An Efficient Region Based Object Detection method using Deep learning Algorithm

R Anitha[1], Dr.S.Jayalakshmi[2]

[1]Research Scholar, Department of Computer Science, V I S T A S, Pallavaram, Chennai, Tamilnadu, India.
scholar.anitha@ gmail.com
[2]Professor, Department of Computer Applications, V I S T A S, Pallavaram, Chennai, Tamilnadu, India.
jai.scs@velsuniv.ac.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract—
In the field of intelligent machine learning and computer vision techniques, the research of object detection has been improving step by step for several years. It also has been used across many of our real-time surveillance applications like Traffic detection, Vehicle detection, Road detection, face detection, Pedestrian detection, Fruit detection, Object tracking etc. The Roll of dataset and their size in all these detection plays a vital role. The proposed Smart-Region based detection method actively pre-process, classifies, validates and stores the new images. The SRBD method combining with YOLO-v3 algorithm achieves better results in the detection of small objects where the YOLO-v3 alone unable to give satisfactory results. Therefore the proposed SRBD method gives better results with good accuracy in vehicle detection over the other detection methods.

Keywords— Object Detection, Faster R-CNN, SSD, YOLO-V1, YOLO-V2, YOLO-V3, Smart-Region, Detection.

I. INTRODUCTION

The concept of object detection techniques spy rocketed along with many advancements of Convolution Neural Network (CNN). Further, the CNN had developed into the Region based Convolution Neural Networks(R-CNN). The R-CNN combines the proposals of Regions of Interest and CNN achieves better speed and accuracy in object detection real-time environment. Based on Region Proposal Network (RPN), the object detection algorithm is made better and faster. At each detection, the RPN simultaneously predicts object boundaries and object scores. The RPN is a fully convolution network and it is being used in various public safety aspects like identifying the road accidents, traffic snarls, pedestrian/vehicle detection, and License number plate detection, etc.

With the rapid development of Object detection techniques, it has been deployed in more Real-Time applications, such as (i) human-computer interaction (HCI) [1], (ii) Hospital service robotics (e.g., COVID-19 robots), (iii) security (e.g., Object Recognition, Multiple Object Tracking)[4], (iv) Scale-invariant Feature transformation(SIFT)[2] (v) retrieval (e.g., search engines) [2], and (vi) transportation (e.g., autonomous and assisted driving)[3]. These applications has different requirements, including: speed, accuracy, processing time (off-line, on-line, or real-time), robustness to occlusions, invariance to rotations (e.g., in-plane rotations) [2], and vehicle detection under pose changes [5].

While many applications consider the detection of a single object class (e.g., Face detection) and from a single view (e.g., frontal faces), others require the detection of multiple object classes (humans, vehicles, etc.), or a single class from multiple views (e.g., side and frontal view of vehicles). In general, most systems can detect only a single object class from a restricted set of views and poses [1].

II. RELATED WORK

Ross Girshick et al[5][2014] proposed a simple and scalable detection algorithm and improved mean average precision (mAP) of objects. Rautaray et al.[1][2015] focused on detecting, tracking, and recognition of objects using gesture recognition systems for human computer interaction. Joseph Redmon et al.[12][2016] introduced the combination of YOLO and Fast YOLO in the unified architecture. And it detected mainly the real-time objects. Joseph Redmon et. al[13][2016] proposed the improved model of the YOLO algorithm and improved the training of object detection and classification using 9000 images. Liang Zheng et al [2][2017]focused on the retrieval of the image in SIFT-based and CNN-based models. Ankita et al[3][2017] proposed a new technique to detect driver's face movement, gaze movement, and closure of eyes to detect the

condition of the drivers. Yongzheng Xu et al[7][2017]proposed the car detection framework using the Faster R-CNN algorithm and the author tried to improve the speed of detection in Unmanned aerial vehicles. Mingyu Li et al[4][2018] proposed the pose estimation of the object in foreground and background boundaries. LongshengFu et al[6][2018] proposed and implemented a new technique using images acquired from natural conditions with multiple clusters and got good detection accuracy. Kyoungmin Lee et. al[10][2018]proposed a single-stage detector using multi-scale features in the region proposal network. The author introduced the 3-way Resblock method to achieve a higher score of prediction than SSD on PASCAL VOC and MS-COCO. It has maintained the fast computation in single-stage detectors. Joseph Redmon[14][2018] proposed the new faster object detection network which was derived from YOLO-v2. It has got three times improvement than the existing algorithm in speed. Also, it got step-by-step improvements in various dataset training. Chen Linkai et al[15][2018] proposed the effective framework for vehicle detection and classification from traffic surveillance, the Faster-R-CNN has achieved better average precision. Songmin Jia et.al[9][2019]proposed a new framework to overcome the problem of poor detection of small objects and lower system robustness. It efficiently integrated the global feature depth model and image texture. Yi-Qi Huang et al[16][2020] proposed to overcome the YOLO-v3 algorithm methods and optimized the new algorithm to increased the 3.86% of detection accuracy rate.

II.CURRENT SCENARIO

In today's scenario, the object detection technologies follow the fastest algorithms such as Fast R-CNN, Faster R-CNN, Histogram Oriented Gradients(HOG),
Region-based Convolutional Neural Networks(R-CNN), Region-Based Fully Convolutional Neural Networks(R-FCN), Single Shot Detector(SSD), Spatial Pyramid Pooling(SPP-Net), and YOLO.
The fast detection in a single image uses a single layer of Convolutional network by using a Single-shot Multi-box detection algorithm. Similarly, it uses the multi-layer convolutional network to detect multiple images by using YOLO versions algorithms.
In Yolo algorithms versions such as YOLO-V1, YOLO-V2, and YOLO-V3 professionally detect the objects in a real-time environment in various applications like Vehicle detection, Face detection, vehicle type detection, License Number plate detection, etc. Now a day, the versions of YOLO algorithms are more popular in vehicle detection in the traffic surveillance system.

(a) Faster R-CNN:

The Object Detection approach introduces the Faster R-CNN algorithms and it performs better and faster in the Detection and prediction of objects. Also, infixes the object score using the RPN network.

Ross Girshick et al [6] [2014] proposed an object detection model which has been used in Region selection. It provides a scalable object detection algorithm which is also simple and gives a 30% relative improvement over the previous results on PASCAL VOC 2012 giving an mAP of 53.8% accuracy.

Longsheng Fu et al[7][2018] proposed and implemented the Faster R-CNN-based kiwifruit recognition model which was developed and evaluated by images using the ZF-Net framework. The results obtained a good overall detection accuracy of 92.3% in the kiwifruit images that captured in the daytime.

Yongzheng Xu[8][2017] proposed the framework of Faster R-CNN for car detection from low altitude UAV images captured over the traffic signals. It evaluated the illumination changes in-plane rotation and detection speed in the constant frame and it achieved 96.4% completeness and 98.43% correctness with a real-time Detection speed of 2.10 frame/second.

Chen Linkai[15][2018] implemented to improved convolutional network and fast detection of the vehicle in highway, it uses the new techniques in feature concatenation to Faster-R-CNN and SSD.

(b) ***SINGLE-SHOT MULTI-BOX DETECTOR (SSD):***

Wei Liu et al [9][2016] proposed the Single-Shot Multi-box detector which is the first-stage detectors to achieve an accuracy of 74.3% on the pascal VOC 2007 dataset at 59 FPS. It works on the various dataset and it achieves better results which include timing and accuracy analysis on models with varying input sizes evaluated on PASCAL VOC, COCO, and ILSVRC.

The following table shows the results of SSD 300 and SSD 512 architecture models which are trained on the coco, Pascal VOC 2007, and Pascal VOC 2012 dataset with new data augmentation, and it significantly detects small objects.

Method	VOC2007 test		VOC2012 test		COCO test-dev2015		
	07+12	07+12+COCO	07++12	07++12+COCO	trainval35k		
	0.5	0.5	0.5	0.5	0.5:0.95	0.5	0.75
SSD300	74.3	79.6	72.4	77.5	23.2	41.2	23.4
SSD512	76.8	81.6	74.9	80.0	26.8	46.5	27.8
SSD300*	77.2	81.2	75.8	79.3	25.1	43.1	25.8
SSD512*	**79.8**	**83.2**	**78.5**	**82.2**	**28.8**	**48.5**	**30.3**

Table : 1 shows , Results on Pascal voc2007,2012, COCO datasets. [9]

Songmin Jia et al[10][2019] proposed the deep learning algorithm using the SSD method to try to detect the small objects and improve the robustness of object accuracy. This study introduces the shallow layer with high resolution and strong structure and thus the detection of small-scale objects were effectively improved.

.

Kyoungmin Lee et al [11][2018] proposed a method to represent power increase in feature maps using Resblock and Deconvolution layers, by creating multi-scale feature maps with 3-way residual features and applied a unified prediction module to generate the output, which increased the power of prediction. It achieves a higher score than SSD on Pascal Voc and MS coco datasets.

(c) YOLO:

Joseph Redmon [11] [2016] proposed a unified model for object detection and constructed the trained image to predict bounding boxes and class probabilities, optimizing end-to-end detection performance. This method makes more localization errors and a low level of prediction in false positives on the background images and processes images in real-time at 45 frames per second.
Hendry et al [14][2018] proposed a modified tiny-YOLO in single class detector using the YOLO's 7 convolution layers and sliding window to detect and fast predict the car license plate utilized by the Yolo-darknet framework. The sliding window method works for all classes, but it struggles in small object detection.

(d) YOLO-V2:

Joseph Redmon et al [12][2016] proposed YOLO-V2 with various improvements in the YOLO detection method, referred to as a single-stage real-time environment. The author concludes, the YOLO v2, which is an improved model, is well-trained jointly in optimizing detection and classifications of images in PASCAL VOC and COCO datasets. And achieves a detection accuracy of 67 Fps with 76.8 mAP on VOC 2007.

To improve the YOLO-v2, the author used various techniques such as Batch normalization, High-resolution Classifier, Convolutional networks with anchor boxes, Dimensional cluster boxes, Direct location prediction, Fine-Grained techniques, etc.

The original input resolution of 448 x 448 was changed to 416 x 416 with the addition of anchor boxes on the fly and the modified YOLO predicts detection on a 13 x13 feature map.

In multi-scale training, in YOLO-v2 in the Low-resolution detection output has detected 288 x 288 image at 91 Fps with 69 mAP and in the High-resolution detection, output has detected 544 x 544 image at 40Fps with 78.6 mAP.

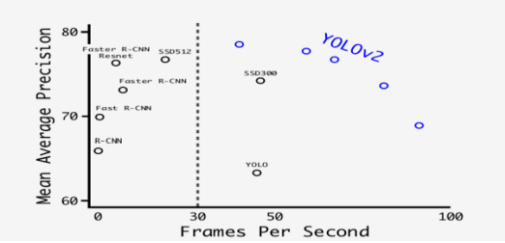


Figure: 1. Graph shows Accuracy and Speed on VOC 2007

In Figure: 2 show the speed and Accuracy on Pascal Voc 2007 dataset performing in various type of algorithms and it gets the better mAP.

(e) YOLO-V3:

YOLO- v3 is having balanced features like the highest speed, accuracy, and implementation intricacy as well. And it has a good powerful feature extraction capability along with good accuracy and faster detection. But in small-sized objects, its performance is somewhat less.

YOLO v3 has success in real-time object detection with accurately and classifying the objects in different scales. In the following important attributes in predictions and detection

1. *Bounding box predictions*
2. *Class Predictions*
3. *Feature Pyramid Network*

i) Bounding Box predictions

YOLO v3 gives the score for the objects for each bounding box. It uses logistic regression to predict the objectiveness score.

The network predicts the 4 coordinates for each bounding box such as tx, ty, tw, th, where tx, ty is the co-ordinate of starting point in boundary box and tw, th represents width and height of boundary box. And the top left corner coordinates of the image cell is referred to as, cx, cy and the Prediction of boundary box is Pw, Ph.

$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + c_y \qquad (1)$$
$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h} \qquad (2)$$

Whereas, eqn (1) and eqn (2) shows a bounding box is represented by a 6 number of attributes such as, $p_c, b_x, b_y, b_h, b_w, c$. The network expanded the confidence of bounding box into dimensional vector.

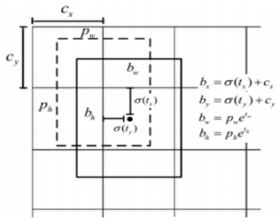


Figure 2: Boundary value calculation with confidence

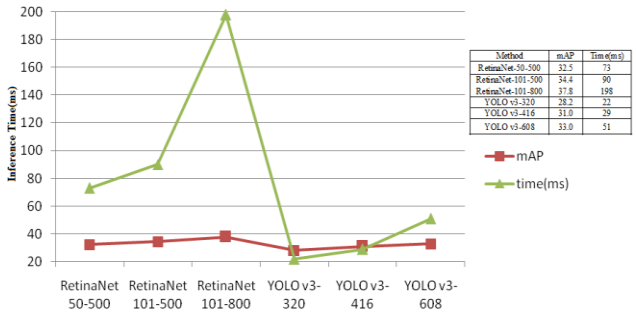
ii) Class prediction:

In YOLO v3 utilizes strategic classifiers for each class rather than softmax. YOLO v3 we can have multi-mark characterization. The location head is a multi-scale identification head thus, we would have to separate highlights at numerous scales also.

iii) Feature pyramid Networks(FPN):

The FPN takes the multiple feature mapping in an image. It is mainly used to improve the better resolution of an image. Also helps to detect the image in small, medium and large size. It has connected in between conventional neural network layers and original feature maps of an image. It obtained to improve the object localization.

Yolo-v3 achieves 33.0 mAP in small images, 57.9 mAP in medium images, and 34.4 mAP in large images prediction in 51 ms on Titan x and RetinaNet it achieves 57.5 mAP in medium images in 198 ms Which shows its better performance 3.8 times faster.



Method	mAP	Time(ms)
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198
YOLO v3-320	28.2	22
YOLO v3-416	31.0	29
YOLO v3-608	33.0	51

Figure: 3 Graph shows YOLO-v3 performance

Figure: 3 shows the highest mAP of 37.8 is achieved by RetinaNet-101-800 in 198ms and the next highest mAP of 34.4 is achieved by RetinaNet-101-500 in 90ms. The YOLO -V3-608 achieves mAP of 33.0 in 51ms.

Even though RetinaNet-101-800 has scored more mAP(37.8) it took 198ms. But RetinaNet-101-500 achieved a mAP of 34.4 in 90ms, whereas YOLO-v3-608 achieved a more or less equal mAP of 33.0 in 51ms. Hence, when more accuracy is solicited it takes more time. But by compromising a little bit on the mAP the YOLO-V3 achieves it in 51ms. Hence, it can be seen that YOLO-v3 strongly runs faster than other detection methods and performs better.YOLO-v3 works best in Single-stage detectors and produce good accuracy in traditional object detection model.

III. Fine tuned of YOLO versions

Algorithms	***Improvements***	***Limitations***	***Datasets with accuracy***
YOLO –V1	*Speed is improved in real time detection at 45 frames per second* *Use of Fast YOLO network to increase the speed at 155 frames per second.*	*YOLO can detect only 49 objects using in 7 X 7 grids.* *YOLO can predicts, only one class at each grid and limits the number of nearby objects* *Moderately High in localization errors occur in small objects*	*Pascal voc 2007+2012 dataset with 52.7 accuracy using 155 frames per second*
YOLO –V2	*Furthermore, it can be run at a variety of image sizes to provide a smooth tradeoff between speed and accuracy.*	*It achieves classification loss, back propagate loss at the time of labeling the image.* *The localization error is reduced, but*	*COCO dataset and Image classification dataset with 68.7 mAP* *Pascal Voc dataset with 78.6 mAP.*

		not accurate. Most weak in image segmentation.	
YOLO –V3	*It has increased more improvement in real time object detection with speed, accuracy and classifying the objects. The Prediction of the object at aspect ratios or different scales. Easily to detect the small object in good accuracy rate. With the increase in MAP, there was a decrease in the localization errors.*	*It has most struggles with small objects that appear in groups of images It gets Low mAP in different categories of object. Minimum localization errors.*	*COCO dataset with 57.9 mAP*

Proposed Smart-Region Detection Method:

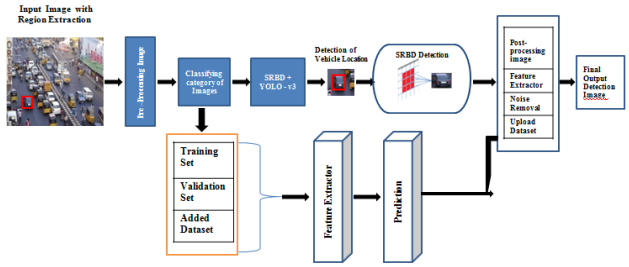


Figure: 3 Smart-Region Detection Method

IV. Methodology

The proposed method *captures the image from streaming videos and preprocesses it. The architecture classifies the various sizes of images into the small medium and large objects and then subjects them to training and validation process whereas new images are added to dataset.*

Overall pipeline architecture of proposed Smart-Region Detection method:

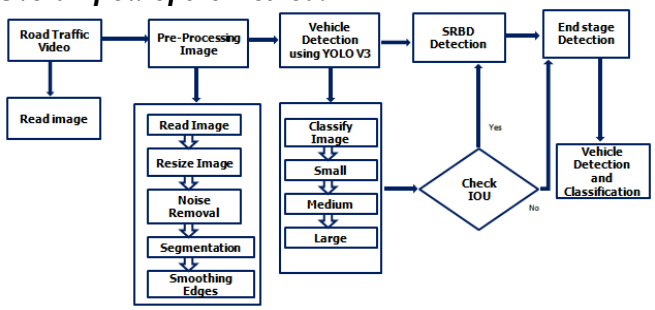
Steps: 1 In the initial stage capturing the image from streaming videos

Steps 2: An initial stage of data pre-processing model is created and finds the corrections in image annotation, when the image is resizing and updating the new images.

Steps 3: Next stage of Selecting and classifying Image objects is started with the new layer of CNN network and Image objects are Classified into small, medium and large objects.

Steps 4: Each Region of Interest is detected and the selection is entered into the grid cells with varying the sizes of images and the location is culled out.
Steps 5: To analyze the feature extraction of the detected image and to check the condition in the layer.
Steps 6: The Image data is trained and validated using the image data in dataset.
Steps 7: Available images from the datasets are matched and predicted. Otherwise repeat the step 4 and step 5.
Steps 8: New images are labeled and added into the dataset.
Step 9: Final detection of the results with classification value and localization value of image data is displayed.

Overall flow of the method:



It yielded good results and higher performances of vehicle detection using vehicle dataset and it had overcome the problem of small, medium and large object detections, which was achieved by combining YOLO-v3 algorithm with our proposed SRBD algorithm. Since YOLO-v3 struggles with small object detection, the amalgamation of YOLO-v3 and SRBD algorithm which gives better results than other deep learning methods.

Conclusion:

This paper analyses various technologies present in object detection algorithms. Object detection based on the state-of-the-art methods has some limitations in detection, which corresponds to various sizes of images in the dataset. The proposed Smart-Region-based Detection Method preprocesses, classifies, and validates the images obtained from the ROI extractor and adds the new images to the dataset after labeling them. This Smart-Region-based Detection method works on all sizes of images irrespective of the underlying architecture. Our proposed method minimizes the detection error, tries to detect small objects, and runs faster than other detection methods by providing good performance in streaming videos and images.

References

[1] Rautaray, Siddharth S., and Anupam Agrawal. "Vision-basedhand gesture recognition for human-computer interaction: a survey." Artificial Intelligence Review 43, no. 1 (2015): 1-54. Doi: 10.1007/s10462-012-9356-9 1
[2] Liang Zheng, Yi Yang, and Qi Tian, Fellow, IEEE "SIFT Meets CNN: A Decade Survey of Instance Retrieval", arXiv:1608.01807v2 [cs.CV] 23 May 2017. 2
[3] Ankita.S. Kulkarni and Sagar. B. Shinde "A Review Paper on Monitoring Driver Distraction in Real Time using Computer Vision System", 2017 International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE2017),2017. 3
*[4]Mingyu Li * ID and Koichi Hashimoto ID, "Accurate Object Pose Estimation Using Depth Only", **MDPI and ACS Style** Li, M.; Hashimoto, K. Accurate Object Pose Estimation Using Depth Only. Sensors **2018**, 18, 1045. 5*
[5] Ross Girshick Jeff Donahue Trevor Darrell and Jitendra Malik," Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5)", arXiv:1311.2524v5 [cs.CV] 22 Oct 2014 6
[6] LongshengFu, YaliFeng, YaqoobMajeed, XinZhang, JingZhang, ManojKarkee and QinZhang," Kiwifruit detection in field images using Faster R-CNN with ZFNet", in IFAC conference paper(IFAC-PapersOnLine), 2018. 7

[7]Yongzheng Xu, Guizhen Yu, Yunpeng Wang, Xinkai Wu, and Yalong Ma,"Car Detection from Low-Altitude UAV Imagery with the Faster R-CNN", in Journal of Advanced Transportation , Volume 2017, Article ID 2823617,https://doi.org/10.1155/2017/2823617 8

[8] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu1 and Alexander C. Berg1, "Single-Shot Multibox Detector "in arXiv:1512.02325v5 [cs.CV] 29 Dec 2016 9

[9] Songmin Jia, Chentao Diao, Guoliang Zhang, Ao Dun, Yanjun Sun, Xiuzhi Li, and Xiangyin Zhang1,2
"Object Detection Based on the Improved Single Shot MultiBox Detector" in IOP Conf. Series: Journal of Physics: Conf. Series 1187 (2019), doi:10.1088/1742-6596/1187/4/042041 10

[10] Kyoungmin Lee, Jaeseok Choi, Jisoo Jeong and Nojun Kwak, "Residual Features and Unified Prediction Network for Single Stage Detection", in arXiv:1707.05031v4 [cs.CV] 5 Jan 2018 11

[11]Joseph Redmon , Santosh Divvala, Ross Girshick, Ali Farhadi "You Only Look Once: Unified, Real-Time Object Detection" in Computer vision and pattern Recognition, 2016, https://arxiv.org/abs/1506.02640 13

[12]Joseph Redmon, Ali Farhadi," YOLO9000: Better, Faster, Stronger", in ***Computer Vision and Pattern Recognition,*** *25 Dec 2016,https:// arXiv:1612.08242 14*

[13]Joseph Redmon, Ali Farhadi,"YOLOv3" An Incremental Improvement", in ***Computer Vision and Pattern Recognition,*** *8 Apr 2018, https://arxiv.org/abs/1804.02767. 15*

[14] Hendry, Rung-Ching Chen," A New Method for License Plate Character Detection and Recognition", in ICIT 2018: Proceedings of the 6th International Conference on Information Technology: IoT and Smart CityDecember 2018 Pages 204–208 https://doi.org/10.1145/3301551.3301592

[15] Chen, Linkai Ye, Feiyue, Ruan, Yaduan, Fan, Honghui, Chen, Qimei,"An algorithm for highway vehicle detection based on convolutional neural network", EURASIP Journal on Image and Video Processing
Proc. 2018, 109 (2018). https://doi.org/10.1186/s13640-018-0350-2

[16] Yi-Qi Huang, Jia-Chun Zheng, Shi-Dan Sun, Cheng-Fu Yang, and Jing Liu," Optimized YOLOv3 Algorithm and Its Application in Traffic Flow Detections" April 2020 pages 1-15.