NVEO
Natural Volatiles & Essential Oils

Intelligent Object detection with classification and localization using Deep Learning

Akhilesh Kumar Srivastav[1], Azad Ali[2], Atif Khan[3], Danish[4], Himanshu Tripathi[5]
[1]*Associate Professor, Computer Science & Engineering Department*
[2,3,,4,5]*Computer Science & Engineering Department*
ABES Engineering College, Ghaziabad, U.P., 201009, India
akhilesh.srivastava@abes.ac.in[1]

Abstract:

Human eye is blessed to have an advantage of being enabled to differentiae and recognize the objects visually. Computer vision is enabling the same for the computer machines. Applications of human interaction with computer & object detection is enormous. The era of past few years has been of scientific achievements in the area of computer vision for the purpose of object detection. One of the primary goals of exisiting AI technology is the intelligent human computer interaction. Many Object Detection techniques were proposed in the earlier years, was presented as a summary in [1] [2]. Many a researchers have been shifting their trend to utilize diferent multifaceted classification and feature extraction methods just to enhance the correctness of system that does the object detectio. It has always been a challenging task to implement such system in real time applications. The earlier researches have reached to their threshold in the accuracy in the problems in the computer vision. Ever since the Deep learning technique took its way, the improvement in the accuracy of these problems is worth noticing. Image classification to predict the class of the images is among the major problems. Another one, a bit tougher problem is that of the Image Localization. Here in this problem, the images may contain the single object and the expected predicton is the class of the location of the object in the image (it expectes a box surrounding the object known as bounding box). In this article the problem of localization and classification both is taken into picture in the object detection. A bounding box for all the objects in the given image and the class of the object prediction technique has been proposed in the current article.

Keywords*: Bilinear Neural Network, Facial Recognition, Receiver Operating Characteristics, Convolutional Neural Network,Linear, Binary Pattern Histogram, Computer Vision Principal Component Analysis.*

I. INTRODUCTION

One of the major areas of Computer vision, the Object Detection has been researched interest area for a long time. There are many a scenarios where it is applied widely e.g. security using biometric feature, tagging of known faces in the photos automatically, etc. Many research laboratories and professional companies have been paying a lot of attention to this. In the past few years face detection has been researched widely. A person's face can be very well recognized without any support of manual exercise. In current article, authors have proposed a system to evaluate the dection of the known objects. It actually is the specific case of dection of object-class.

The earlier researches have reached to their threshold in the accuracy in the problems in the computer vision. Ever since the Deep learning technique took its way, the improvement in the accuracy of these problems is worth noticing. Image classification is among the major problems that intends to predict the class of the image. Image localization is another problem, which is a bit complex. In this the one of the object is being contained in the image and it is expected that the location class of the underlying image should be forecasted by the system (it expectes a box surrounding the object known as bounding box) given in Figure-1 e.g.

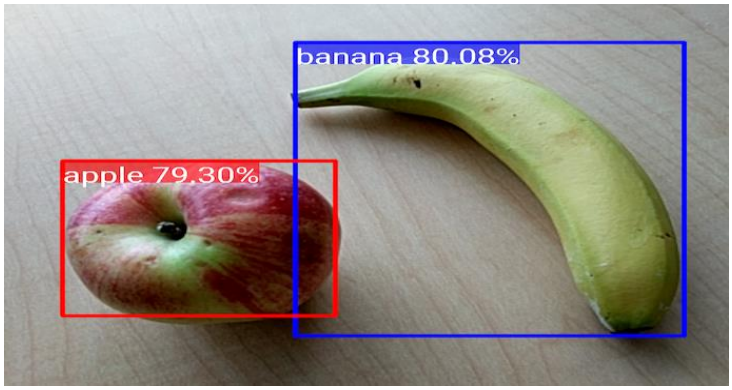


Fig.1. Object Detection in bounding box.

There are several motivatios for the object detection e.g.

- A system that ensures the secureity of citizens of the state/country. The places like Railway station, Borderline of control, and airports. Here the verification of the identity is extremely important. Facial recognition, one of the applications of object detection, has the ability of identify such faces to prevent any possible terror outfits and cases.
- In the search of the individuals with criminal records, surveillance cameras with the ability of face recognition can help the state machinery.

Objectives of the authors of this article are to build an object detection comprehensive system that is capable dealing with variety of the images alog with contant updations for improvement. The improvements are expected to be continuous and prediction capability needs to be better with the progress of the time. The proposed system is capable of working in the real time for the updations and improvement as the time for training such system is the key. Recogniton of images in the uncontrolled environment is a very tedious task. The authors of this article tried to use the existing research and carry out the improvements rather than starting everything from the scratch. This was undertaken to save the time of building the system from scratch and make the system feasible to work and produce quality results

Among the many known applications of object detection is the face detection used in cell phone cameras. A variety of the object is required to be identified in the independent driving, called multi-class application. Alongside, it has a significant role to play in surveillance systems. Integration of various tasks like estimation of pose after the object dection is the multi-stage obect detection. Similar applications of the same is tracking of objects that can be utilized in medical imaging and robotics as shown in Figure-2.

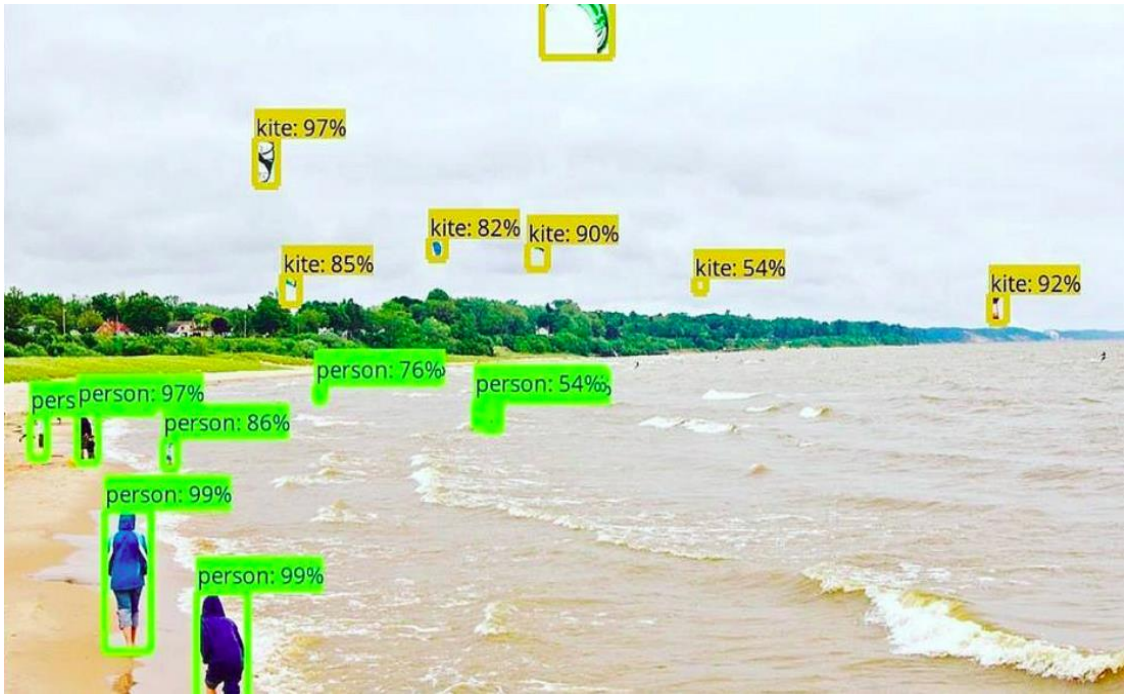


Fig.2. Various applications of Object Detection.

II Literature Survey

Object detection using Customary techniques of Computer vision has beed worked upon rigourously. Commonly used technique include Deformable part(DP) Model & sliding windows. These techniques have poor accuracy as compared to the Deep learning enabled methods. Out of the various deep learning enabled methods, the two class of methods which are dominant are: unified detection (Yolo[4], SSD [5]) & 2-stage detection(RCNN[1], Fast RCNN[2], Faster RCNN[3]). Theory behind these techniques are listed as under

A. Bounding Box

It is a surrounding rectangle that surrounds the underlying image that grips objects inside the image. There has to be a bounding box for each of the objects, even in the case of multiple instances of the same object in the image. A 4 parameter box is predcted for each object that includes (x-centrod, y-centroid, height and width). Training of the same can be done using actual and predicted bouding box's difference as a feedback. Jaccard distance is computed for the distance measured. Figure.3 explains the process.

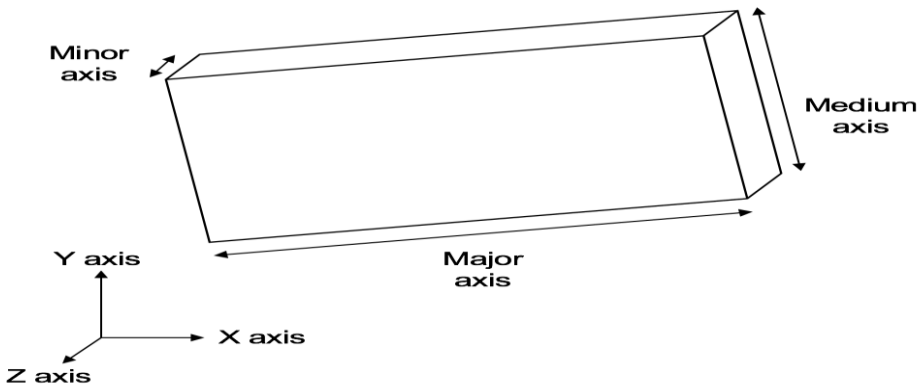


Fig.3. Bounding Box having x, y and z axes.

B. Regression & Classification

Regression is used for predicting the bounding box. Classification is used to predict the class inside the bounding box. Architecture of the same is shown in Figure 4

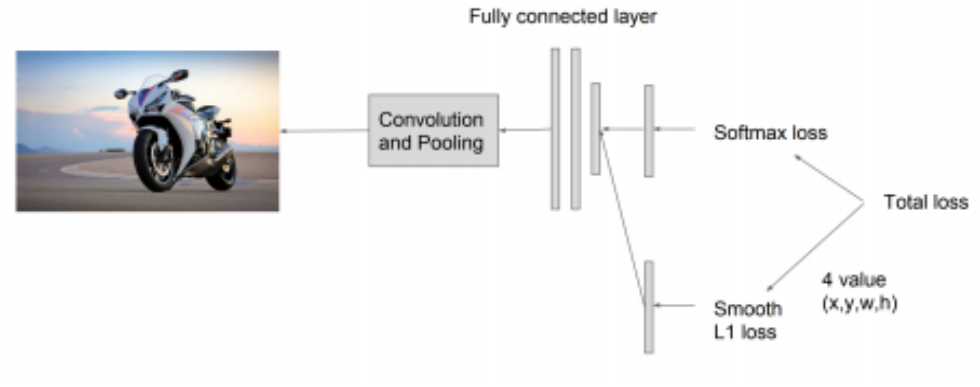


Fig.4. Classification and Regression of object

C. The 2-Stage Method

Firsty the proposals are extracted by making use of different computer vision technique. In the second step, for the classification network, they are resized to stationary I/P. This beheaves as the feature extractor. A SVM is then trained do the classification of object from background, for each class one SVM is used. Along with this a regressor for the bounding box is trained. This is done for the purpose of the corrections in the proposed box. Idea of the same has been presented in Figure 5. Accuracy rate of the methods are very good but lag in fps.

D. Unified Method

Major difference in this is rather than production of proposals, a set of pre-defined boxes are used to identify objects are used. Use of the convolutional features maps from post layers of the neural N/W guides additional N/W over the given feature map in order of prediction of offsets of bouding box and scores of class. Listed below are the steps that is required to be peroformed for the same:

- Firstly, train a Convolutional Neural N/W for the purpose of classification and regression.
- Then try to collect activation from post layers to conclude location and classification with convolutional layers.
- In the process of training, suggested the usage of jaccard distance in order to find the relation between the ground truth with that of the predictions.
- In the process of inference, usage of suppression other than maxima to extract manifold boxes nearby the same object is suggested.

III PROPOSED METHODOLOGY

The Methodology used for the object recognition is given as under

Object Detection → Feature Extraction → Data Comparison → Object Recognition

A. Object Detection: See the image and predict an Object in it.

B. Feature Extraction: characteristics which are unique in nature need to be extracted such that it can be differentiated from other similar things, like length, breadth, height etc.
C. Data Comparison: In spite of the variations in expression or light, comparion of these uinqe features with that of all the features of the rest of the objects which is known.
D. Object Recognition: It will determine “that is apple or any other things”.

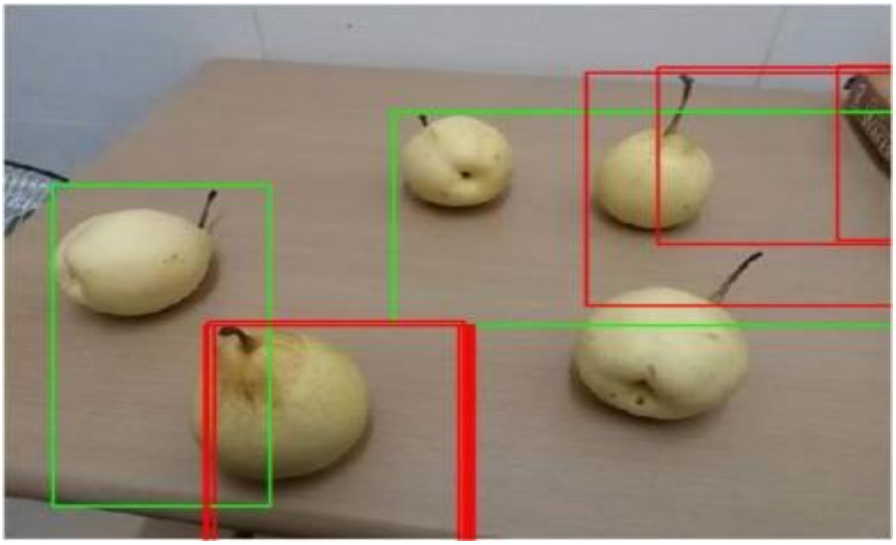


Fig.5. Localisatio and Extraction of feature.

E. Methods

Object detection methodology being developed can majorly be catagorised in two of the approaches, the Approach of Machine learning or approach of Deep learning. All the Approaches which are Machine Learning based, At the first very first features are defined using VJODF or SIFT of HOG. Sevondly the classification is done using different technique, SVM e.g.. Whereas the techniques which are deep learning based make use of Convolutional Neural N/W that make use of end to end object detection without feature definition specifically.

Approaches based on Deep Learning:

- R-CNN, Fast R-CNN, Faster R-CNN, the Region Proposals
- SSD, the Single Shot MultiBox Detector
- YOLO, YouOnlyLookOnce.

Approaches based on Machine Learning:

- Haar features based Viola–Jones object detection framework
- SIFT – Scale invariant feature transform
- HOG features– Histogram of oriented gradients.

F. The Deal

Authors present the entire YOLO v3 scenario. Authors have taken many ideas from similar work and trained a new classifier network, which is improved version than the rest. In the current article authors, present the entire system from scratch as shown Fig 6.

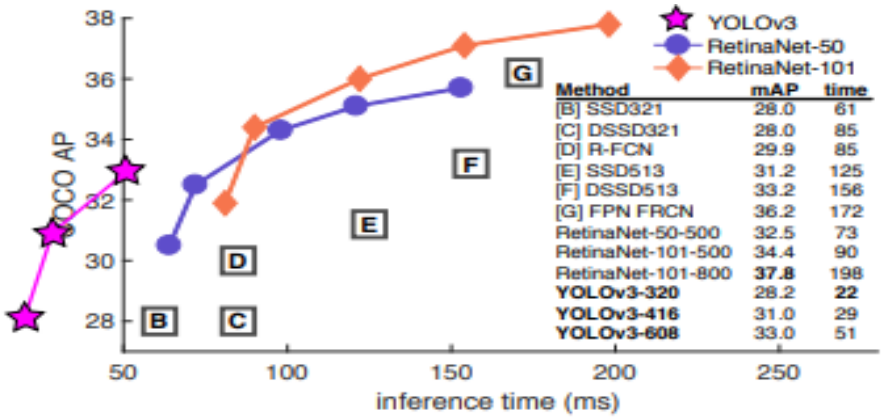


Fig.6. Yolo3 before training of data set.

H. Bounding Box Prediction

Following YOLO9000, our system is able to predict the bounding boxes by the usage of anchor box i.e. dimension clusters [10]. The N/W is able to predict, for every bounding box, four parameters namely t_x, t_y, t_w, t_h. In case the cell has an offset by (c_x,c_y) from the left top corner and priorly the bounding box had a height and width of p_h, p_w, then the predictions can be given as under:

$b_x = \sigma(t_x)+c_x$
$b_y = \sigma(t_y)+c_y$
$b_w = p_w * e^{tw}$
$b_h = p_h * e^{th}$

In the process of training, authors have made use of squared error loss. In case the ground truth with respect to few prediction coordinate is tˆ *, the gradient remains same as the ground truth value, which is calculated from the ground truth box, substracted with own prediction: tˆ * − t*. By inverting the equations given above, it is very easy to compute the ground truth-value. For every bounding box, YOLOv3 does the prediction of objectness score. This is done by the use of logistic regression.

If the bounding box has an ovelaps with the actual object as in the GTO (Ground truth object), this value should be 1. If the prior bounding box has an overlap over the GTO by over a defined threshold, prediction are ignoredas in [7]. Here a threshold of 0.5 is used. The proposed system does an assignment of exactly 1 bounding box prior with respect to each ground GTO. In case a bounding box prior has not been assigned to a GTO, there is no for the coordinate. Working of bounding box shown in Figure 7.

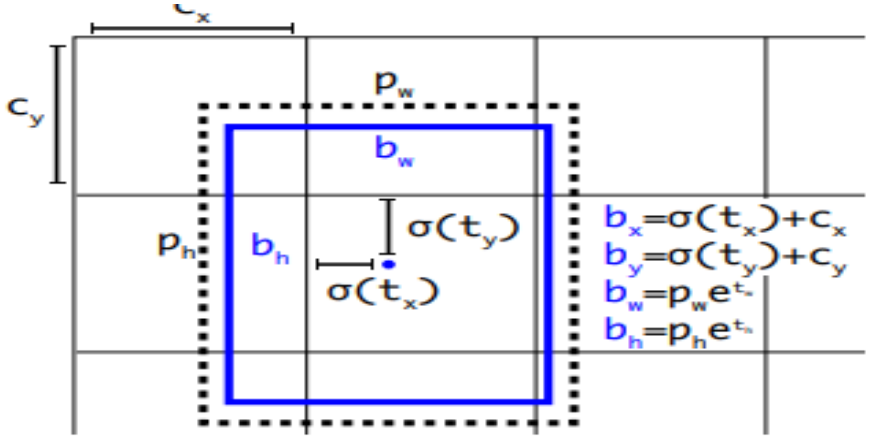


Fig.7. Working of Bounding Box.

I. Prediction of Class

Using the multi-label classification, each of the box does the prediction of the classes the bounding box may encompass. The authors of this article do not prefer a softmax as it has not delivered the good performance. Indeprendent logistic classfiew are used in the methods used by the authors.

For the purpose of the class predicions, binary cross entropy loss is used. While moving to the open image dataset [7], this formulation is very useful. There are plenty of the overlapping labels in thes dataset. Use of the Softmax makes an assumption that only one class is associated with each box which is not correct in most of the cases. A better approach to model the data is multi-label approach.

J. Predictions across Scales

There are 3 different scales at which YOLOv3 predicts the boxes. The system proposed by the authors of this article tries to extract the features, by making use of these scales to nearly same concept to feature pyramid network [8]. Many a convolutional layers are added in the base feature extractor. A 3-dimensional TE bounding box, its objectness and class are predicted as an outcome. Authors have experimented with COCO[10] that predicted three different boxes at each of the scale. Hence the Tensor has a size of is n*n*[3*(4+1+80)]. This is for offsets of 4 bounding boes. These outcomes with a objectiveness prediction and eighty class predictions. Upsampling of feature map by 2 times is done which is taken from previous 2 layres. Concatenation of earlier feature map with that of the upsampled feature map is also done by the authors of this article.

Underlying method has allowed to extract significant semantics because of the upsampled features. This also fine tune the information extracted from feature map found earlier. Some convolutional layers are added to the concatenated feature map. It predicts almost the same tensor but size gets bigger by 2 times. A final design has been performed again for the final prediction of the boxes at the new scale.

In order to determine the bound box priors, k-means clustering algorithm has been used by the authors. Division of the clusters evenly across the scale has been done for arbitrarily chosen 9 clusters & 3 scales. The nine clusters are: (10*13), (16*30), (33*23), (30*61), (62*45), (59*119), (116*90), (156*198), (373*326) on the COCO dataset.

K. Feature Extractor

An all-new N/w was used by the Authors in order to perform the feature extraction. The new N/w was a hybrid approach of the N/W in Darknet-19, nrn & YOLOv2. The proposed network uses CNN layes of size 3×3 and 1×1. It has alternative connections too and because of that, it is very large. Darkenet53 is the name of this because of usage of convolutional layers.

L Training

Authors have performed the multi-scale training on full images. Batch normalization, data augmentation were also performed. For the purpose of training and testing authors used the DNN, Darknet Neural Network framework.

IV Results and Discussions

YOLOv3 is comparative with COCOs. COCOs have unusual average. In terms of mean AP metric YOLOv3 is more or less same as the SSD. It is faster by more than 3 times. This one is not the fastest one though as RetinaNet Model overshadows YOLOv3. It is extremely strong in the metrics over the mAP old detection metric. The comparsion shows that YOLOv3 is preferred version for being used as producing the bounding boxes near the objects. If the IOU threshold is increased, the performance of YOLOv3 downgrades and it finds difficult to yield boxes for the underlying objects.

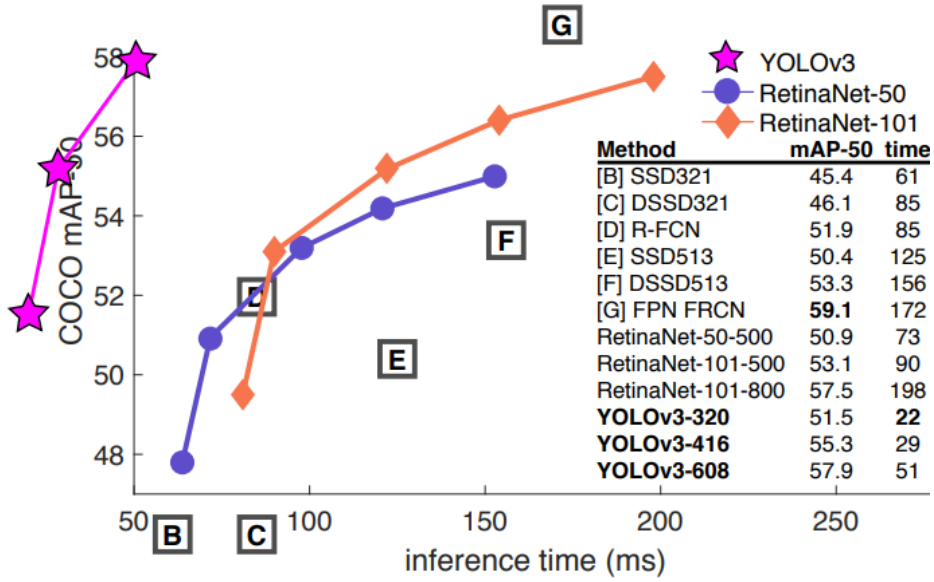
Although YOLO was not able to box the small objects earlier but its recent versions have overshadowed the demerits. YOLOv3 uses the multi scale predictions that significantly higher APS

performance. When the plot of the speed versus accuracy is drawn (see figure 8), It can well be seen that YOLOv3 is significantly better than existing object detection techniques.

The mechanism of usual anchor box prediction was used in the experiments. Using linear activation, it predicts the x, y offset that is the multiple of box's heigh & width. Linear activation stability of the model. A drop of some points in mAP was observed in the experiments. Authors have experimented with the focal loss as well. It caused a drop of mAP by 2 pt approximately.

YOLOv3 is immune to the focal loss problem as it has conditional class predictions and objectness predictions separately. With YOLOv3, it is difficult to assure that there would be no loss from the class predictions.

Truth assignment and Dual IOU thresholds: During the training, 2 IOU thresholds are used in Faster RCNN. The acceptable range of the preduction ovelap of bonding box with ground truth is 0.7 or above. Any value in the range of 0.3 and 0.7 is ignored and less than 0.3 is treated as the negative overlap. Figure 8 shows the outcome of YOLOv3 after data set training.



Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

i

Fig 8. Improvement in performance after training of data set

The output of the trained dataset on our image is shown in Figure 9, Figure 10 and Figure 11

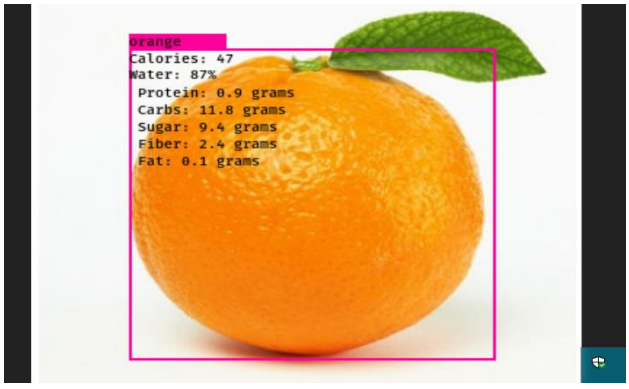


Fig 9. Output showing the properties of Orange

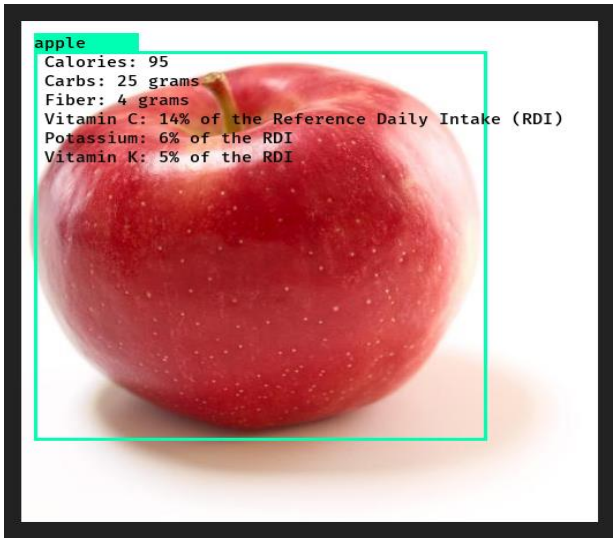


Fig 10. Output showing the properties of Apple

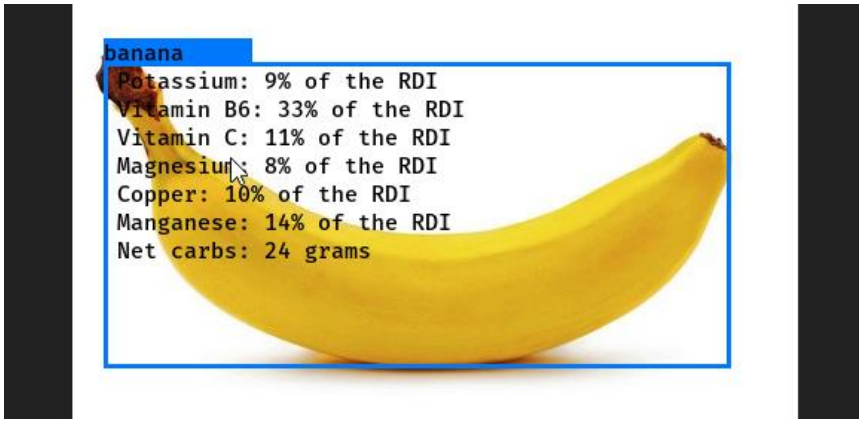


Fig.11. Output showing the properties of Banana

IV Conclusion

The authors have tried to develop an efficient Object detection methodology, which is accurate in the nature as compared to previous researches. The metrics hence obtained are better in many a sense from existing state of the art system of object detection. The suggested methodology makes use of Deep learning and computer vision fundamentals. During the process of labeling, a custom dataset was built. The evaluation turned out to be consistent. The real time scenarios in which object detection is used in the step of preprocessing pipeline, the suggested methology can be used directly. Video sequences for the trafficking is an important application where the suggested technique can be used. It can be concluded that YOLOv3 is a reasonably good object detector. This has not only higher accuracy but is faster as well. YOLOv3 has been proven comparative on the old detection metric of 0.5 IOU.

References

[1]. Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), 303-338.

[2]. Fu, C. Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A. C. (2017). Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

[3]. Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., & Farhadi, A. (2018). Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4089-4098).
[4]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
[5]. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., & Murphy, K (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7310-7311).
[6]. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A. & Belongie, S (2017). Openimages: A public dataset for large scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, *2*, 3
[7]. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117-2125).
[8]. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement *arXiv preprint arXiv:1804.02767*.
[9]. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.
[10]. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.
[11]. Newton, I. (1999). *The Principia: mathematical principles of natural philosophy*. Univ of California Press.