Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

EFFICIENTLY DETECT THE VIDEO OBJECT THROUGH MULTI-LEVEL ATTENTION-BASED RPN NETWORK

AMITESH KUMAR JHA

Assistant Professor, Computer Science and Information Technology, Guru Ghasidas Vishwavidyalaya, Koni, Bilaspur, CG.

Abstract:

In recent years, deep learning algorithms have grown in prominence in the field of video object detection. However, in existing systems, Sobel, canny, and wavelet filter-based approaches are employed for edge recognition, but they are unable to identify edges that are blurred or occlusion. Furthermore, the inclusion of undesired data, as well as a large amount of spatiotemporal data in the video, makes video object recognition difficult. Hence, in this research, a novel Gradient Convolution Based Edge Detection has been proposed, in which the pixel connectivity map has been derived in central, radial, and angular directions to capture the rich information on the blurred or occlusion edges. Hence identified the edges even at blurred or occlusion. Moreover, the existing models like model YOLO and Faster RCNN suffered a trade-off between accuracy and speed by extracting and learning the unrelated features from images without knowing their importance level, which provokes time complexity, thus reducing the speed. To solve this, a novel multi-level attention-based RPN has been proposed, in which three attention networks are integrated with RPN to extract all the features and effectively detect the object from the video. As a result, the proposed model effectively detects the objects on the video frames by efficiently extracting and learning the features, and increasing learning speed while reducing time complexity.

Keywords: attention networks, blur and occlusion edges, gradient, pixel difference, region proposal network.

1. INTRODUCTION

The detection and recognition of objects with a machine require a lot of computation to extract knowledge about the shapes and objects in an image. Object detection applies in computer vision to the detection of a photo or video for an object [1]. Video recognition varies in many aspects from still-image detection in which video defocus, movement blur and component occlusion remain often difficulties [2]. Current deep learning models are utilized to create useful detectors for such problems, particularly differences in object items including scale, color, shape, and texture. At present, deep neural networks are the basis of the most accurate detection models [3]. Deep convolutional neural networks (CNNs) need high volumes of information for processing but deep features can be processed in a model and used to identify or recognize various artifacts with enough processing in the video detection framework [4]. CNN is the most frequently applied technology for object identification, recognition, and tracking, especially in computer fields [5]. During the monitoring process, object correlation interprets the motions of the subsequent video frames. The detector will also observe and deduce object behavior in the video using the data [6]. Normally, a salient object detector creates a mask for the position of the most attractive region(s) and object(s) in an image. Consequently, video-

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

based detection of salient objects is more difficult while video data normally include more sophisticated spatial scenes, movement cues, cluttered backgrounds, etc., compared to image data. It is challenging to sequential model, both spatial and temporal details for video sequence data [7]. Video object recognition focuses on the exploitation or modeling of spatial-temporal information to ensure continuity between frames [8].In VSOD, moreover, the non-local action will rely instead of motion cues on global context in previous methods [9].

Videos are even less explored for saliency detection. This is attributed largely to the difficulty and absence of large video datasets in temporal character extraction [10]. Many models of handcraft-based video detector objects are based on optical flow for temporal cues, fuse presence, and temporal to produce maps of saliency in prior methods [11]. Optical flow can be used to define the motion cue of moving objects immediately and provide explicit movement details. To capture the temporal motion cue, the spatiotemporal information learning module can be used in the prior method but the accuracy is not enhanced as well as the detection speed is not increased [12]. While the FCN methods have accomplished major success, errors will still occur for classical SOD methods based on FCN in low-contrast areas with outstanding artifacts and backgrounds. In deeper networks, conversely, the background border structure of the salient items is demolished and the region space for progress is limited. The analysis was based on the Previous approach is more precise, but in low - resolution regions near the border, it cannot perform very well [13]. In addition, Moving Object Recognition (MOR) determines the position and description of moving objects in videos [14]. Discrimination against moving objects and video backgrounds is a key challenge for many machine vision applications. Yet various scenarios from the actual universe – such as complex background changes, lighting differences, shading, obstacles such as precipitation, smoke, etc. – are a difficult task to distinguish moving objects [15].In moving object detection, the related motion data is defined in the foreground and background regions via the class-agnostic classification of video frames. These approaches however do not classify the foreground regions into their respective groups of objects [16].

In the field of object recognition, the instances of objects are located and classified accordingly into each type, i.e. vehicles, humans, cars, dogs, etc. in a specific class [17]. They only operate on static images but do not consider the temporal behavior to distinguish between moving and not moving objects. It is critical for many applications that only moving objects are identified and categorized [18]. Most research works were proposed to tackle motion cues, rapid movement, background changes, not sufficient spatiotemporal fusion features, and distinguish between moving and non-moving object detection. Therefore, to detect the object proficiently from videos a creative strategy has to be developed. The main contributions of this paper are as follows:

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

- Gradient-based edge detection is proposed to identify the appropriate edges even at blurred or occlusion.
- Multilevel attention-based RPN is proposed to detect the object from the video as well as extract and learn all the features of the video frame.

The content of the paper is organized as follows: section 1 represents the introduction; section 2 presents the related work; the novel solutions are presented in section 3; the implementation results and its comparison are provided in section 4; finally, section 5 concludes the paper.

2. LITERATURE SURVEY

Geng et.al [19] presented a video object detection Object-aware Feature Aggregation (OFA) module (VID). This approach is motivated by the fascinating property that knowledge of video objects can be used as an effective semantic before helping to recognize objects. As a result, increasing characteristics of the previous knowledge will effectively boost classification and location to allow the features to access more content in the whole video, first acquire object-conscious information about concepts and integrate this knowledge into existing pair-wise contexts. However, in this method object tracking and action classification of objects in the video may not be focused on.

Qi et.al [20] proposed a high-speed salient video detection of objects at 0.5s per frame (including an average of 0.32 s for optical flow computation). It consists primarily of two modules: the original Spatio-temporal salience module and the salient temporal propagation module based on the correlation filter. The former combines spatial output with the high minimum obstacle gap and the border contrast indicator with the data of temporal output from the motion region. The above requires similarity filters to ensure the continuity of the nearby frames. Finally, the two above modules are adaptively combined. However, in this method, the spatiotemporal fusing features are not sufficient. Dong et.al [21] presented that end-to-end SOD video algorithm that showed efficient knowledge about the movement of objects. This algorithm is comprised of two main elements: the 3D-convolution X-shape structure that essentially depicts the move details in successive video frames and a pyramid-based 2D densely connected convolutional neural network (DenseNet). This approach not only preserved a small amount but also uniformly represented spatiotemporal details, which allows the end-to-end training of 2D convolution network parameters. However, in this method motion cues in video frames cannot capture in the static image as well as long-term motion cues cannot integrate.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Guo et.al [22] proposed a computer-efficient and sufficiently reliable approach for defining the most visible object in a video series of spatiotemporal objects. The fundamental motion in a video is intuitively a more stable measure of saliency than the visible color signals, often involving major differences and complex structures. On this basis, developed an efficient and accurate space-temporal saliency detection method that uses motion data to identify the most dynamic regions in a video series. To achieve preliminary priors and then integrate spatial characteristics such as appearance contrasts and compactness in a multi-cue integration system, evaluate the optical flow field first and merge different saliency markings to achieve temporal compatibility. However, in this method, a rapid movement of objection detection is not explored.

Ray et.al [23] proposed a novel technique in variable context videos that are taken by moving cameras with no extra sensor for sensing and tracking objects. In videos with a variable background, the efficiency of monitoring depends on the effective identification of an object in the variable background. To detect an object in a variable context, the most desirable aspect is that it does not rely on prior scene information. The backdrop and foreground shift in every frame of the picture series when the footage was taken by a moving camera. For these images, it is also difficult to model a single background using conventional background approaches and, therefore, to detect real moving objects in a variable background. Then the individual objects that are part of an entity are fused into the moving object in the variable background using the minimal spanning tree. In this work, Linear Assignment Problem resolves the problems of the data association during the monitoring and the use of a kalman filter handles occlusion. The key advantage of this approach over most current tracking algorithms is that in the first frame or training for the sample data, the proposed method does not need to be initialized. However, in this method complex variable background cannot be detected.

Zhu et.al [24] proposed a novel for achieving high performance of video detection. Their deep convolutional network is only used to extract image features in key frames, and the optical flow network is combined to fuse motion information between key frames, while non-key frame features are obtained by updating the motion part's features based on previous key frame features using the optical flow network. At the same time, the key frame selection is adaptively determined based on the feature map's quality. However, this method is needed to steadily improve the detection performance.

From the literature survey, [19] results must focus on object tracking and action classification and, in [20] not sufficient spatiotemporal fusing features and in [21] video frames cannot capture the static image and in [22] object detection is not explored and in [23] cannot detect the complex variable background and in [24] must improve the detection performance. The description of the proposed technique is discussed in the following section.

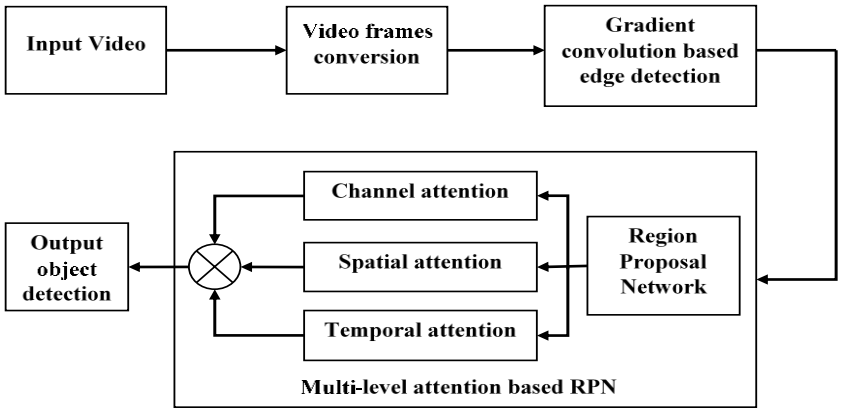
Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

3. VIDEO OBJECT DETECTION THROUGH MULTI-LEVEL ATTENTION-BASED RPN NETWORK

Videos, as one of the most engaging mediums, have a strong emotional impact on people. However, the existing systems are uses Sobel, canny, wavelet filter-based methods for edge detection but they are not able to detect the edges on blurred or occlusion edges. Moreover, the presence of unwanted information and a vast number of spatiotemporal information in video data causes video object detection complicated. To tackle this issue, a novel gradient-based convolution edge detection has been proposed in which each pixel pair has been convolved with the kernels, and the pixel connection map has been produced in central, radial, and angular directions to capture the rich gradient information on the blurred or occluded regions. By element-wise multiplication, those connectivity maps are convolved with the kernel weights, transforming several connectivity maps into a single map. As a result, even at blurred and obfuscated borders, the necessary edges can be determined. Consequently, in existing object detection systems, a trade-off between accuracy and speed has been presented, such as YOLO, which can achieve high speed while sacrificing accuracy, whereas others, such as RCNN, may obtain accuracy while sacrificing speed. The cause for this may be traced back to the extraction and learning of unrelated characteristics from images without knowing their relevance level, which adds time complexity and slows down the process. To overcome this issue, a novel multi-level attention-based RPN has been introduced, in which learn the RGB channel, space region, and time sequence properties of the frame sequences, the channel, spatial, and temporal attention networks are combined with the region proposal network. Whereas attention networks may learn features simultaneously with knowledge of feature relevance, then concatenate with a correlation of all discriminant characteristics to recognize the object by maintaining the key information while decreasing irrelevant feature learning. As a result, the proposed model effectively identifies the edges on the video frames and efficiently extracts and learns the features from the video frames, improved detection of the accuracy of the object on the video as well minimizes the irrelevant features and increases the learning speed. The proposed system architecture is presented in figure 1.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Fig 1: Architecture of proposed system

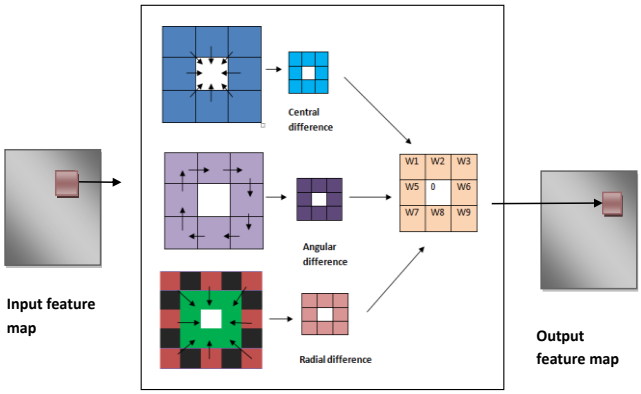


3.1 Gradient Convolution Based Edge Detection

Initially, in the proposed system, the input video clips are converted into several frames. Then these frames are considered for a pre-processing technique that uses a Median filter to remove noises and other unpleasant qualities during the extraction process. If the noises are removed from the image, then the proposed system uses the gradient convolution method for edge detection because the existing systems use Sobel, canny, wavelet filters to detect the object but which cannot be able to detect the edges on blurred or occlusion edges for that reason the proposed system is utilized a novel gradient convolution-based edge detection. Here each pixel pair of the image is convolved with the kernels. To capture the rich gradient information on the blurred or occlusion regions in the image, the proposed system select the pixel pairs in different directions such as central, angular, and radial directions so it only captures the rich information from the image as well as avoid the unnecessary information. Because the existing systems cause to detect the object from the image in the view of the fact that presence the unnecessary information so it takes a long time to complete the process. Therefore, the proposed system captures the rich information to select the pixel pairs in different strategies such as central, angular, and radial differences so the detection time of the edge process is decreased. The three pixels' difference convolution is depicted in figure 2.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Fig 2: Three-pixel difference convolution



The convolution products of the picture pixels with varied masks result in the computation of the horizontal and vertical gradient in image edge detection. The differences between neighboring pixels are used to compute the two gradients. The central difference are expressed by the following first order derivate:

$$\frac{\partial L}{\partial x} \approx \frac{L(x+1,y)-L(x-1,y)}{2} \quad (1)$$

$$\frac{\partial L}{\partial y} = \frac{L(x,y+1)-L(x,y-1)}{2} \quad (2)$$

The two derivates are equivalent to a convolution kernel consisting of the horizontal convolution {-1, 0, +1} and the vertical convolution {-1, 0, 1}. Then the horizontal (Vx) and vertical (Vy) gradients are obtained by applying these convolutions to a grey-scale picture. To conduct convolution on the entire image, build an n x n (usually 3 x 3) number matrix called kernel mask and multiply it with a part of the image of the same dimension. The center pixel value is then calculated by summing all of the products. The n x n matrix is given below.

1	0	-1
1	0	-1
1	0	-1

V(x)

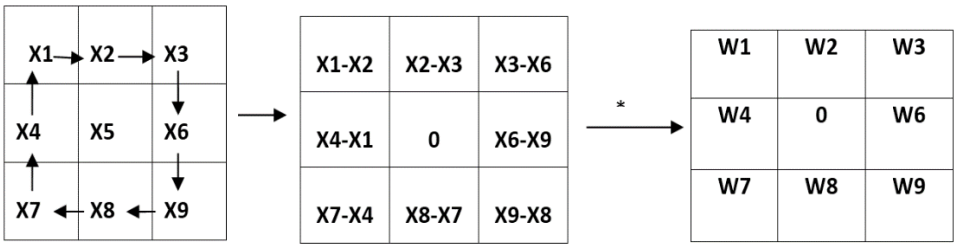
1	1	1
0	0	0
-1	-1	-1

V(y)

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

The pixel differences from the pairs are then convolved with the kernel by performing element-wise multiplication with the kernel weights, followed by a summation, to produce the value in the output feature map. While the beneficial encodings of pixel relations are kept in the trained convolution kernels, as with CNN, these convolution kernels are encouraging to have a larger inner product with those key encodings to achieve greater activation responses. In this way, the proposed system easily identifies the appropriate edges even at blurred or occlusion edges. The selection process of pixel pairs in the angular difference convolution is presented in figure 3.

Fig 3: Selection of pixel pair and convolution in angular difference convolution



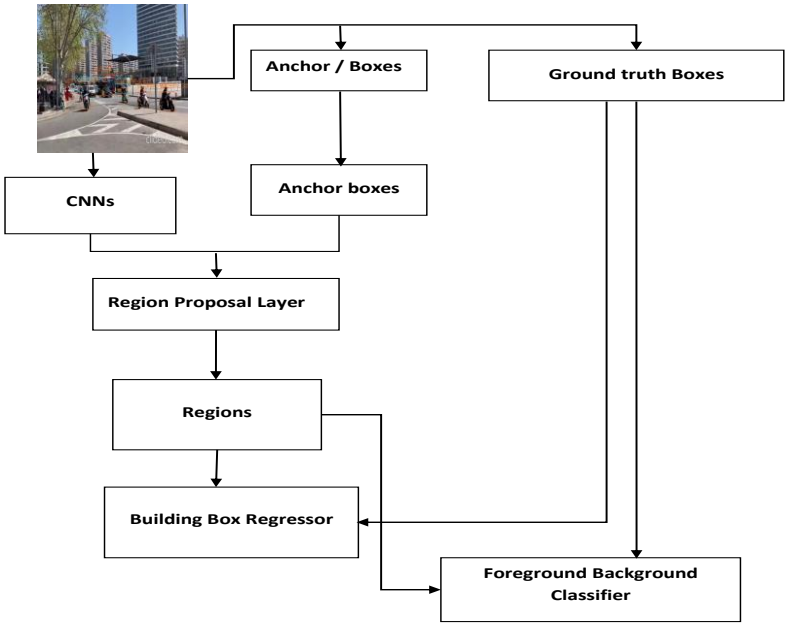
Once the edges of the objects are properly identified through the proposed method, then efficiently extract, learn all the features and detect the object from the video frames in this work using multilevel attention-based RPN.

3.2. Multi-level Attention-based RPN

Like the faster RCNN, the proposed system uses a region proposal network (RPN) for detecting the object from the videos. The region proposal network is taking images as input and produces rectangle regions proposals with an objectness score for each proposal. The Region Proposal Network does not anticipate region proposals' boundaries from ground zero; rather, it predicts offsets using established reference boxes.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Fig 4: RPN in training Network



The region proposal network training process is shown in fig. 4. Detect the object from the video using the region proposal network is described step by step is given below.

Step 1:

The input image is sent through the convolution neural network its last layer gives the features maps as output.

Step 2:

In this step, a sliding window is running through the feature maps acquired in the final stage. The size of the sliding window is n × n. And a unique set of anchors is created for each sliding window, although they have varied aspect ratios and scales. Then these anchor boxes are centered at the image's anchor point, which corresponds to the feature map's anchor point. With a feature map of size W*H and K anchors for each position of the feature map, the total number of anchor boxes is calculated using W*H*K.

Step 3:

Each point of the feature map has n × n anchor boxes. However, there might be a lot of boxes that are empty. So in this step, the model is trained to learn which anchor box has the object as well as which anchor box is empty. If an anchor box contains an object or a portion of an object, it is referred to as foreground; if the anchor box does not include an object, it is referred to as background. At the same time, the model is learning the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

offsets for the foreground boxes to modify for object-fit.

Step 4:

In step 4, the intersection over union (IoU) score of a Ground Truth Box with anchor boxes is calculated by the Bounding Box Classifier, which then classifies the Anchor box as foreground or background with a specified likelihood which is referred to as objectness score. Then the bounding box regressor layer is learning the offsets concerning ground truth box for anchor box it is classified as a foreground as well as its predicts the labels. If the training process is done, then the model is tested. Then the three different attention networks such as channel attention model, spatial attention model, and temporal attention models are concatenated with the region proposal network for learning the features from the video.

Spatial Attention Module:

In the spatial attention module, the coordinates of the bounding boxes are projected on the $\mathrm{n} \times \mathrm{n} \times \mathrm{d}$ activation map to extract the spatial features. For the projected region, the proposed system takes a scalar feature by taking a weighted average over each feature map. Then the features are extracted from the bounding boxes detected object including the vehicle, buildings, trees, and the other things of the object so the spatial attention module is learning only the vehicle features from the video and avoid the unnecessary features, as well as the spatial attention module, gives the pure features of the diverse types of car, motor-cycle, and bus, etc. In this work, focused on a most significant region of a video sequence the spatial attention mechanism to a dynamic weighted sum of the top-n local features to obtain a single spatial local feature on each frame such that,

$$\mathrm{S}_{\mathrm{i}}^{\mathrm{t}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_{\mathrm{ij}}^{\mathrm{t}}\, \mathrm{vl}_{\mathrm{ij}} \qquad (3)$$

Where

- $\delta_{\mathrm{ij}}^{\mathrm{t}}$ represents the spatial attention weight
- $\mathrm{vl}_{\mathrm{i}} = \mathrm{vl}_{\mathrm{i1}}, \mathrm{vl}_{\mathrm{i2}}, \ldots, \mathrm{vl}_{\mathrm{in}}$ is the top-n local features

The weight of the spatial attention weight is calculated at each time and $\sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_{\mathrm{ij}}^{\mathrm{t}} = 1$. The spatial attention module is catch and learns the most significant features according to increasing the weights.

Temporal Attention Module:

The temporal attention module generates the feature vector for each video frame. Then the attention score is obtained by applying a linear mapping followed by a sigmoid function. The attention score is calculated by using,

$$\beta_{\mathrm{i}}^{\mathrm{T}} = \sigma(\emptyset^{\mathrm{T}} \mathrm{y}_{\mathrm{i}}), \mathrm{i} = 1, \ldots, \mathrm{T} \qquad (4)$$

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Where:

- $\sigma(.)$ is the sigmoid function
- β_i^T is the temporal attention matrix
- $\emptyset$ is the vector parameters of the linear mapping

If the proposed system obtains an attention score of each frame in the video then combine the attention scores (i=1, 2… T) With frame-level feature vectors to create a weighted feature vector as follows,

$$q^t = \sum_{i=1}^{T} \beta_i^T y_i \quad (5)$$

Where

- q^t is the temporal feature vector for the entire video which takes into account the importance of each frame in the video.

Channel Attention Module:

The channel attention module extracts all the features from the input frame. The channel attention module is used to learn how to weight the individual detected object. The final prediction is generated as,

$$X = \sum_{a=1}^{A} w_a \,.\, \mathrm{softmax}(Y_a) \quad (6)$$

Where

- Y_a is the feature map for each detected object $a \in \{1, \dots, A\}$
- w_a is the weight of the channel attention module.

Finally, all the attention networks are combined and concatenate to the region proposal network. The combined network is expressed by,

$$AN = S_i^t + q^t + X \quad (7)$$

Where

- S_i^t is the spatial attention module
- q^t is the temporal attention module
- X is the channel attention module

The channel, spatial, and temporal attention models are including long-term information into the learning of action models and they learn actions models well organized with segment-based sampling and aggregation scheme. The channel attention model is focused on 'what' is relevant in the context of an input image as well as it improves the performance of the segmentation. Simultaneously, spatial attention models improve in terms of performance and interpretability in visual tasks such as image categorization. In both spatial and temporal streams, has a simple but effective module channel

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

attention unit (CAU) that can selectively highlight useful features while preserving the discriminative qualities of the original features, as well as the three different attention models, are efficient in learning the RGB channels, space regions and time sequence features of the frame sequences. For the reason that the attention networks are trained features simultaneously with knowledge of feature relevance, then concatenated with a correlation of all discriminant characteristics to recognize the object by retaining the vital information and eliminating unnecessary feature learning.

As a result, RPN is reduce time complexity and increases speed as well as detects the objects from the video. Here the RPN is efficiently built using a fully convolutional approach, with the convolutional feature map.

4. RESULTS AND DISCUSSION

This segment provides a detailed description of the implementation results as well as the performance of the proposed system and a comparison section to ensure that the proposed system performs valuable.

4.1 Experimental Setup

This work has been implemented in the working platform python with the following system specification and the simulation results are discussed below.

Platform	: Python
OS	: Windows 10
Processor	: 64-bit Intel processor
RAM	: 8 GB RAM
Dataset	**:** Road traffic video monitoring.

4.2 Dataset Description

The footage was captured over for two days from a stationary camera overlooking I-5 in Seattle, Washington. Light, medium, and heavy traffic were manually tagged on the film, corresponding to free-flowing traffic, reduced-speed traffic, and stopped or very slow-moving traffic, respectively. Because the first frame of the original video has been contaminated by another video signal, you must begin processing each video from the second frame.

4.3 Simulation Results

The simulation results of the proposed system are given below.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Fig 5: Filtered video frame

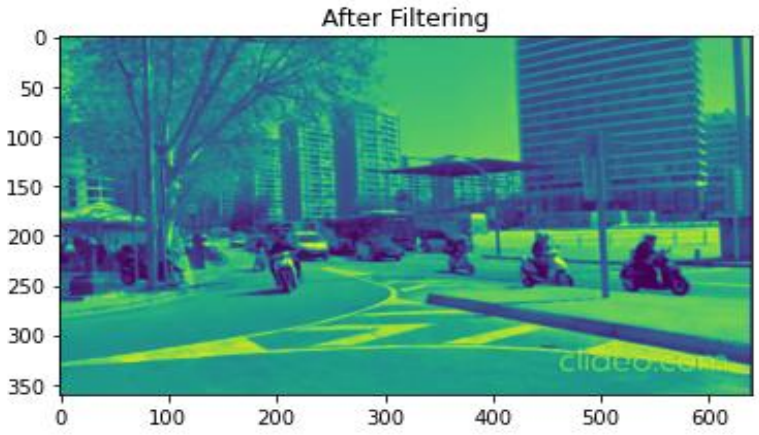
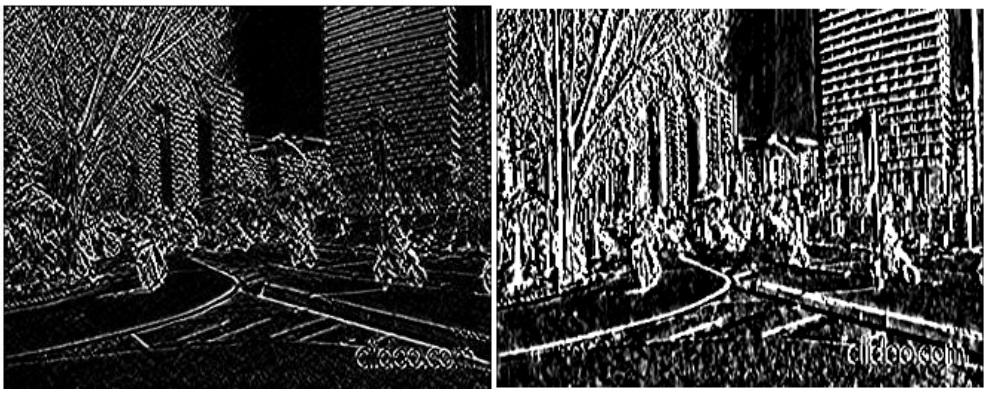


Figure.5. show the simulation result of the filtered video frame. The proposed system used pre-processing method to remove the noise. During pre-processing, the median filter is used to remove the impulse noise and salt and pepper noise. In the median filter, the pixel's intensity value is replaced with the median of the pixel's neighborhood. The main purpose of the median filter is it reduces the impulse noise. The impulse noise may be minimum value (0) or maximum value (255). Here the median filter also reduces the salt and pepper noise. In this work median filter aids to reduce noise and protect the edge.

Fig 6: Simulation view of (a) central, (b) angular, (c) radial connectivity maps, and (d) detected edges in a video frame



(a) (b)

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

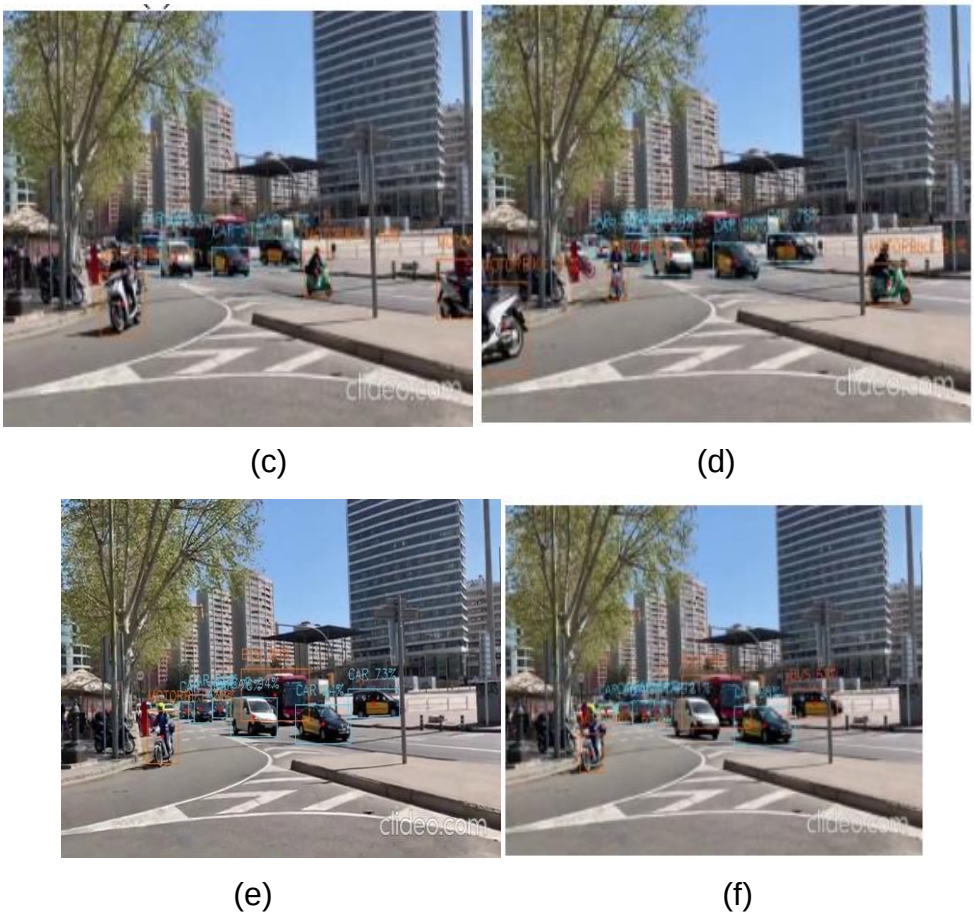
(c) (d)

Figure 6. Shows the three types of pixel difference. A novel gradient-based edge detection is proposed to identify the edges on the video frame in which each pixel pair are convolved with the kernels, and the pixel connection map is produced in central (a), radial (c), and angular (b) directions to capture the rich gradient information on the blurred or occluded regions. By element-wise multiplication, those connectivity maps are convolved with the kernel weights, transforming several connectivity maps into a single map. As a result, the proposed system identifies the edges even the video frame is blurred or occluded (d).

Fig 7: Simulation result of object detection

(a) (b)

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF



(c) (d)

(e) (f)

Figure 7. Shows the simulation outputs of the object detection using multilevel attention based RPN in which all the features are extracted and learned from the video frames using three attention models such as spatial, temporal, and channel then the three attention models are integrated with the region proposal network for detecting the object in the video frames.

4.4 Performance metrics of the proposed method

The performance of the proposed system implemented is evaluated with the following metrics as described below.

4.4.1 Accuracy

The accuracy of the input data is calculated using,

$$\text{Accuracy} = \left[\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}\right] \quad (8)$$

TP- True Positive Value

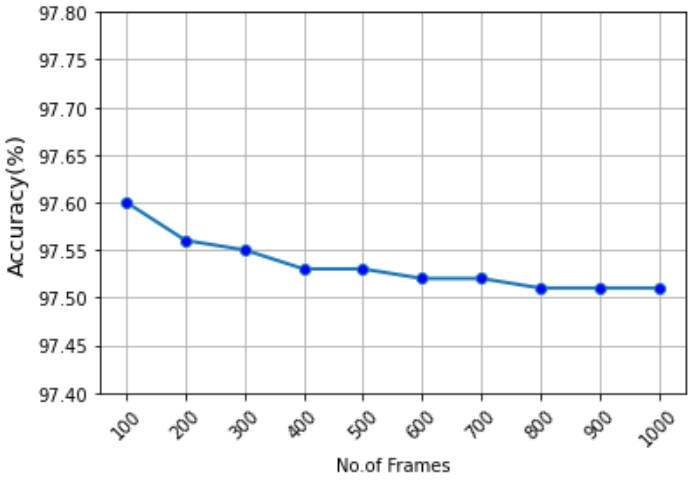
Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

TN- True Negative Value

FP- False Positive Value

FN- False Negative Value

Fig 8: Accuracy of the proposed system



The above-mentioned figure 8 clearly explains the accuracy of the proposed system. The accuracy of the proposed system attains a value of 97.60%. The proposed system's accuracy is improved by extracting the features from the spatial, temporal, and channel networks for detecting the objects.

4.4.2 Recall

The recall of the input data is calculated using,

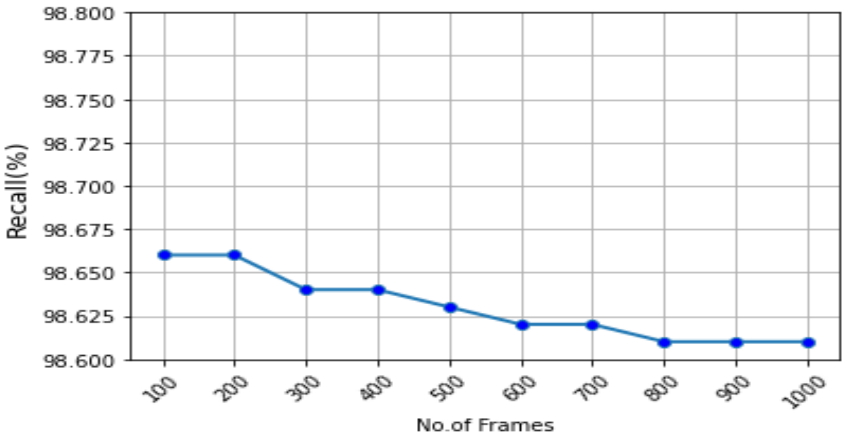
$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (9)$$

TP- True Positive Value

FN- False Negative Value

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Fig 9: Recall of the proposed system



The proposed approach achieves a recall rate of 98.660 percent which is depicted in fig 9. The recall of the proposed system is improved by using multilevel attention-based RPN with feature extraction from the spatial, temporal, and channel networks.

4.4.3 Precision

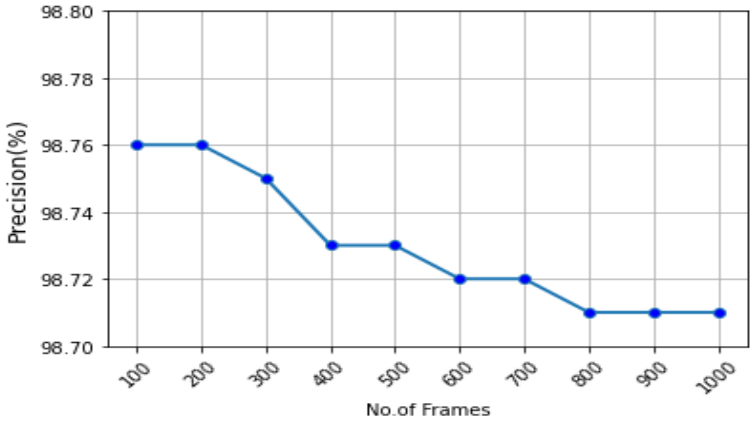
The precision of the input data is calculated using,

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

TP- True Positive Value

FP- False Positive Value

Fig 10: Precision of the proposed system



The proposed system's precision is explained in the graph above. The suggested system has a 98.76% precision rate. Multilevel attention-based RPN improves the precision of

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

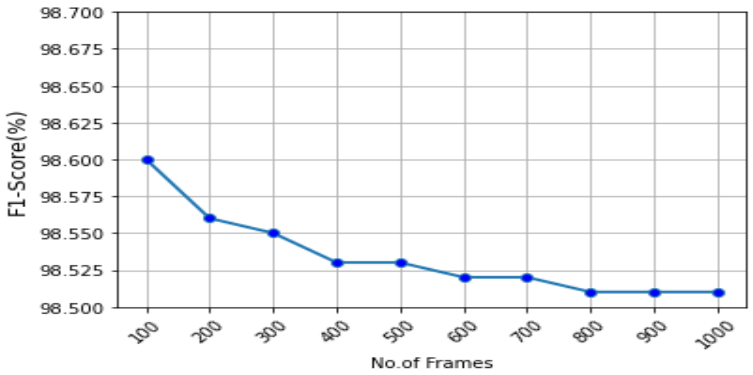
the proposed system lowering time complexity and enhancing speed. It is also clear that as the number of frames gets increased the precision is reduced.

4.4.4 F1-Score

The F1-Score of the input data is calculated using,

$$F1=2*\frac{Precision*Recall}{Precision+Recall} \quad (11)$$

Fig 11: F1-score of the proposed system



The F1 score of the proposed approach is easily explained in the graph above. The proposed method achieves an F1 score of 98.600 percent by using multilevel attention-based RPN that extracts all features from the spatial, temporal, and channel networks, then combines the three attention networks to form a region proposal network to detect objects.

4.4.5 Sensitivity

Precision of the input data is calculated using,

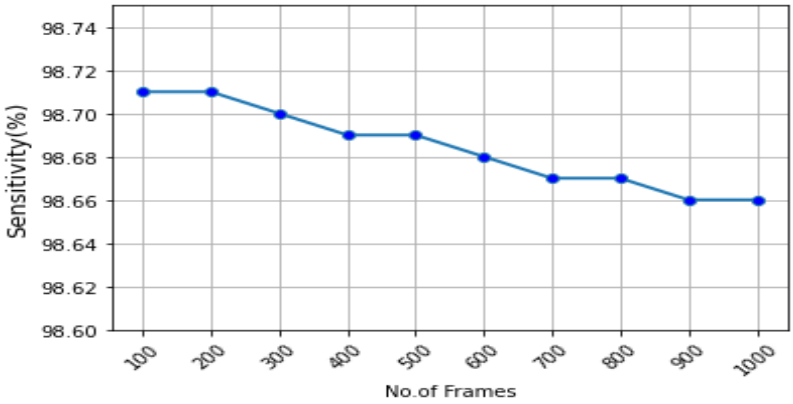
$$Precision = \frac{TP}{TP+FN} \quad (12)$$

TP- True Positive Value

FN- False Negative Value

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Fig 12: Sensitivity of the proposed system

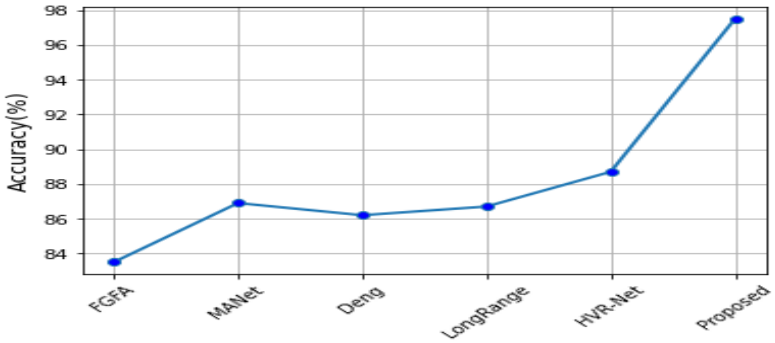


In the graph above, the sensitivity of the suggested strategy is explained with frames ranging from 100 to 1000. The proposed system's sensitivity is calculated to be 98.71%. The sensitivity of the proposed system is improved by employing multilevel attention-based RPN by combining the three attention networks.

4.5 Performance comparison of the proposed method

This section describes the various performances of the proposed method comparing with the results of previous methodologies and depicting their results based on various metrics.

Fig 13: Accuracy Comparison

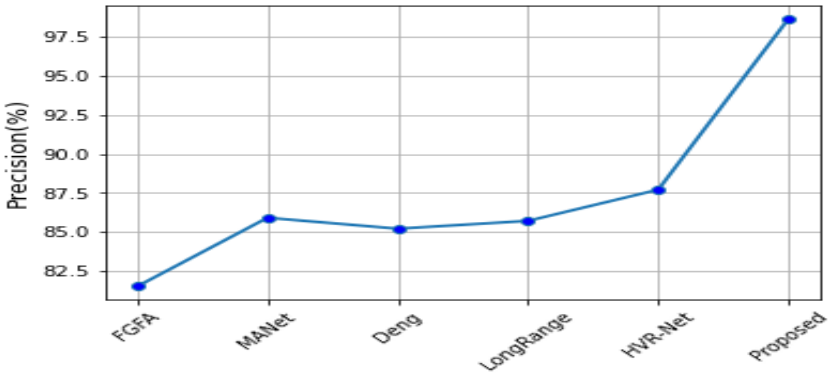


The accuracy of the proposed system is compared with the accuracy of the various previously proposed accuracies such as Filament-guided filament assembly (FGFA), Mobile ad hoc network (MANet), Deng, Long-range, and Hierarchical video relation

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

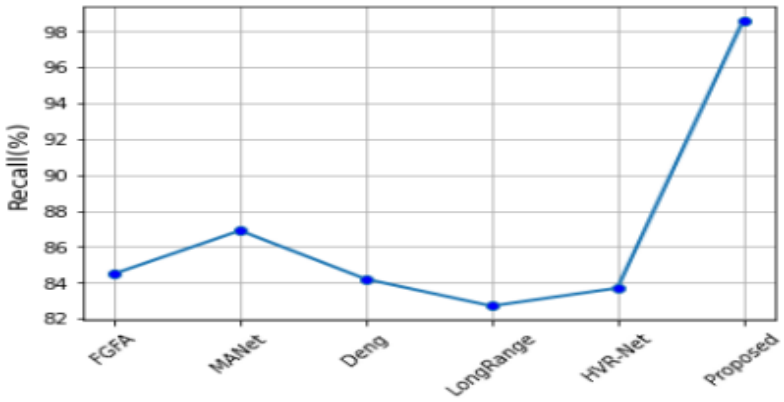
network (HVR-Net). From the graph, it is clear that the accuracy of the proposed system is high that is 97.60% than the existing output when compared with the accuracy of FGFA [25] which is 82%, MANet [26] is 87%, Deng [32] is 86%, Long-range [28] is 87%, and HVR-Net [31] is 88.92% and from the conclusion, it is noted that FGFA has the lowest accuracy whereas our proposed system has the highest accuracy.

Fig 14: Precision Comparison



When compared to the precision of FGFA [25] which is 81 percent, MANet [27] which is 86 percent, Deng [32] which is 85 percent, Long-range [30] which is 86 percent, and HVR-Net [31] which is 87.5 percent, it is clear that the proposed system has the highest precision (98.660 percent) than the existing output.

Fig 15: Recall Comparison

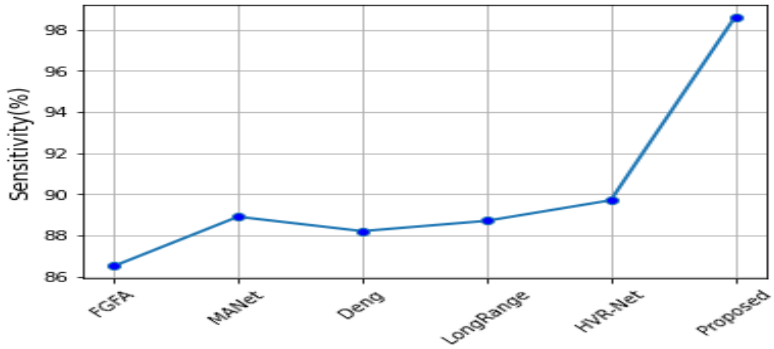


The proposed system has a high recall of 98.660 percent when compared to the recall of FGFA [25] which is 84.52 percent, MANet [27] which is 87 percent, Deng [32] which is 84 percent, Long-range [30] which is 82.99 percent, and HVR-Net [31] which is 83.99 percent, and the conclusion is that long-range has the lowest recall whereas our

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

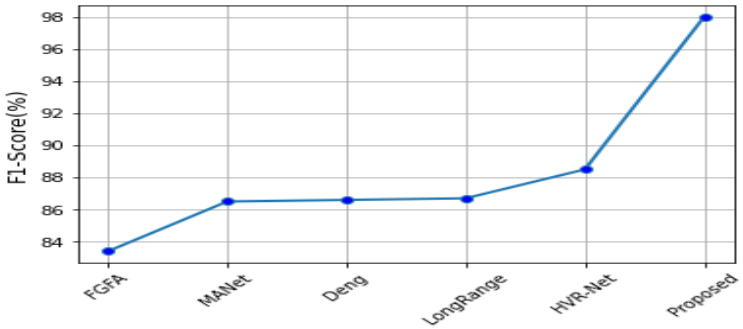
proposed system has the highest recall.

Fig 16: Sensitivity Comparison



When compared to the sensitivity of FGFA [25], which is 86.88 percent, MANet [26], which is 88.99 percent, Deng [32], Long-range [28], and HVR-Net [31], which is 89.99 percent, it is obvious that the suggested system has the maximum sensitivity, which is 98.71 percent.

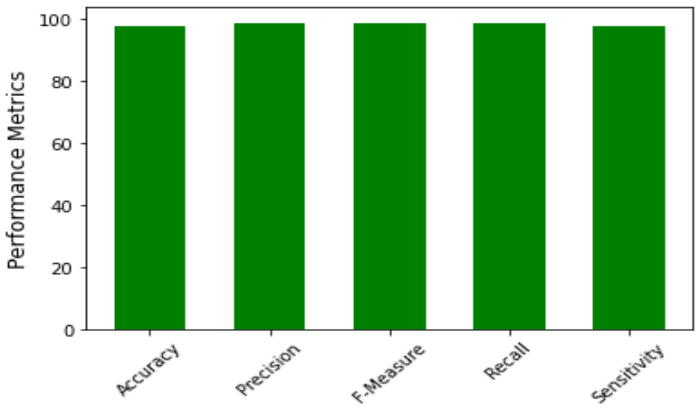
Fig 17: F1 score Comparison



The f1 scores of FGFA [25] which is 83 percent, MANet [27] which is 86.88 percent, Deng [32] which is 86.88 percent, Long-range [29] which is 86.90 percent, and HVR-Net [31] which is 89.92 percent, it is evident that the suggested system f1 score is high, at 98.71 percent.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

Fig 18: The performance bar chart



In figure17, the bar chart clearly explains the performance of the proposed system. From the bar chart, it is clear that the performance of the proposed system is high in which performance metrics are accuracy is 97.60%, precision is 98.76%, recall is 98.660%, sensitivity is 98.71%, and F1-score is 98.600%. The performance of the proposed system is increased by using Multilayer attention-based RPN.

5. CONCLUSION

In this research, the proposed system used a novel gradient-based edge detection for identifying the appropriate edges even at blurred or occulated in which the preprocessing method is used to divide the video frames and the median filter removes the noise from the input video frames and three types of pixel difference are combined to a single map, therefore, the appropriate edges are easily identified. Then the extracted features from the video frame and detection of an object in the video clips are done by using novel multilevel attention based RPN in which all the features are extracted from three different attention networks such as spatial, temporal, and channel then the three different attentions networks are combined with the region proposal network for detecting the objects. Thus the proposed system reduced the time complexity and increased the speed and the performance metrics in terms of accuracy is improved by 97.60%, recall is improved by 98.66%, precision is improved by 98.76%, the F1 score is improved by 98.60%, and sensitivity is improved by 98.71% and also showed better performance when compared with other techniques.

REFERENCES

1) T. Ahmad, Y. Ma, M. Yahya, B. Ahmad, and S. Nazir, “Object detection through modified YOLO neural network,” Scientific Programming, 2020.

2) D. Li, R. Wang, C. Xie, L. Liu, J. Zhang, R. Li, F. Wang, M. Zhou, and W. Liu, “A recognition method for rice

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

plant diseases and pests video detection based on a deep convolutional neural network," Sensors, vol. 20, no. 3, pp. 578, 2020.

3) F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita and F. Herrera, "Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," Knowledge-Based Systems, vol. 194, pp.105590, 2020.

4) C.C. Poon, Y. Jiang, R. Zhang, W.W. Lo, M.S. Cheung, R. Yu, Y. Zheng, J.C. Wong, Q. Liu, S.H. Wong, and T.W. Mak, "AI-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices," NPJ Digital Medicine, vol. 3, no. 1, pp.1-8, 2020.

5) Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," Neurocomputing, vol. 323, pp.37-51, 2019.

6) Z. Kadim, M.A. Zulkifley, and N. Hamzah, "Deep-learning based single object tracker for night surveillance," International Journal of Electrical & Computer Engineering, (2088-8708), vol. 10, 2020.

7) Y. Ji, H. Zhang, Z. Jie, L. Ma, and Q.J. Wu, "Casnet: a cross-attention siamese network for video salient object detection," IEEE transactions on neural networks and learning systems, 2020.

8) G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3243-3252, 2018.

9) Y. Gu, L. Wang, Z. Wang, Y. Liu, M.M. Cheng, and S.P. Lu, "Pyramid constrained self-attention network for fast video salient object detection," In Proceedings of the AAAI on Artificial Intelligence, vol. 34, no. 07, pp. 10869-10876, 2020, April.

10) Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis, "SCOM: Spatiotemporal constrained optimization for salient object detection," IEEE Transactions on Image Processing, vol. 27, no. 7, pp.3345-3357, 2018.

11) K. Huang, G. Li, and S. Liu, "Learning channel-wise spatio-temporal representations for video salient object detection," Neurocomputing, vol. 403, pp.325-336, 2020.

12) Z. Wang, J. Li, and Z. Pan, "Cross Complementary Fusion Network for Video Salient Object Detection," IEEE Access, vol. 8, pp.201259-201270, 2020.

13) Y. Wang, K. Chen, and Y. Song, "Real-time salient object detection with boundary information guidance," Neurocomputing, vol. 412, pp.437-446, 2020.

14) M. Mandal, L.K. Kumar, and M.S. Saran, "Motionrec: A unified deep framework for moving object recognition," In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2734-2743, 2020.

15) D. Zeng, and M. Zhu, "Multiscale fully convolutional network for foreground object detection in infrared videos," IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 4, pp.617-621, 2018.

16) L.A. Lim, and H.Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," Pattern Recognition Letters, vol. 112, pp.256-262, 2018.

17) S. Javed, A. Mahmood, S. Al-Maadeed, T. Bouwmans, and S.K. Jung, "Moving object detection in complex scene using spatiotemporal structured-sparse RPCA," IEEE Transactions on Image Processing, vol. 28, no. 2, pp.1007-1022, 2018.

18) L. Li, Q. Hu, and X. Li, "Moving object detection in video via hierarchical modeling and alternating optimization," IEEE Transactions on Image Processing, vol. 28, no. 4, pp.2021-2036, 2018.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/RPCQF

19) Q. Geng, H. Zhang, N. Jiang, X. Qi, L. Zhang, and Z. Zhou, “Object-aware Feature Aggregation for Video Object Detection,” arXiv preprint arXiv: 2010.12573, 2020.

20) Q. Qi, S. Zhao, W. Zhao, Z. Lei, J. Shen, L. Zhang, and Y. Pang, “High-speed video salient object detection with temporal propagation using correlation filter,” Neurocomputing, vol. 356, pp.107-118, 2019.

21) S. Dong, Z. Gao, S. Pirbhulal, G.B. Bian, H. Zhang, W. Wu, and S. Li, “IoT-based 3D convolution for video salient object detection,” Neural computing and applications, vol. 32, no. 3, pp.735-746, 2020.

22) F. Guo, W. Wang, Z. Shen, J. Shen, L. Shao, and D. Tao, “Motion-aware rapid video saliency detection,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp.4887-4898, 2019.

23) K.S. Ray, and S. Chakraborty, “Object detection by spatio-temporal analysis and tracking of the detected objects in a video with variable background,” Journal of Visual Communication and Image Representation, vol. 58, pp.662-674, 2019.

24) X. Zhu, J. Dai, L. Yuan, Y. Wei, “Towards High Performance Video Object Detection”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7210–7218.

25) Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei, “Flow-guided feature aggregation for video object detection,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 408–417, 2017.

26) Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng, “Fully Motion-Aware Network for Video Object Detection”, ECCV, 2018.

27) Safaa Laqtib, Khalid El Yassini, Moulay Lahcen Hasnaoui, “A Deep Learning Methods for Intrusion Detection Systems based Machine Learning in MANET”, https://www.researchgate.net/publication/338027948, 2019.

28) Mykhailo Shvets, Wei Liu, Alexander C. Berg, “Leveraging Long-Range Temporal Relationships between Proposals for Video Object Detection”, IEEE/CVF, 2019.

29) Chiman Kwan, Bence Budavari, “A high performance approach to detecting small targets in long range low quality infrared videos,” Signal, Image and Video Processing, 2021.

30) Mrunalini Nalamati, Ankit Kapoor, Muhammed Saqib, Muhammed Saqib, Michael Blumenstein, “Drone Detection in Long-range Surveillance Videos,” IEEE, 2019.

31) Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao1Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao, “Mining Inter-Video Proposal Relations for Video Object Detection,” European Conference on Computer, 2020.

32) H. Deng, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, H. Guan, “Object guided external memory network for video object detection”, In: ICCV, pp.6678–6687, 2019.