direct classification and regress anchors. The number and shape of anchors will affect the object detection performance. Anchors are used in object detection for generating

object region proposals. More anchors are generated, leading to decreased detection accuracy, since most of the area surrounded by the anchor frame is the background. Hence, the anchor generation strategy can effectively reduce the number of anchors and will have great importance in optimizing the object detection performance.

In faster RCNN, the input image is passed through the convolution layer for anchor generation and feature extraction. Instead of RPN in the region, proposal generation and anchor generation have been proposed for determining the object presence. Object proposals generated from the anchor generation and features extracted from the feature map will be combined through RoI pooling to classify the objects. As described in Figure 5, initially, the anchor position prediction was made to generate a probability distribution in the feature map. It also indicates the possible location of the object in the image. In a similar manner, the object prediction involves prediction of the presence of objects in the input image. By combining the results of object and location prediction, the generated predicted objects in the images will be classified.

A probability distribution map generated by the anchor position prediction will be similar to the feature map of the input image. The proposed method is involved in determining the ground truth box prediction for the feature map in the training process and the threshold set for the remaining areas for determining the disregard area. The determination of the object prediction is to predict the length and width for the given anchor center point.

Anchor box characteristics are to capture the scale and aspect ratio of object classes required to be detected and chosen based on object sizes in the datasets. In the object detection, the anchor boxes were predefined and arranged across the image. Then, the network predicts the probability of object presence, background, and IoU. Predictions were used to determine each anchor box. In the proposed method, faster RCNN used to generate anchors with suitable size. The parameters (x, y) are used to describe an anchor, where (x, y) signifies the coordinates of the anchor. The shape information of the anchor had been integrated into feature map through which object detected can adapt to the anchor box parameters. As defined in Figure 5, a 3×3 deformable convolution to the feature map, the offset of the convolution had been obtained through a 1×1 convolution.

3.4. Object Detection with Tubelet Generation. Object detection with tubelet generation was made for the localization of the tubelets. Bounding box representation of the objects in the previous and current frame was compared to detect the presence of the same object in the two frames. IoU was used to predict the object location between different frames by using the threshold greater than 0.5, where the bounding box has to be the same. Lower detection scores may result during the bounding boxes tracking when the objects have larger overlap.

3.4.1. Detection Scores. Detection scores were calculated based on the bounding box coordinates and let the coordinates be (x_1, y_1) and (x_2, y_2),

$$S_t = \frac{1}{1 + e^{-0}}, \tag{1}$$

where S_t denotes the tracking confidence and with S_d detection score; object tracking has been achieved. O is the resultant of the average output for the given image.

Real-time elephant movement images as shown in Figure 6 are taken for a study from Hosur areas. Bounding representation of the object with its left, top, and bottom right coordinate representation had been described in Figure 6.

$$S_t' = \left\{ \begin{array}{l} (S_t + S_d) * \dfrac{1}{8} \\ (S_t + S_d) * \dfrac{2}{8} \\ (S_t + S_d) * \dfrac{3}{8} \end{array} \right\}, \tag{2}$$

where $S_d \epsilon [0, 0.25], [0.25, 0.5], [0.5, 0.75]$ and S_t is obtained by using the sigmoid function on tracking for ensuring the object presence based on the true objects detected as bounding boxes in expression (2). The values of the bounding boxes as stated in expression (2) are considered for analyzing the tracking of bounding boxes. In general, bounding boxes were considered true objects by the object tracker without analyzing the detection scores. The lower the detection score, the less likely to be the objects presence; hence, the tracking confidence and detection scores were calculated. There may be different bounding boxes which may have overlap within the frame. IoU was calculated, and if their value is higher than 0.5, then it is merged, since the detected boxes will be of similar objects.

In general, tubelets tend to overlap with each other due to multiple object detection as described in Figure 7 of an elephant image captured from Hosur forest areas. To overcome these types of problems, tubelets which are overlapping satisfying the above condition of value higher than 0.5 were merged. In the process of object tracking, object proposal suppression was performed to minimize the redundant tubelets. The objective of tubelet box distressing and max pooling process is to generate new tubelet boxes by replacing the existing tubelets on each frame randomly.

Figure 8 comprises of the object movement proposals for object classification in images with tubelets. Figure 8(a) is the input image comprised of an object classified, while Figure 8(b) defines the tracking of objects in an image. In the objects detected, tubelets had been replaced with those that have overlaps based on the threshold to perform the conventional NMS process. This process will bring back the positive boxes, if the tubelet detected has been with a lower detection score of positive boxes.

Object tracking and localization have been achieved using the bounding box sequence generated for the given video frames. A bounding box overlap for multiple objects had been determined using IoU overlap. Based on the IoU threshold, object movement has been tracked as described in Algorithm 1.

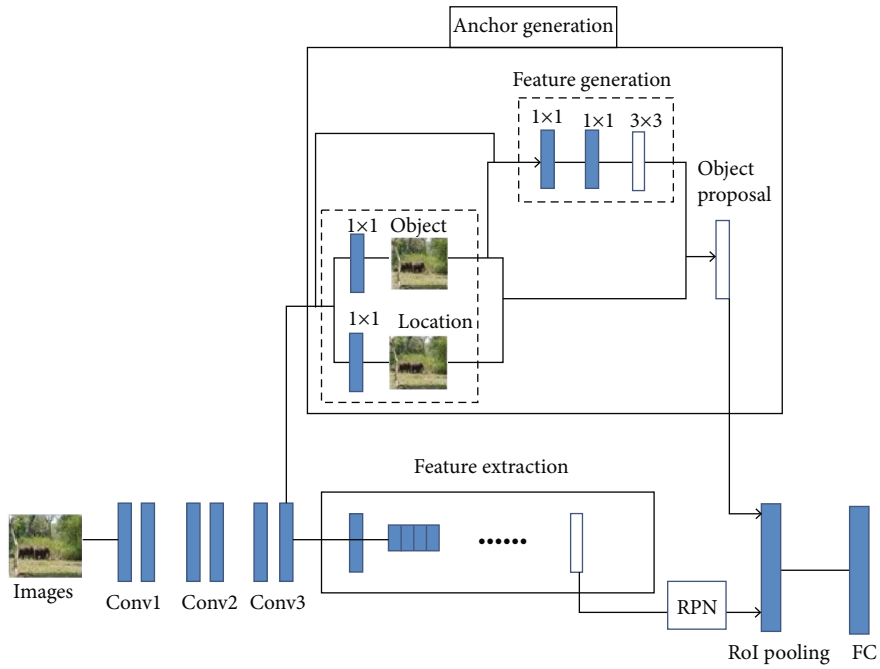

Figure 5: Object generation module using anchor generation to predict objects and location. Elephant images acquired in the forest area of Theppakadu were passed through convolution layers to classify the objects by anchor generation and feature extraction.

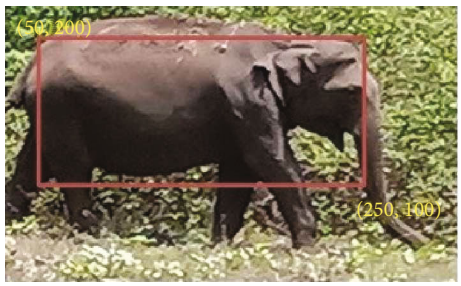

Figure 6: Bounding box representation for the input image. Bounding boxes of the object in the image with the coordinates were represented.

Figure 7: Intersection over union between objects. Bounding boxes were merged based on IoU values considering it as a similar object.

Tubelets were generated from the bounding box sequence for the given object instances as described in Figure 9. Tb_i, Tb_j, and Tb_k were the different tubelets generated for the input video sequence.

3.5. Object Tracking Based on Tubelet Generation. There had been drifting problems in the object detection, due to bounding boxes overlapping in the detections. To overcome the drifting problems and precise object localization, the tubelet generation method was proposed. Object tracking was made based on the features of moving objects, by achieving prediction and tracking of moving objects. In the object tracking based on classification, a number of candidate regions will be extracted. These candidate regions will be sent to the network for classification and result in considerable time computation overhead. Object detections were represented in the form of bounding boxes. Tubelets are generated by applying the tracking algorithm to the static image bounding box proposals.

The proposed approach computes the object locations and classification scores for each frame. The classification score of the tubelet has been computed by combining the classification scores of the objects.

Tubelet proposals detect the objects in the images and can accurately track objects. a_1^i, a_2^i,$a_3^i \cdots ..a_w^i$ are visual features based on the box locations b_t^i. In order to track the movement of the object, a network is trained that effectively evaluates the spatial actions with respect to spatial features.

$$m_1^i, m_2^i, m_3^i \cdots m_w^i = \mathrm{r}\left(a_1^i, a_2^i, a_3^i \cdots ..a_w^i\right). \quad (3)$$

Expression (3) was used to evaluate the movement of the

(a) (b)

Figure 8: Detection of elephant object movement in the Theppakadu forest area with the bounding box proposal. (a) Object classification using tubelet generation. (b) Object tracking approach using tubelet for moving objects.

Step 1: input: $B = (BB_1, BB_2, BB_3 \cdots \cdots \cdots \cdots ..BB_n)$, S_d, O, δ; B—bounding boxes, S_d—detection scores of bounding boxes, O—intersection over union, δ—NMS, BB_n—bounding box numbers.
Step 2: output: β—bounding boxes detected
Step 3: $\beta \longrightarrow \varnothing$
Step 4: while $B \neq \varnothing$
Step 5: do
Step 6: arg max$U(bb_j) \longrightarrow bb_m$
Step 7: $S_d(bb_m) \longrightarrow S$
Step 8: if IoU$(bb_m, bb_j) > \delta$ then
Step 9: $S \longrightarrow \max(S, S_d(bb_j))$
Step 10: end if
Step 11: $\beta \longrightarrow \beta \cup (bb_m, S)$
Step 12: end while
Step13: return β

Algorithm 1: Object tracking and localization using intersection over union (IoU).

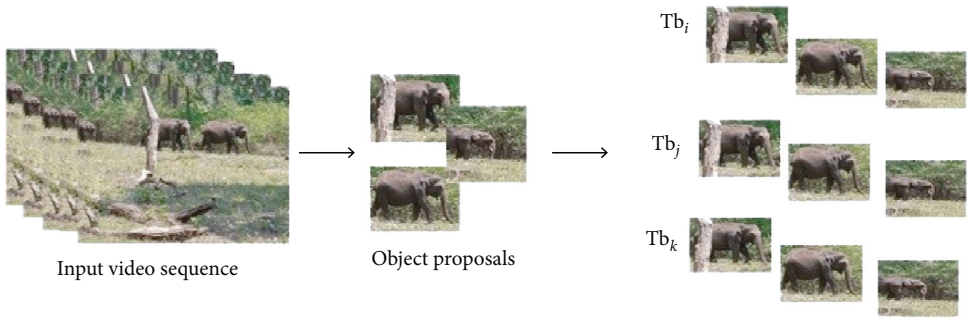

Figure 9: Tubelet generation from object proposals of the elephant movement images captured from Theppakadu forest area. Object proposals were generated by bounding boxes. Sequence of bounding boxes form the tubelets.

objects with the use of visual features. Relative movements are $m_t^i(\Delta x_i^t, \Delta y_i^t, \Delta w_i^t, \Delta h_i^t)$ which has been calculated as follows:

$$\Delta x_t^i = \frac{x_t^i - x_1^t}{w_1^i}, \tag{4}$$

$$\Delta y_t^i = \frac{y_t^i - y_1^t}{h_1^i}, \tag{5}$$

$$\Delta w_t^i = \log\left(\frac{w_t^i}{w_1^i}\right), \tag{6}$$

$$\Delta h_t^i = \log\left(\frac{h_t^i}{h_1^i}\right). \tag{7}$$

By using (4)–(7), the relative movements of objects in the images can be inferred. The input taken as the visual features includes $(a_1^i, a_2^i, a_3^i \cdots ..a_w^i)^t$ and outputs the movement features of objects as 4 W expressed by the tubelet proposal as

$$\left(m_1^i, m_{2,}^i m_3^i \cdots \cdots \cdots m_w^i\right)^t = W_w\left(a_1^i, a_2^i, a_3^i \cdots ..a_w^i\right)^t + b_w, \tag{8}$$

where W_w and b_w are the learning parameters of the

Step 1: input: video sequence $(I_1, I_2, I_3 \cdots \cdots \cdots \cdots ..I_n)$
Step 2: data: O_t—object tracks, $O_t = \{O_1, O_2, \ldots\ldots O_n\}$, $D_n = \{D_1, D_2, \cdots \cdots D_n\}$ where
D_n—object detections, T—represents number of frames detected in the tracking and frames not detected in the tracking
Step 3: output: object track, O
Step 4: **Function** tracking
Step 5: **while**max $A_{ij} > \propto // A_{ij}$—object tracks with intersection over union values
Step 6: **do**
Step 7: **if**$B_{ij} \cup A_{ij} > \propto // B_{ij}$—number of frames in the video sequence without any object detections
Step 8: arg max $A_{ij} \longrightarrow (i, j)$
Step 9: $D_j \longrightarrow O_i$
Step 10: $B_j \longrightarrow True$
Step 11: $O_t \longrightarrow O_{t-1} \cup \{B_{ij} = \text{True} \forall D\}$//object tracking for the detected objects
Step 12: **return**O_t for the given set of video sequences

Algorithm 2: Object tracking algorithm for the given video sequence.

Table 2: Dataset overview.

	Training	Validation	Testing
Video sequences	2758	684	2040
Number of frames	6329	4831	—
Positive snippets	1344	863	—

Table 3: Detection results for elephant object in video sequences.

Range	Detection results
(64, 128)	75.9%
(128, 256)	88.1%
(256, 512)	93.2%

Table 4: Object detection results for the given dataset.

Methods	Detection rate
TCNN	78.8%
Background subtraction	81.6%
Proposed approach	85.7%

Table 5: IoU overlap thresholds for different methods.

Method	0.25	0.50	0.75
Tubelet detection method	0.56	0.49	0.41
Static image video dataset	0.52	0.47	0.39
Proposed approach	0.71	0.63	0.54

layer. Tubelet boxes are generated by the regression layer that has similar movement patterns with the ground truth. The relative movement targets $m_1^i = (x_i^t, y_i^t, w_i^t, h_i^t)$ can be defined on the basis of ground truth boxes at time b_t. Movement patterns of the objects with respect to the bounding boxes are represented through

$$L\left(\{\tilde{M}\}, \{M\}\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{w}\sum_{k\epsilon(x,y,w,h)} d\left(\Delta k_t^i\right), \tag{9}$$

where $\{\tilde{M}\}, \{M\}$ are the normalized movements for the object movement detection outputs. Object tracking based on tubelet generation primarily depends on the detection scores of a frame. Tracking of objects in images were based on the confidence of the object proposals. Anchors are called starting detections of tracking, based on the object proposals. There may be a drift in the objects detected during the object tracking moves away from the anchors. Hence, false tracking can be reduced by stopping the early object tracking while the confidence is below the threshold.

Algorithm 2 describes the process of object tracking in the video sequence. Object tracking represent the tubelet of the objects detected in each frame with the bounding box proposals. In the given video sequence, a set of tracking has been detected and based on the intersection over union measure, and tracking objects were identified. By detecting the objects in the frames, object tracking has been achieved.

4. Datasets

In the proposed approach, a new datasets consisting of real-time video recordings has been used for analysis. A dataset consisting of video sequences is captured during the field visits made at different seasons. During the field visits, elephant movements in different seasons and patterns were observed and recorded manually for future reference. Videos were captured in the format of 1920×1080 pixels at 50 frames per second. The complete dataset comprises of 70 GB of video files, which approximately corresponds to 8.5 lakh frames. To avoid computational complexity in video processing, the video sequence taken for analysis is between 3 and 5 seconds, which includes major factors that may be available in the larger video sequences. For the evaluation of the proposed method, a subset of the video collection was taken for the study. Datasets taken were real time consisting of multiple objects. Images comprising multiple moving object categories were chosen, such as elephant, vehicles,

Table 6: Elephant object localization and tracking for the given dataset.

Method	Object localized
Tubelet detection method	66.3%
Tracking and detection based tubelet (proposed approach)	73.9%
Static image video dataset	71.8%

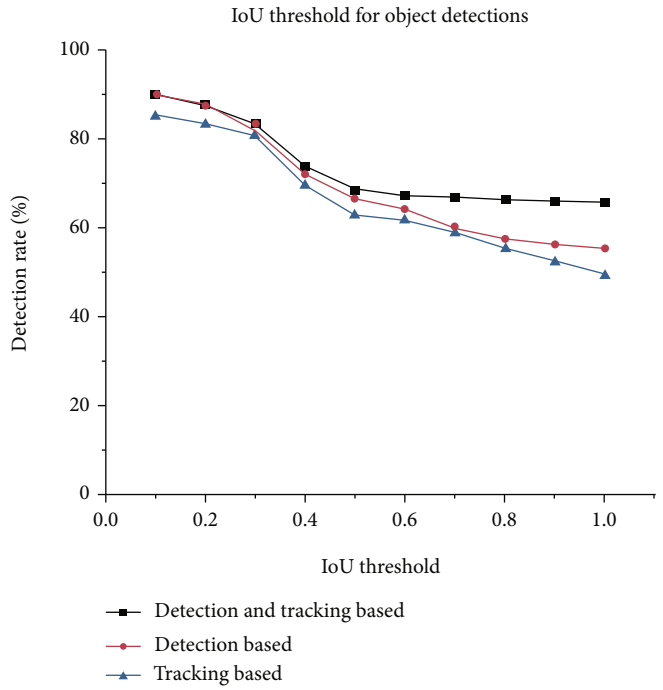

Figure 10: Object detection based on the IoU threshold. Performance comparisons of object detection and tracking methods based on the IoU were described in the figure.

and humans. In each input video, the numbers of frames are different.

Table 2 comprises of real-time datasets used in the process. The dataset is split into a training set and a validation set, containing 2758 video snippets and 684 video snippets, respectively. This mainly includes the category information of all objects which are identifiable and location information. In addition, the proposed object detection model performance was verified on the standard test set that contains 2000 video sequences and calculated the evaluation indicators in the test set.

4.1. Evaluation Metrics. Evaluation metrics were based on the average of precision on different aspects of object detection. The mean average precision was computed based on the score of a tubelet. The tubelet score was based on the average score of detection in the object detection process. Faster RCNN detectors were trained on different real-time datasets and network structures. Objects will be localized exactly based on the ground truth tubelet from a given class of images. An anchor has been considered a positive sample if it satisfies the constraints such as an anchor having the highest IoU overlap measure and an anchor having IoU greater

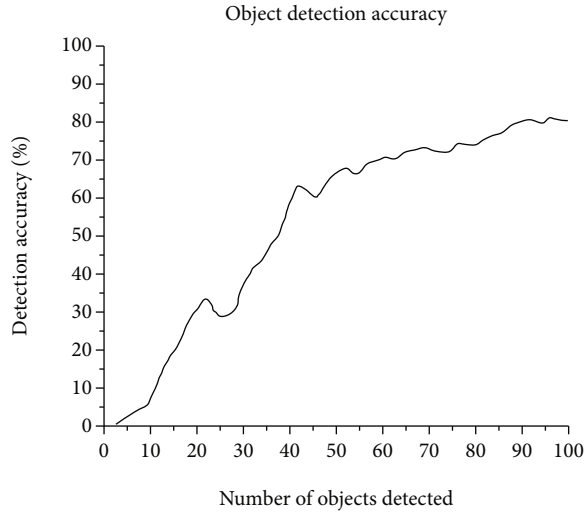

Figure 11: Object detection accuracy based on the number of objects. With the increase in the number of objects presence, the proposed method achieves a high detection rate.

Table 7: Performance comparison of the proposed approach with the existing approaches.

Methods	No. of frames	Specificity (%)	Precision (%)	Recall (%)	Accuracy (%)	TP rate (%)	FP rate (%)	Error detection rate (%)
Faster RCNN	572	0.7384	0.8513	0.8411	0.9235	0.9185	0.8331	0.9076
Kai et al. [25]	463	0.7572	0.8358	0.8824	0.9088	0.8938	0.8578	0.9012
Mihir et al. [9]	546	0.7264	0.8488	0.8517	0.9124	0.8754	0.8841	0.0876
Proposed approach	575	0.7841	0.9133	0.9357	0.9788	0.9341	0.9544	0.0212

Table 8: Object detection performance of the proposed approach.

Methods	Vehicles (%)	Monkey (%)	Deer (%)	Elephant (%)	Bear (%)	Humans (%)	Tiger (%)	Dog (%)
Faster RCNN	0.7813	0.8518	0.8551	0.8419	0.7915	0.8845	0.8688	0.8314
YOLO v3	0.8051	0.8493	0.8918	0.8588	0.8344	0.9145	0.8755	0.8489
HyperNet	0.7988	0.8369	0.8754	0.8635	0.7889	0.8801	0.8655	0.8581
Proposed approach	0.9311	0.8950	0.8845	0.9798	0.9318	0.9841	0.8845	0.8765

than 0.5 of the ground truth boxes. Anchors having IoU threshold less than 0.5 were considered a negative sample, and it have been ignored.

5. Results and Discussion

In the proposed approach, RPN anchors contained three scale (128, 256, and 512) models which were trained. The detection results of this proposed method reach 85.7%. The detection rate has been improved by exploring the model ensemble and computing the detection in different object ranges. Missing objects had been recalled by comparing it with the detection results which has been improved by the model ensemble as found in Table 3.

Based on the detection results and bounding boxes, tubelets are generated for tracking the objects. Object localization depends on the IoU threshold for the bounding box matching. IoU thresholds have multiple values for the bounding box matching such as 0.25, 0.5, and 0.75. The final tubelet generated will be based on an average of the tubelet mAP with different tubelet IoU thresholds. The tubelet detection was based on the actual tubelet localized to the given input video.

$$\mathrm{IoU}(T, T_{at}) = \frac{T \cap T_{at}}{T \cup T_{at}}. \tag{10}$$

IoU thresholds were determined using expression (10) for object localization.

From Table 4, the object detection performance has been compared and the results show that the proposed approach has achieved higher performance. Detection comparisons have been made on the real-time dataset chosen for analysis. By using detection-based tubelet and tracking-based tubelet, object movement has been tracked in the videos.

5.1. Threshold. IoU is an evaluation metric used for describing the object detection model in the datasets. IoU evaluates the bounding boxes predicted by the proposed model. IoU overlap thresholds for the different methods are presented in Table 5. These thresholds were determined based on the experimental results of object detection during overlapping. Based on the overlapping threshold, the object detection process for the given image varies. Threshold values such as 0.25, 0.50, and 0.75 were taken for considering the object detection variations on the images taken for analysis. With the increasing overlapping thresholds, the elephant object detection rate decreases. Increasing the overlapping thresholds of objects had resulted in varying detection results.

IoU overlap was set to the nonmaximum suppression (NMS) threshold to 0.3. Tubelets generated based on the bounding boxes in the video frames and their locations are predicted in the next frame using the optical flow value of bounding boxes. IoU overlaps were computed for each bounding box in the next frame. If the IoU overlap is of maximum value and above the threshold, then it belongs to the same object, else the bounding boxes represents a new object.

Table 6 compares the results of tracking and detection based on tubelet, to the other approaches for tracking and localization of objects in a video dataset. Based on the comparison, it is clear to define that the proposed approach outperformed the static image video dataset and tubelet detection method.

IoU thresholds for the object detection on the different scenarios were presented in Figure 10. Varying object detection rates were based on the bounding boxes for different kinds of tubelets. Detection- and tracking-based approaches were the proposed work, which had been effective in comparison with other approaches.

In Figure 11, the object detection performance based on the number of objects detected had been shown. When the number of objects to be detected is minimum, the detection rate has been higher. The minimum number of objects to be detected varies, since real-time scenario was given for the detection process. When the number of objects to be detected is less than 20, detection accuracy has been achieved above 80%.

Figure 12: Elephant movement tracking on the images and videos captured from the Hosur forest area and Theppakadu elephant camp by the proposed approach using tubelets. Elephants were detected in the first column of the images, and then, its movements were localized and tracked.

Tables 7 and 8 describe the performance of the object detection accuracy. The proposed approach performance has been described by comparing it with the existing approaches. Different objects detected using the real-time datasets have been tabulated below. From the results in Tables 7 and 8, it is conclusive that the proposed approach has been effective in object detection.

In Figure 12, an example of tubelet generation for elephant objects at different scenarios has been presented. Bounding boxes represent the object instance, and the similar object presence in the consecutive frames has been denoted by the same color. Hence, through the proposed approach, object instance detection and tracking have been achieved.

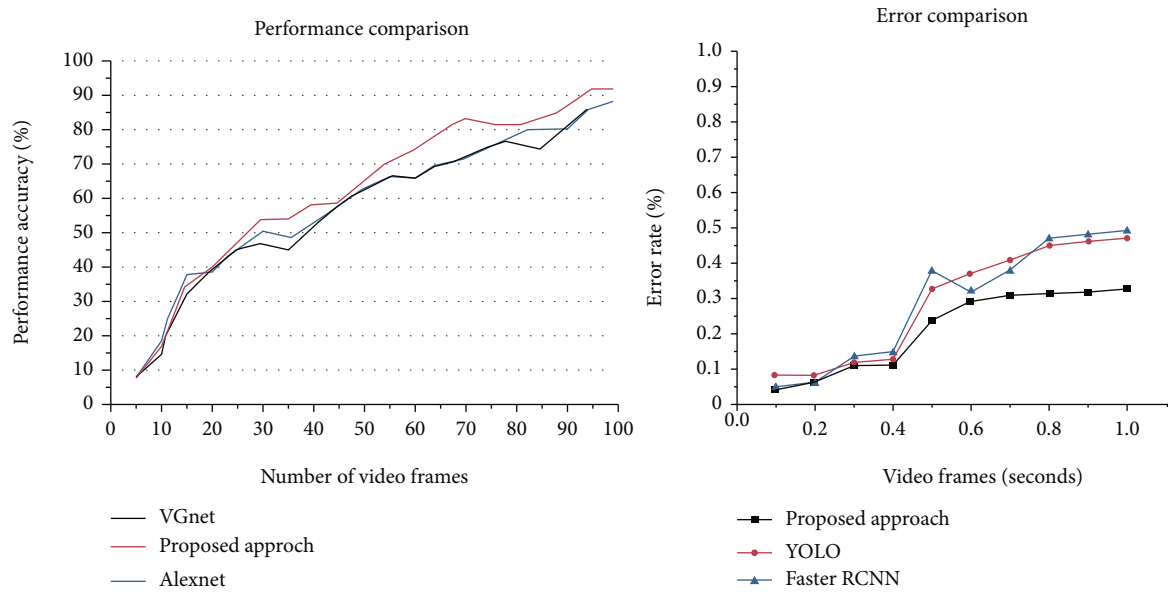

Figure 13: Performance and error rate comparison of the proposed approach. (a) Performance accuracy of the proposed approach compared with the existing approaches. (b) Error rate of the proposed approach compared with the existing approaches.

Figures 13(a) and 13(b) describe the performance and error rate comparisons of the proposed approach with the existing approaches. Based on the above comparison, it is clear that the proposed approach has been effective in object detection. The performance comparison of the proposed approach with other methods found data duplication due to the image pixel overlapping leading to misclassifications. Comparatively, other methods have been time consuming and require more memory space.

Faster RCNN has RPN to generate bounding boxes called region of interests (ROIs) which has high probability of containing objects. Hence, the number of bounding boxes generated indicates the presence of objects. The IoU threshold value was used in determining the object presence even though overlapping of bounding boxes exists which had been efficient in comparison to image pixel overlapping. In the proposed method, the object detection process had determined using a bounding box with a IoU threshold value. Then, the detected objects had been tracked through the tubelet generation method.

In the error comparison, localization errors had been found in the YOLO detections, and faster RCNN had a mean average precision (mAP) error variations in the object detections. The error rate in the object detections had been lower in the proposed approach while comparing to the existing approaches as described in Figure 13(b). Hence, the proposed approach will be an effective in the detection and tracking of elephants.

6. Conclusion

In this paper, the video object detection analysis is made through the machine learning approach. In the proposed approach, through faster RCNN and tubelet generation method, different objects in real time have been detected and classified. The proposed approach has achieved 73.9% of detection and tracking of elephant objects which differs based on the image scale. Elephant objects have been tracked and classified using the IoU overlap of the anchor, where the different objects occlude. Using the detection results, elephant objects in the video had been localized and its movements had been tracked. The proposed approach has analyzed the elephant as an object for classification, and future work needs to be investigated on detection accuracy for multiple object detection. Furthermore, future study has to be made on object detection and tracking based on the object size variations.

Data Availability

Data is available on authors request. Elephant images and videos were acquired from Hosur forest and Theppakadu elephant camp of Tamilnadu forest department.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The work was supported and sponsored by the Department of Science and Technology, Science Engineering and Research Board (DST-SERB). The authors would like to express their gratitude to the forest officials and villagers from Hosur forest area and Theppakadu elephant camp for their help and support during the data collection procedure. The images of the elephant used in the paper are captured from the Hosur forest areas and from Theppakadu elephant camp. Images were captured and used with the consent from the authorities of Tamil Nadu forest department.

References

[1] A. Gayathri, A. Sulaiman, D. Kumar, S. Phalke, and A. Krishnan, "Status of elephant proof barrier mechanisms in Bannerghatta National Park," *Technical report*, A Rocha India, Bengaluru, 2016.

[2] S. J. Sugumar and R. Jayaparvathy, "An improved real time image detection system for elephant intrusion along the forest border areas," *The Scientific World Journal*, vol. 2014, Article ID 393958, 10 pages, 2014.

[3] H. Zhu, H. Wei, B. Li, X. Yuan, and N. Kehtarnavaz, "A review of video object detection: datasets, metrics and methods," *Applied Sciences*, vol. 10, no. 21, p. 7834, 2020.

[4] H. T. Nguyen, M. Worring, and A. Dev, "Detection of moving objects in video using a robust motion similarity measure," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 137–141, 2000.

[5] A. Arora, A. Grover, R. Chugh, and S. S. Reka, "Real time multi object detection for blind using single shot multibox detector," *Wireless Personal Communications*, vol. 107, pp. 1–11, 2019.

[6] V. Rodrigo and R. D. S. Javier, "Object detection: current and future directions," *Frontiers in Robotics and AI*, vol. 2, pp. 1–7, 2015.

[7] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Computer Vision and Pattern Recognition*, pp. 3457–3464, IEEE, 2011.

[8] C. Rd, J. Fernando, and G. Narciso, "An efficient multiple object detection and tracking framework for automatic counting and video surveillance applications," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 857–862, 2012.

[9] J. Mihir, V. G. Jan, J. Herve, B. Patrick, and G. M. S. Cees, "Tubelets: unsupervised action proposals from spatiotemporal super-voxels," *International Journal of Computer Vision*, vol. 124, pp. 287–311, 2017.

[10] J. R. R. Uijlings, K. E. A. Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[11] O. B. Eshed and M. T. Mohan, "Multi-scale volumes for deep object detection and localization," *Pattern Recognition*, vol. 61, pp. 557–572, 2016.

[12] M. Rochan, S. Rahman, N. D. Bruce, and Y. Wang, "Segmenting objects in weakly labeled videos," in *2014 Canadian Conference on Computer and Robot Vision*, pp. 119–126, 2014.

[13] D. Jerline Sheebha Anni and S. Arun Kumar, "A wireless sensor network based on unmanned boundary sensing technique for minimizing human elephant conflicts," *Studies in Informatics and Control*, vol. 26, pp. 459–468, 2017.

[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (China, 2005)*, pp. 65–72, 2005.

[15] J. Mao and L. Yu, "Convolutional neural network based bi-prediction utilizing spatial and temporal information in video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1856–1870, 2020.

[16] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, and J. Wang, "Object detection in videos by high quality object linking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1272–1278, 2020.

[17] S. I. Mostafa, A. B. Amr, R. A. Mostafa, and F. E. Ibrahim, "Bounding box object localization based on image superpixelization," in *Proceedings of the International Neural Network Society Winter Conference (INNS-WC, 2012)*, pp. 108–119, 2012.

[18] I. Everts, J. Gemert, and T. Gevers, "Evaluation of color spatio-temporal interest points for human action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1569–1580, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[20] L. Li, O. Wanli, W. Xiaogang et al., "Deep learning for generic object detection: a survey," *International Journal of Computer Vision*, vol. 128, pp. 261–318, 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference Learn Represent (ICLR, 2014)*, pp. 1–14, 2014.

[22] A. Namphol, S. H. Chin, and M. Arozullah, "Image compression with a hierarchical neural network," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 1, pp. 326–338, 1996.

[23] J. Y. Cheong and I. K. Park, "Deep CNN-based super-resolution using external and internal examples," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1252–1256, 2017.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[25] K. Kai, L. Hongsheng, Y. Junjie et al., "T-CNN: tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 2896–2907, 2017.

[26] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *2013 IEEE International Conference on Computer Vision*, pp. 2056–2063, 2013.

[27] C. Han, Y. Duan, X. Tao, M. Xu, and J. Lu, "Toward variable-rate generative compression by reducing the channel redundancy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1789–1802, 2020.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 91–99, 2015.

[29] S. V. Kothiya and K. B. Mistree, "A review on real-time object tracking in video sequences," in *2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO, 2015)*, pp. 1–4, 2015.

[30] J. Li, B. Li, J. Xu, R. Xiong, and W. Gao, "Fully connected network-based intra prediction for image coding," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3236–3247, 2018.

[31] R. Christian, R. Yuan, and I. S. Nazrul, "Object detection and tracking in real time videos international," *Journal of Information Systems in the Service Sector*, vol. 11, no. 2, pp. 1–17, 2019.

[32] B. Erik, E. Volker, and S. Tomas, "Training a convolutional neural network for multi-class object detection using solely virtual world data," in *13th IEEE International Conference on*

Advanced Video and Signal Based Surveillance, pp. 278–285, 2016.

[33] W. Bin, T. Sheng, X. Jun-Bin, Y. Quang-Feng, and Z. Yong-Dong, "Detection and tracking based tubelet generation for video object detection," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 1–14, 2018.

[34] L. Chengyou and T. Hua, "Human action recognition based on template matching," in *Advanced in Control Engineering and Information Science*, pp. 2824–2830, Procedia Engineering, 2011.

[35] Y. Xiao, X. Wang, P. Zhang, F. Meng, and F. Shao, "Object detection based on faster R-CNN algorithm with skip pooling and fusion of contextual information," *Sensors*, vol. 20, no. 19, p. 5490, 2020.

[36] K. Jia, X. Wang, and X. Tang, "Image transformation based on learning dictionaries across image spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 367–380, 2013.

[37] L. Yong-Hwan, K. Bonam, and K. Heung-Jun, "Efficient object identification and localization for image retrieval using query-by-region," *Computers and Mathematics with Applications*, vol. 63, no. 2, pp. 511–517, 2012.

[38] I. Schiopu and A. Munteanu, "Deep-learning-based lossless image coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1829–1842, 2019.