A Comparative Study of Object Detection Algorithms in A Scene

Prince Kumar[1], Vaibhav Garg[2], Pavan Somvanshi[3], Pathanjali C[4]
Dept. of CSE, BNMIT
Bengaluru, India
[4]Assistant Professor, Dept. of CSE, BNMIT

***Abstract*—Object detection is a major field of interest in the domain of Computer Science, Computer Vision, Artificial Intelligence and Machine Learning. It plays a very important role in many kinds of systems and products which are extensively used by people around the globe. Some of the applications are in home automation, self-driving vehicles, shopping complexes, traffic monitoring, video surveillance, sports, manufacturing, robotics, and many others. As this plays a very important role in these applications it has to be accurate and fast with a very less margin for errors in the least amount of resources possible. In this paper we are comparing few of the image detection algorithms for our project which is to develop a system for visually impaired people to help them in their daily activities and make them independent.** *(Abstract)*

Keywords—Object Detection, R-CNN, Fast R-CNN, Faster R-CNN, YOLO

I. INTRODUCTION

Object detection is a major field of computer vision. It deals with detecting instances and classification of semantic objects of a certain class in images and videos. It has various applications such as face detection and recognition, biometric detection, medical imaging, pedestrian detection, optical character recognition. Object detection works by matching features from the test subject to the features extracted from the training data. Based on the result of object detection the objects are classified and appropriate objects are shown as output.

However, detection of multiple objects within a single frame is a challenging task. The limitation and uncertainty of the object classifiers are overcome by the convolutional neural network approaches. As a solution, many algorithms are proposed but the most prominent, robust, and accurate methods proposed over the time are Region based Convolutional Neural Networks (R-CNN), You Only Look Once (YOLO), (SSD), RetinaNet. These solutions take multiples regions of the frame for the feature extraction and perform the object detection over these regions.

The advantage of neural network methodology is that the model is self-learning and the training process itself determines the best features from the data sets. The disadvantage of the neural network is that the training process is time consuming and computational intensive.

The main goal is to identify and detect multiple objects in a video frame. The multiple objects are identified and distinguished by the boundaries and the data set are used to classify them. Since multiple objects detention is a tedious and computational intensive process and hence smart methods are used on the CNN networks to improve the performance and accuracy.

The objects are stored in the database and are compared with the objects obtain from the regions of the video frame. The system identified the objects based on some threshold value example: 35% feature matching criteria and then the objects are identified from the database.

II. LITERATURE SURVEY

A. RCNN

Object detection system using RCNN [1] consist of three modules- generates category independent region proposals, large convolutional neural network and set of class specific linear SVMs. Different papers offer different methods for generating category-independent region proposals such as objectness, selective search [6], category independent object proposals etc., but selective search is widely used as it enable controlled comparison with prior detection work. Selective search run on test image to extract around 2000 region proposals which are wrapped and forward propagated through the CNN to find the features.

B. FAST RCNN

It is a successor of RCNN. FAST RCNN [2] takes an entire image and a set of (around 2000) object proposals as input which are passes through multiple convolutional and max pooling layers to give a featured map. For each test ROI (region of interest), it extracts a featured vector from featured map is given as input to final fully connected layers and outputs a class containing the probability distribution and set of predicted boundaries. A unsophisticated approach to solve this problem is to take distinct regions of interest from the image, and use a CNN to segregate the presence of the objects within that region.

C. FASTER RCNN

It was proposed by Ross B. Girshick [3] in the year 2016. The feature extraction, proposal extraction and rectification are integrated in a network in faster RCNN. It creatively uses the convolutional network and share it with Object detection network which reduces the proposed frame to 300 instead of 2000, Hence the performance is greatly increased.

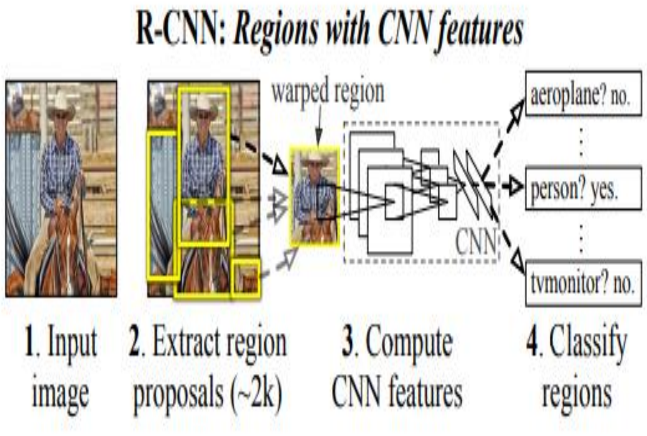

Figure 1: OBJECT DETECTION OVERVIEW WITH RCNN

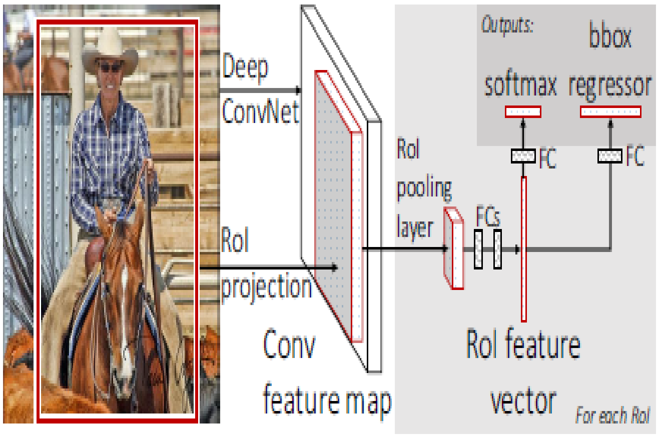

Figure 2: OBJECT DETECTION OVERVIEW WITH FAST RCNN

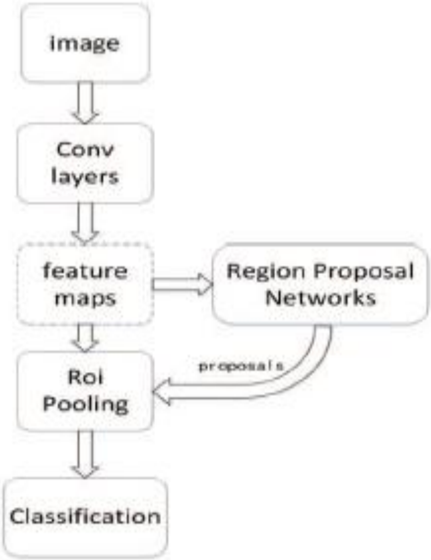

Figure 3: FASTER RCNN STRUCTURE

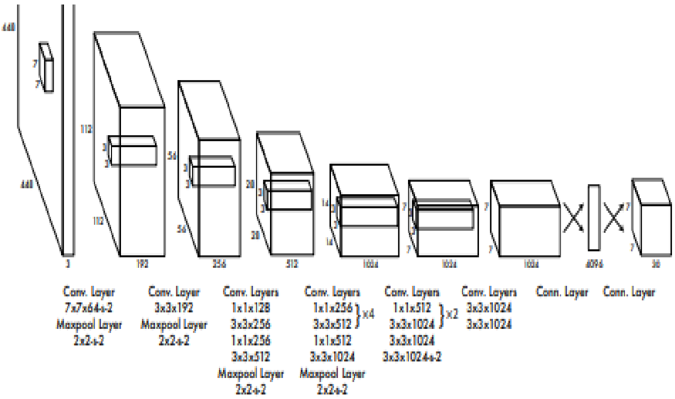

Figure 4: YOLO STRUCTURE

D. *YOLO*

YOLO [4] acronyms for You Only Look Once. It is proposed in the year 2016. Here instead of using the classifier to perform the detection the whole process is done in one network. A single neural network determines the bounding boxes and class probabilities in one analysis, which brings its performance to a whole new level. Since whole process is a single pipeline it can be optimized further.

III. RESULT ANALYSIS

The result analysis of a system or algorithm is based upon some set of parameters. Most common parameters are performance, time taken, resources needed, accuracy etc. which are undertaken in almost all analysis. Where the performance is parameter indicating how well the algorithm does perform. Time taken is the parameter which represents the time taken by the algorithm to output the result. Resources needed are defined as the amount of resources required by the algorithm. Accuracy defined the promising factor of the algorithm which is the percentage of the correct output generated by the algorithm.

On applying the general parameters over the R-CNN method of the object detection proposed by the Ross Girshick, the results show that it is much faster than the old methods based on the classification methods. Instead of huge number of regions, RCNN use the selected search to extract just 2000 regions per image. So the feature extraction will run over only 2000 regions. After such drastic reduction in the computation R-CNN still have its limitation. Firstly, it still takes lot of time to train the network as it classifies 2000 regions for each image. Secondly, it is not applicable in real time as it takes around 40 seconds for each test image. And lastly, since it uses selective search algorithm which is fixed, it cannot learn from the experiences.

A new version of R-CNN is proposed later proposed by the Ross Girshick to solve the drawbacks of the R-CNN and it is called as Fast R-CNN. The implementation of this method is different from the R-CNN implementation as the input image is fed to the CNN instead of the region proposals. The CNN generate a convolutional feature map which is then used to identify the proposal region in the image. A feature vector is generated from the feature map for each object and on this vector; softmax layer is used to anticipate the class of the proposed region. This method is far better than R-CNN as we don’t have to feed 2000 region proposals to CNN every time. Instead the CNN operation is done once per image.

Similar to the R-CNN and Fast R-CNN another method is proposed. The implementation of the method is similar to the previous methods but instead of selective search algorithm, an independent network is used to anticipate the proposed regions. These proposed regions are reshaped using the ROI polling layer and then used to segregate and identify the classes and boundaries. This method is much faster than Fast R-CNN as an independent network is used to anticipate the proposed region instead of a fixed algorithm.

A new method You Only Look Once (YOLO) is proposed for the recognition of the objects. As the above methods use proposed regions to identify the object in the image, it actually never considers the full image. Rather the regions with high probability of having the objects are passed in the system for the object detection. But in YOLO, it has only one Convolutional network and the whole image is analysed by this network. It divides the image into SxS grid and take m bounding boxes. For each box the network outputs a class probability and the classes with chance higher than the threshold value are used to locate the object. This method has many advantages due to its single convolutional neural network. Firstly, it predicts bounding boxes and class probability directly from the whole image in one evaluation. Secondly, the whole detection process is done in a single network; hence it is easy to optimise the network. It is much faster than the RCNN, Fast RCNN and Faster RCNN as it has only one convolutional neural network.

Table 1: Comparison result of different models(R-CNN, Fast-RCNN Faster-RCNN, YOLO)

Model	**Latency**	**mAP**	**FPS**	**Real Time**
R-CNN	High	~60	<1	NO
Fast R-CNN	Medium	~70	<1	NO
Faster R-CNN	Medium	~70	7	NO
YOLO	Low	~60	46	YES

The table 1 shows the comparison of different models with respect to latency, mean Average Precision (mAP), Frames Per Second (FPS), and whether they can be used for real time applications or not. The above table clearly shows that YOLO is better than the R-CNN based algorithms having low latency and higher FPS. It can be clearly seen that to gain this speed a trade-off has been made in precision. Even after having low mAP, YOLO has acceptable mAP to be able to be used for real time applications and when taken together with the high FPS and latency, it becomes clear it is the best algorithms in its class.

IV. CONCLUSION

As, the speed and accuracy are an important parameter in object detection applications, it is necessary to have less computation time and also process the input fast to provide the result to the user. YOLO is best for real time object detection as it has only one convolutional network and with fast computational speed it can deliver the result faster as compared to the other methods and the accuracy of the method can be managed as per the requirement of the system.

REFERENCE

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, UC Berkeley, "Rich feature hierarchies for accurate object detection and semantic segmentationTech report (v5)", 22 OCT 2014.

[2] Ross Girshick, "Fast R-CNN", IEEE International Conference on Computer Vision, 2015.

[3] BIN LIU, Wencang ZHAO and Qiaoqiao SUN, "Study of Object Detection Based On Faster R-CNN”, Chinese Automation Congress (CAC), 2017.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] Shaoqing Ren Kaiming He Ross Girshick Jian Sun,” Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017

[6] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition", IJCV, 2013.

[7] Joseph Redmon, Ali Farhadi,"YOLO9000: Better, Faster, Stronger", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

[8] Joseph Redmon, Ali Farhadi,"Yolov3: An incremental improvement", arXiv preprint arXiv:1804.02767, 2018