Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

A Study on Real Time Object Detection using Deep Learning

Pradyuman Tomar
Dept. of Electronics and Communication Engineering
Meerut Institute of Engineering and technology
Meerut, India

Sagar
Dept. of Electronics and Communication Engineering
Meerut Institute of Engineering and technology
Meerut, India

Sameer Haider
Dept. of Electronics and Communication Engineering
Meerut Institute of Engineering and technology
Meerut, India

Abstract **— Object Detection is very closely connected with the Field of Computer Vision. Object detection empowers recognizing instance of different objects in images and videos or video recordings. It identifies the different characteristics of Images rather than object detection techniques and produces an intelligent and effective understanding of pictures very much like human vision works. In this paper, We will starts with the concise presentation of introduction of deep learning and famous object detection system like CNN(Convolutional Neural Network), R-CNN, RNN(Recurrent brain network), Faster RNN, YOLO(You Only look once). Then, at that point, we center around our proposed object detection model architecture along for certain advancements and modifications. The conventional model recognizes a little object in pictures. Our proposed model gives the right outcome with precision.**

Keywords — YOLOv4, CNN, Real-time object detection, Deep Learning, **RNN.**

I . INTRODUCTION

Although the human eye can instantaneously and exactly recognize a given visual, including its content, location, & nearby visuals by interacting with it, computer vision-enabled robotic systems are sometimes and somehow to slow and inaccurate. Any developments in this field will lead to increased efficiency and performance may open the way of more intelligent systems, similar to humans. As a result, systems such as advanced technology, which allow humans to accomplish tasks with little to no conscious thought, will definitely make our life lot easier.

For example, even if the driver is not aware of their activities, driving a car equipped with computer vision enabled assistive technology could foresee and notify a driving crash before to the incidence. As a result, real-time object identification has become a critical component in continuing to automate or replace human operations. Computer vision and object detection are very important and crucial fields in machine learning, and they are expected to help to unleash the hidden potential of general-purpose robotic systems in the future.

With the ongoing innovation in current technology, making transparency and feasibility of information to and from everybody associated with it has turned into a simple errand. Most of the humans have standard PCs (laptops), and cell phones, made this global expansion significantly more open. Alongside this internet globalization, the development of information, data and pictures accessible on the web/cloud has become to the mark of millions every day. Use of electronic devices to use this data and make important acknowledgments and cycles is indispensable because of people's difficulty performing same iterative assignments or tasks. The underlying advance of most such cycles might incorporate perceiving a particular article or region on a picture. Because of the unconventionality of the accessibility, area, size, or state of a thing in each picture, the acknowledgment interaction is incomprehensibly hard to be performed through a conventional modified PC calculation.

Deep learning is essential for ML. An excessive number of Methods have been proposed for object detection. Methods and Techniques of object detection comes under deep learning. Object Detection is an important field of Machine Learning and broadly utilized in Computer vision. Deep learning has been becoming well known beginning around 2006.

Various Techniques have been proposed to tackle the issue of Object Identification over time. These techniques center around the solutions through different stages. Specifically, these center stages incorporate recognition, classification, localization, and object detection. Alongside the advancement in present technology throughout the long term, these Techniques have been confronting difficulties, for example, output accuracy, resource cost, processing speed and complexity issues. With the creation of the main Convolutional Neural Network(CNN) algorithms during the 1990s roused by Yann LeCun et al. [1] and very important research and innovations like AlexNet [2], CNN algorithms have been fit for giving answers for the item recognition issue in various methodologies. With the goal of ease of human, improving accuracy and speed of recognition and detection, optimization focused algorithms are continuously being developed and improved over time, for example, Deep Residual Learning (ResNet) [5], VGGNet [3] and GoogLeNet [4] have been developed throughout the long term.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

However these Algorithms improved over the long time, window selection or recognizing various objects from a given picture or image was as yet an issue. To carry answers for this issue, algorithms having region proposals, crop/wrap feauture, bounding boxes regressions like Regions with CNN (R-CNN)SVM classification were presented. Despite the fact that R-CNN was very high in precision with the past innovations, its high utilization of existence later prompted the creation of Spatial Pyramid Pooling System (SPPNet)[6].
Regardless of SPPNet's speed, to remove the same problem it was imparted to R-CNN; Faster R-CNN was presented. However Faster R-CNN could arrive at ongoing paces utilizing exceptionally profound organizations, it held a computational bottleneck. Later Faster R-CNN, Algorithms, is heavily based on previous algorithm ResNet, was presented. Because of Faster R-CNN not yet fit for outperforming results, YOLO was presented. This paper will review You Only Look Once Algorithm for Object detection.

A. Abbreviations

Abbreviations used:
CNN – Convolutional Neural Network.
ResNet50 - Residual Neural Network (50 layers).
ResNet152 - Residual Neural Network (152 layers).
YOLO – You Only Look Once.
RNN – Recurrent Neural Network .
RCNN – Region Based CNN.

A. Overview of object detection(CNN).

Object Detection is a study of Computer Vision Field. Object location is a huge exploration region in Computer Vision, can be applied to numerous applications, for example, Driver less vehicles, security, reconnaissance, machine examination, and so forth. Object Detection is utilized to distinguish the area of the object in a picture, Face detection, medical imaging, etc.
Evolution of Deep learning have changed the old methods of object detection and tracking system. Computer Vision recognizes characteristics in pictures, Classifying Object in the picture, Classifying objects along with localization, drawing a bounding box around object Present in the picture, Object segmentation or semantic segmentation, Neural style Transfer. Deep learning strategies are the most grounded strategy for object detection.

II. LITERATURE SURVEY

Three are various methodologies has introduced by numerous researchers . An algorithms for the first face detector was concocted by Paul Viola and Michael Jones 2001. The face had identified and detected continuously on Webcam feed. It was developed out by Opencv and Face Detection. This couldn't distinguish some direction like up down, named, wearing a mask, and so on. Because of the advancement of Object detection in Deep Learning, it can be further classified into two model (1) Model based of region proposal; (2) Model based on regression/Classification.

A. Model-based on Region

a. **CNN**: This network was presented by Creators: Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton in 2012.
The network comprises of five convolutional layers. It accepts input as a picture which is a 2D array of a pixel with RGB channel. Then Channels or elements indicator apply to the information picture and get yield highlights maps. Numerous convolutional are acted in lined up by applying the ReLU work. CNN works for just a single object at a time so it doesn't work successfully in different objects in an image. CNN turned into a decent norm for image classification after Kriszhevsky's CNN's performance. We can't recognize objects which are overlapping and various background and don't order these various objects yet in addition don't distinguish boundries, contrasts and relations in other.

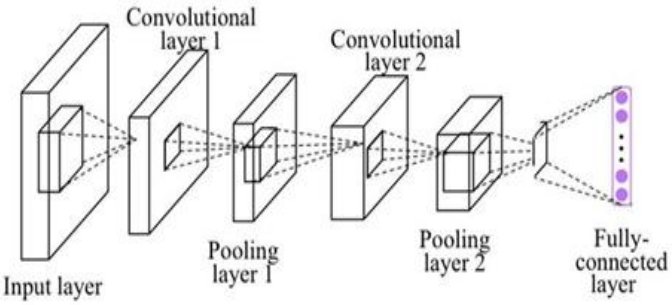


Figure 1. CNN layer diagram

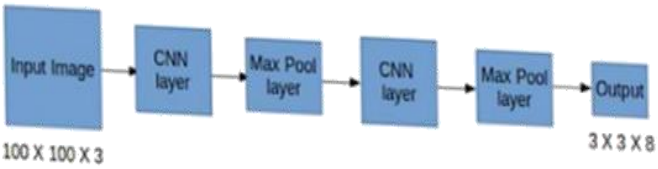


Figure 2. CNN Flow Diagram

b**. RCNN** : This network is presented by Creators: Ross Girshick, Jeff Donahue, Trevor in 2013this network motivated by overfeat. This network incorporates three principal parts, first is region extractor, second is feature extractor and last is classifier. It involves a selective search algorithms for object detection to create region proposal. Extricate 2000 small regions for each picture. Here 2000 convolutional networks utilized for every small regions of the pictures. So have one Convolutional network expected to handle RCNN different regions with CNN characteristics partitions the picture into a few regions. Run pictures through pre-prepared AlexNet lastly apply the SVM algorithm.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

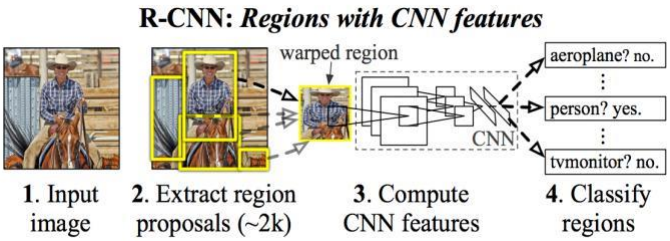


Figure 3. RCNN Flow Diagram

c. **Fast R-CNN**: This network is a superior adaptation of R-CNN which is presented by Ross Girshick. The article guarantees that Quick R-CNN multiple times quicker than past R-CNN which is nine times. Network select different sets /arrangements of bounding boxes then use feature extractor by CNN network then, at that point, use classifier or regression for yield the class of each containers.

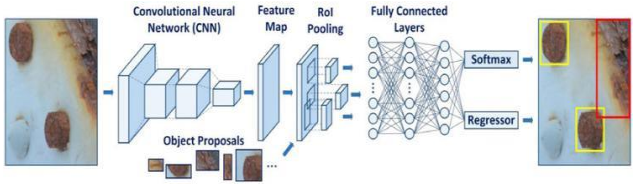


Figure 4. Fast RCNN Flow Diagram

d. **Faster R-CNN**: This is a better form of Faster RCNN which presented by Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun in 2015. Picture is given as input to a convolutional network that gives convolutional map. To recognize the different regions here the different network is utilized to foresee the region proposition.

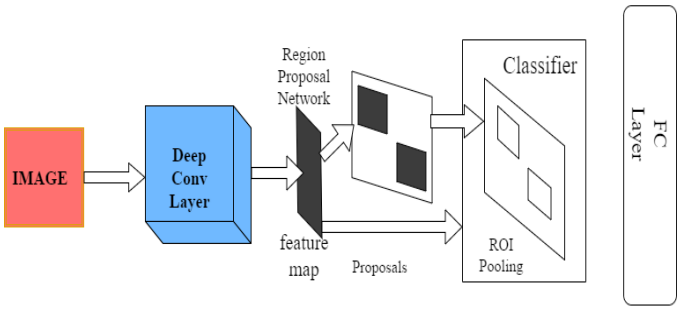


Figure 5. Faster RCNN Flow Diagram

B. Model based on regression/Classification.

a. **YOLO:** YOLO (You just check out once) at a picture to anticipate what are those object and where objects are available. A single convolutional network at the same time predicts numerous bounding boxes and class and probabilities for those crates. Regards detection as a relapse issue. Incredibly quick and precise YOLO takes a picture and split it into networks. Every lattice cell predicts just a single object. YOLO is very quick at test time and it requires single network assessment and performs feature extraction, bounding box predict, non max suppression, and contextual reasoning all simultaneously. Just go for it isn't pertinent for little items that shows up in gatherings like rushes of birds. Consequences be damned has a few variation like quick YOLO. Consequences be damned is something else altogether. It looks just once however in clear ways. Assuming a basic picture gives through the convolutional network in a single pass and comes out the opposite end as a 13×13×125 tensor portraying the bounding boxes for the framework cells. All you really want to do process then, at that point, is predict the last scores for the bounding boxes and discard the ones scoring lower than 30%.

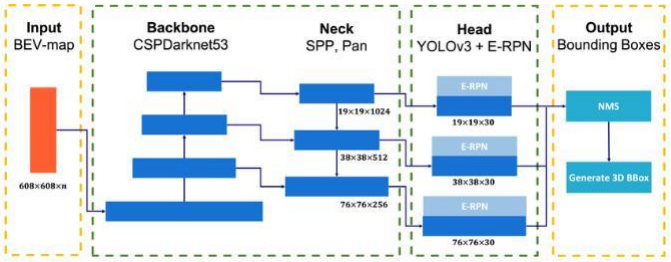


Figure 6. YOLO Network Arcchitecture

b. **SSD:** SSD (Single Shot MultiBox Detector) aim of classifications and localization are done in a single forward pass . The main benefit is quickness with relevel accuracy or, with great exactness. it runs a convolutional network on input pictures just a single time and processes a characterstic map. Histograms of Oriented Gradients are imagined by Navneet Dalal and Bill Triggs concocted in 2005. We need to take a glance at every pixel that straightforwardly encompassing it. Here contrast current pixel with each encompassing pixel. It flopped in more summed up object detection with commotion and interruptions behind the scenes or noise in the background.

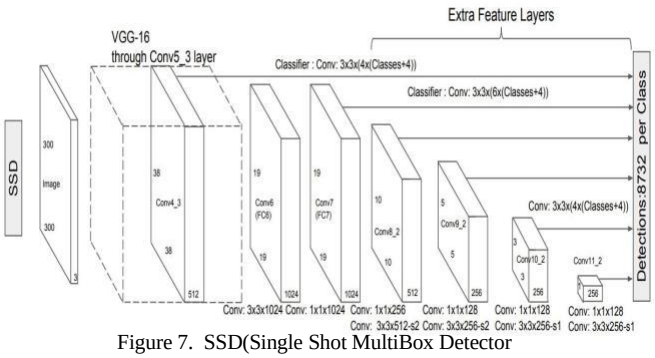


Figure 7. SSD(Single Shot MultiBox Detector

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

III. PROPOSED SYSTEM METHODOLOGY

A. PROBLEM STATEMENT

"To implement real - time object detection and recognition in an images captured by webcam and videos in dynamic environment using deep learning model."

B. PROBLEM ELABORATION

The primary goal is to detect and recognize Objects in Real-time. We require rich data, all things considered. We need to observe the different types of objects which are moving in respect to the camera. It will help us with perceiving and in recognizing different objects collaboration and interaction. We center around accuracy in this paper.

C. PROPOSED METHODOLOGY

This model incorporates include extractor with Darknet53 with features & highlight maps to upsampling and perform Concatenation on images . Proposed Model have different up-gradation for object detection techniques.

a. **Darknet 53:** This proposed framework utilizes a variation of Darknet which has initially 53 layers network and prepared on Imagenet or tested on Imagenet. For the detection more 53 layers are utilizing onto it, an absolutely of 106 layers of convolutional fundamental for proposed framework. This is the explanation the proposed framework turns out to be slow.

b. **Detection of three scales:**This model makes identification and detecting at three distinct scales. Here Detection is produced by applying 1 x 1 identification portions or kernels on highlight feature maps for three different sizes on three different places in the network 1 x 1(M x (5 +N)) is the state of the detection piece.

c. Here M is the quantity of bounding boxes on the component guide and N is the quantityof classes. Highlight Feature map produced by this kernel has a similar level and width of the past component map additionally distinguish ascribes alongside profundity. Three distinct scales are utilized. The principal identification and detection is made by the 82nd layer. The initial 61 layers of the picture are inspected by the network. On the off chance that we have a picture X416, the component guide will be of size 13 x 13. Identification is made by utilizing the 1 x 1 portion, and the resultant feature map will be 13 x 13 x 255. The subsequent identification is made by the 94th layer of the model and the resultant component guide will be 26 x 26 x255. Then, at that point, last identification and detection done by the 106th layer and producing Feature map size 52 x 52 x 255.

d. **Detecting smaller objects:** In this model three layer has distinct purposes, whereas 13 x 13 layer will detects objects which are larger , 52 x 52 layer is answerable for distinguishing more modest objects with the assistance of 26 x 26 layer recognize medium objects.

e. **Choice of anchor boxes :**This model total purposes 9 anchor boxes for the identification and detection of an object. We are utilizing k-means clustering to produce 9 anchors. For clustering orchestrate all anchors in decreasing manner as indicated by the aspects and relegate large anchors for the principal scales following three anchors for the subsequent scale, and the last three anchors for the third scale. This model predicts additional bouncing boxes. This model predicts boxes at 3 distinct scales, for the pictures of 416 *416, the quantity of predicted
Boxes are completed 10647 class prediction , softmax is not utilized. Independent logistic classifier and single binary cross entropy loss are utilized.

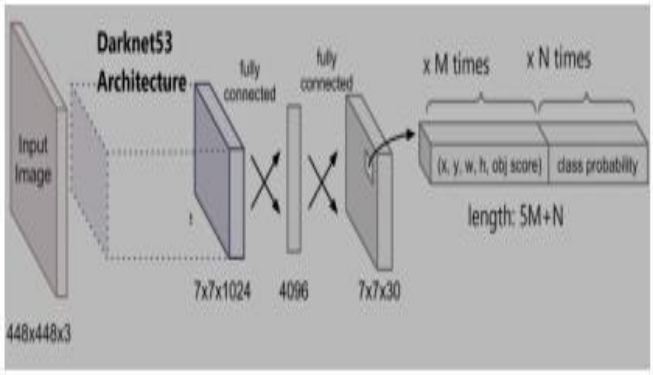
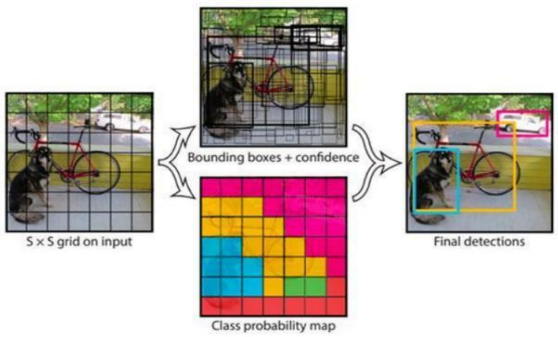


Figure 8. Prediction Model

D. PROPOSED FRAMEWORK

a. **YOLO Algorithm:**YOLO is Abbrevation of (you only look once). Older Object detection algorithms utilize the districts grid parts to distinguish and identifies the objects however don't utilize the whole picture , a few regions might contain the objects . YOLO is an object detection algorithm entirely different from the district based algorithms seen previously. In YOLO a convolutional network predicts the bounding boxes and the class probabilities for these containers. It's challenging for everybody to contain the assets for the Deep Learning So that is where this Yolo came into the picture. Furthermore, bunches of pre-prepared models and datasets are accessible at this point.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

Figure 9. YOLO Algorithm Process

YOLO stores the information in Vector Form:

YOLO = (pc, bx, by, bh, bw, c1, c2, c3),

Where pc characterizes the Probability and demonstrates in the event that object is available or not bx, by, bh, bw determines whether objects for the classes c1,c2,c3.

So on the off chance that there is any object concerning class c1, it will have the worth 1 generally 0.
It utilizes the non max suppression the bounding box with more exactness, precision is chosen and remaining are disregarded.

Equation for Non Max Suppression is :-

$$IoU = \frac{\text{Area of the crossing point or interaction}}{\text{Area of the association or union}}$$

Where , IoU = Intersection Of Union.

IV. DATASETS & PERFORMANCE COMPARISON AMONG VARIOUS ALGORITHMS:

The advancement of detection models is firmly connected with the blast of information volume. This is on the grounds that the performance test and algorithms assessment should be acquired through dataset, what's more, dataset is additionally a strong main impetus to advance the exploration field of detection.

Models	backbone	Size/pixel	Test	mAP%	fps
YOLOv1	VGG16	448*448	VOC 2007	67.2	46
SSD	VGG16	300*300	VOC 2007	78.1	47
YOLOV2	Darknet-19	544*544	V0C 2007	78.6	40
YOLOv3	Darknet-53	608*608	MS COCO	35	51
YOLOV4	CSP darknet-53	610*610	MS COCO	42.1	67.5
RCNN	VGG16	1000*600	VOC 2007	65	0.6
SPP-Net	ZF-5	1000*600	VOC 2007	55.4	-
Fast RCNN	VGG16	1000*600	VOC 2007	70.2	7
Faster RCNN	ResNet-101	1000*600	VOC 2007	76.5	6

Table 1- Comparison of object detection algorithm with performance

V. RESULT AND ANALYSIS WITH ACCURACY AND PERFORMANCE:

a. On MS COCO Dataset

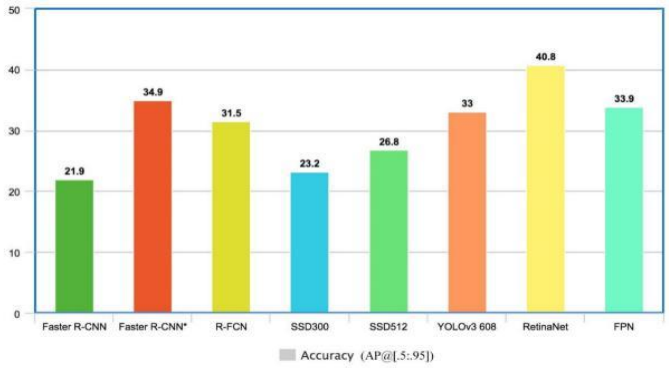


Figure 10. MS COCO Dataset Performance.

b. On PASCAL VOC 2007 & 2012

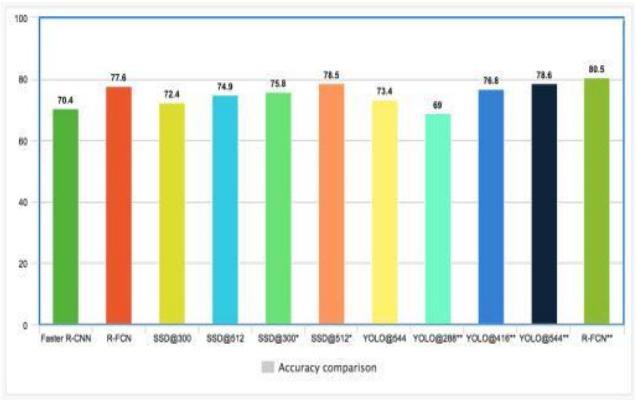


Figure 11. PASCAL VOC Dataset Performance.

c. Real-Time Detection : YOLO is a quick, precise object detection model, making it ideal for various application in the field of Computer Vision. We interface YOLO to a webcam and confirm that it keeps up with continuous execution in real-time.

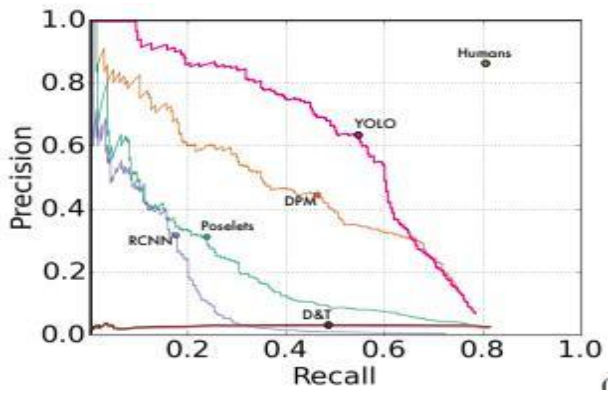


Figure 12. Picasso Dataset precision-recall curves

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

Model/dataset	VOC 2007 AP	Picasso AP	Picasso Best F1	People-Art AP
YOLOv4	59.5	53.4	0.595	45
DPM	43.2	37.9	0.460	35
RCNN	54.2	10.5	0.230	28
Poslets	36.5	17.9	0.275	
D&T	---	2.0	0.055	

Table 2. Results on VOC 2007, Picasso, & People-Art Dataset

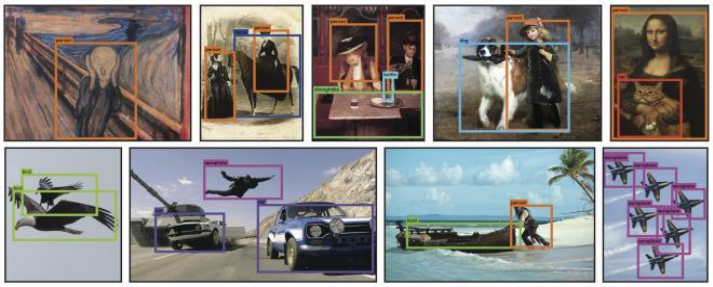


Figure 12. Qualitative sample Results

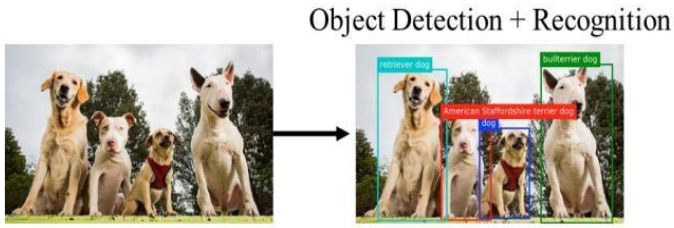


Figure 13. Test Image Detection and Identification

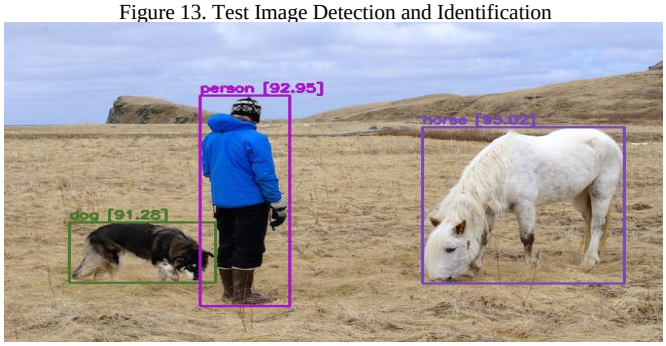


Figure 14 . YOLOv4 On Test Image

YOLO Algorithm running on Sample artwork from internet and natural images from the internet and website.

VI. CONCLUSION & FUTURE SCOPE:

As Object detection and recognition in today's world can be considered as one of the most challenging, complex and most important task in computer vision Field. As we know that this project is been developed with the underlying purpose of real-time object in pictures, videos captured streaming cameras. or web cams.

• Pre-processing methods proposed here .i.e. edge detection techniques to increase the contrast of the image which improve our model accuracy.

• It can be improve and innovate in the future by anybody without worrying about complexity.

• Future enhancements can be focused by implementing the project on the system having GPU for faster results and better accuracy.

• Like, for small object detection which is done by MS COCO in some face detection application and task. For improvement of localization of small objects under partial barrier. So that we will improve the network architecture with some modifications.

• By using that we can reduce the dependency of data network.

• For achieving efficient recognition of small objects with better accuracy.

So it is finally concluded that for enhance the accuracy and performance by using pre processing techniques like edge detection and increase image augmentation and contrast so that we get better results in output.

REFERENCES:

[1] T. Guo, J. Dong, H. Li, and Y. Gao, "Simple convolutional neural network on image classification," 2017 IEEE 2nd Int. Conf. Big Data Anal. ICBDA 2017, pp. 721– 724, 2017, doi: 10.1109/ICBDA.2017.8078730.

[2] J. Du, "Understanding of Object Detection Based on CNN Family and YOLO," J. Phys. Conf. Ser., vol. 1004, no. 1, 2018, doi: 10.1088/1742- 6596/1004/1/012029.

[3] Mauricio Menegaz, "Understanding YOLO – Hacker Noon," Hackernoon. 2018.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016, doi: 10.1109/CVPR.2016.91.

[5] Wu, R.B. Research on Application of Intelligent Video Surveillance and Face Recognition Technology in Prison Security. China Security Technology and Application. 2019,6: 16-19.

[6] Tian, J.X., Liu, G.C., Gu, S.S., Ju, Z.J., Liu, J.G., Gu, D.D. Research and Challenge of Deep Learning Methods for Medical Image Analysis. Acta Automatica Sinica,2018, 44: 401-424.

[7] Jiang, S.Z., Bai, X. Research status and development trend of industrial robot target recognition and intelligent detection technology. Guangxi Journal of Light Industry, 2020, 36: 65-66. [4] Krizhevsky, A., Sutskever, I., Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems,2012, 25: 1097-1105.

[8] Russakovsky, O., Deng, J., Su, H., et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision,2015, 115: 211-252.

[9] Girshick, R., Donahue, J., Darrel, T.,Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Computer Vision and Pattern Recognition. Columbus.2014, pp. 580-587.

[10] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015, 37: 1904-1916.

[11] Girshick, R. Fast R-CNN.In: Proceedings of the IEEE international conference on computer vision. Santiago.2015, pp. 1440-1448.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

[12] Ren, S.Q., He, K.M., Girshick, R., Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. Montreal.2016, pp. 91-99.
[13] Redmon, J., Divvala, S., Grishick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In: Computer Vision and Pattern Recognition. Las Vegas.2016, pp. 779- 788.
[14] Pushkar Shukla, Beena Rautela and Ankush Mittal, "A Computer Vision Framework for Automatic Description of Indian Monuments", 2017 13th International Conference on Signal Image Technology and Internet- Based Systems(SITIS), Jaipur, India, ISBN (e): 978-1-5386-4283-2, December-2015.
[15] M. Buric, M. Pobar and Ivasic-Kos, "Object Detection in Sports Videos", 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics(MIPRO), Opatija, Croatia, ISBN (e): 978-953-233-095-3, May-2018.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)