ISSN 2063-5346

MULTIPLE OBJECT DETECTION USING YOLOV3 MODEL

Yogapriya M[1], S.Kanimozhi[2] , Merlin CD[3] ,Pokala rohith[4] ,Panabakam Sri Vignesh[5] ,Velagala Hiranya Soma Sekhar Reddy[6]

Article History: Received: 01.02.2023 Revised: 07.03.2023 Accepted: 10.04.2023

Abstract

— Conventional object detection techniques are based on shallow, trainable structures and handmade characteristics. Building intricate ensembles that integrate several low-level image features with high-level information from scene classifiers and object detector is able to stabilize their accuracy. More potent tools that can learn high-level, semantic and deeper features are being introduced as a result of the rapid advancement of deep learning to solve issues with conventional architectures. The paper provides an innovative method for identifying multiple objects using a publicly accessible image dataset. The dataset includes far-exposed views, such as those taken in bright sunshine, and the intrinsic characteristics are not particularly trustworthy for training, making it challenging to construct detection in it. As a result, we suggest adopting a YOLOv3 model that has already been pretrained for training, increasing its basic accuracy using various regularization and loss techniques. We also suggest a solution for numerous object detection problems, particularly in real time, based on our findings. Many applications, including autonomous vehicles and sophisticated systems for driver assistance, are targeted by these strategies.

Keywords— Multiple Object, Object Detection, Deep Learning and YOLOv3..

m.yogapriya@velhightech.com[1], Kanimozhi274@gmail.com[2] ,
merlin@velhightech.com[3] , rohithchet30@gmail.com[4] ,
panabakamsrivignesh@gmail.com[5],somu93739@gmail.com[6]

[1,2,3,] Assistant Professor, Department of Electronics and Communication Engineering

(Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, India)

[4,5,6,] UG Student, Department of Electronics and Communication Engineering

(Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, India)

DOI:10.31838/ecb/2023.12.s1-B.323

I. INTRODUCTION

Humans are able to quickly detect and locate things in a scene. The visual system in humans is rapid and accurate, and it can carry out difficult tasks like recognizing many objects and spotting impediments without much conscious effort. Thanks to the accessibility of large amounts of information, faster GPUs, and improved algorithms, we are able to quickly training machines to detect and categorize multiple object in an image along with high precision. Using this form of localization and detection, object detection is capable of being utilized to count the items in a scene, as well as to locate and track them in real time while precisely labelling them. Using this form of localization and identification, object detection is capable of being utilized to count the items in a scene as well as finding and tracking them in real time while precisely labelling them.

The task of object detection in computer vision is to locate and identify things in videos and images. It is a crucial component of numerous applications, including robotics, self-driving automobiles, and surveillance. The object recognition bounding box (BBox), which locates the boundaries of an object marked with a square, rectangle, or other clear-cut quadrilateral, is an essential component. They have a label on them that defines the intended target, which might be a person, a dog, or a car. Bounding rectangles may overlap in order to display multiple objects in a single image, provided that the model knows about the objects it is labelling.

The two primary types of object detection algorithms are single-shot detectors and two-stage detectors. Previous detection outlines used image classification methods to find objects while looking at numerous regions of the image repeatedly at various scales. This strategy is ineffective and slow.

Convolutional neural networks are used, which sets object detection using deep learning apart from competing methods (CNN). Given that CNNs can be trained automatically with minimal manual programming, deep neural networks used for object detection produce the fastest and most accurate results for both single and multiple object detection. An object detection model known as the YOLOv3 is presented in this paper. YOLO pursues an entirely different course of action. The entire image is only briefly viewed, and only one network scan is performed to look for objects. It comes rapidly.

The paper is organized as follows: Section Two does a thorough literature review to identify research gaps. A methodology for object detection is proposed in Section III. The simulation analysis and findings are covered in Section IV. The conclusion is covered in Section V.

II. RELATED WORKS

With the Tile convolution neural network and its recursive mode, applications for automated driving may find objects more easily (DAS). Unsupervised training is a component of the approach, which uses a variety of training data to help learn and modify weights. In order to lower the number of valid detections, obstacle validation algorithms are incorporated [1]. To study motion of objects that is not visible to the naked eye, ideas like a histogram of magnitudes and optical flow are used. Localization and classification enable the site setting to distinguish among usual and unusual occurrences, enabling identification of normal and abnormal occurrences [2]. Pretrained networks are used to extract features, while SVM is used to distinguish between outcomes from classification. The route for ITS is guided by the approach [3]. Several methods, such as feature extraction centred on colour and gradients, fall short in providing spatial placement in the image. Using the principal components analysis [4], image undistortion pipeline, classification, picture registration, then detections depending on coordinates and speeds allows for the

resolution of the issues. The method employs detectors such as FAST and FREAK classifiers, followed by Squeeze Net classification [5]. The processes of creating candidate targets, mining features from those targets, and placing ground truth Boxes around objects all help with tracing. VGGNet is used to classify the objects [6]. Convolutional neural networks (CNN), which was created to categories pictures, was modified to carry out object localization and detection. The method views the detection of objects as a return for the object class to the initial detection of objects by their bounds. Instead of repeatedly evaluating an image, as CNN does, it is scanned once to process more frames per second (fps). Unlike the conventional classification technique, YOLO is learned based on losses experienced [7]. The paper discusses the use of video analytics for traffic. Vehicle counting is another important application area in addition to vehicle detection and tracking. Single Shot Detector (SSD), a cutting-edge algorithm, is used. Features like binary big items are handled by the algorithm. In applications like object categorization, it produces superior outcomes. Topics such as background subtraction and the virtual coil approach are used in object tracking. SSDs perform better than YOLO versions in terms of precision. While choosing the best algorithm for object identification at a speed of 58 fps with a performance measure for accuracy exceeding 85% [8], swiftness and accuracy have always been tradeoffs. The study discusses the upgrade to YOLO that was developed. During the course of the chains of YOLO types, i.e. YOLOv1, YOLOv2, and YOLOv3, gradual updating has been observed. Modern technology, such as YOLOv3, is used. Thinner bounding boxes are one example of an upgrade that doesn't affect nearby pixels. The COCO dataset implementation in YOLOv3 shows that mAP is just as effective as SSD. YOLOv3 produces results three times more quickly. YOLOv3 claims to be capable of finding tiny things [9]. Single-object monitoring [10], [12], and [14] will become obsolete as the number of vehicles in urban areas rises. Kernelized correlation filter (KCF) is used to track many objects. Many KCFs are operated concurrently. KCF works better with photos containing occlusions. If used in conjunction with background subtraction, KCF produces accurate outcomes regarding urban traffic.

III. PROPOSED METHODOLOGY

The phenomenon of object detection in computer vision includes detecting different objects in digital photos or movies. Several methods are utilized for object localization and detection, like Retina-Net, fast R-CNN, and faster R-CNN. These methods have overcome the difficulties associated with limited data and modelling in detecting objects, but they still have not been capable of identifying objects in one algorithm's running. An YOLOv3 deep learning model for object detection is proposed in this work. Overall proposed methodology is shown in Fig.1. YOLO is a method for real-time object detection that makes use of neural networks.

The following processes make up the proposed system:

- From a public database, testing videos and images are obtained.
- After loading the pretrained YOLOv3 model, which was trained using the common objects in context (COCO) dataset, which contains many items such as a human, bicycle, vehicle, motorbike, aero plane, bus, train, etc., use the YOLOv3 model to localize and detect the multiple objects in the scene.
- In the YOLOv3 detection module, use the loaded and trained YOLOv3 model to detect several objects.

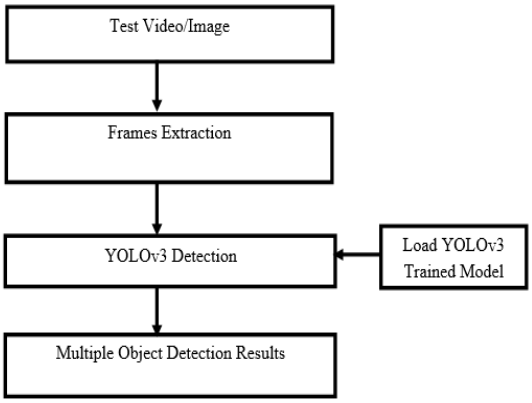

Figure 1. Overall proposed methodology for object detection.

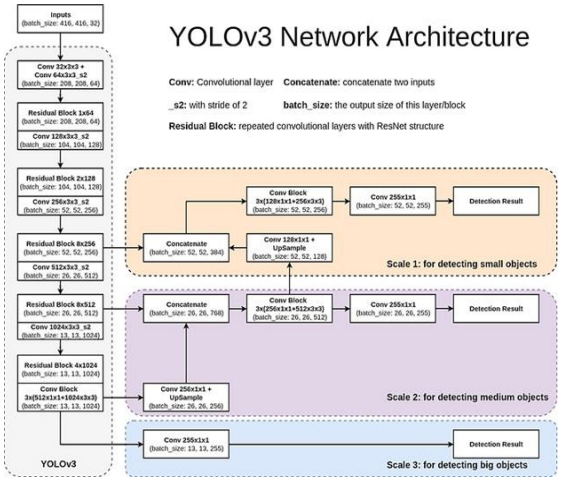

Figure 2. YOLOv3 architecture for multiple object detection.

A. You Only Look Once (YOLOv3) model

The YOLO algorithm that implies "you only look once," forecasts employing 11 convolutions, which indicates that the dimension of the prediction map is precisely the same as the size of earlier feature map. YOLO is a Convolutional Neural Network (CNN) that can quickly detect objects. Due to the fast propagation of low-level information from the early convolutional layers to the subsequent convolutional layers in a deep CNN, the CNN performs impressively when extracting features from visual input. The difficulty in this situation is in precisely identifying multiple objects and pinpointing their exact locations within one visual input. Two key characteristics of CNN, the sharing of parameters and the number of filters, are capable of successfully addressing this object detection issue. The advantage of YOLO is that it is more accurate and faster than other networks.

Versions 1 and 2 of YOLO use softmax functions to convert scores to probabilities. When only opposing items may coexist, this strategy is viable. Multi-label classification is used by YOLOv3. YOLOv3 architecture is depicted in Fig.2. To determine the likelihood that an input belongs to a particular label, an unsupervised logistic classifier is utilized. Using the binary-cross entropy of each label, the loss is determined. Complexity is diminished since the softmax function is left out. Darknet-53 is used by YOLOv3. Darknet-53 model is shown in Fig.3. The YOLO creators Joseph Redmon and Ali Farhadi also created the backbone known as Darknet-53. Each convolution process for darkent53 was followed by batch normalization and then leaky RELU. Due to the fact that Darknet-53 uses 53 convolutional layers instead of the previous 19 (ResNet-101 or ResNet-152) layers, it is more powerful than Darknet-19 as well as more efficient than competing backbones.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 3. Backbone darknet53 network model.

1) Bounding box optimization

The score of object presence is predicted by YOLO v3's logistic regression. All objects are determined by a ground truth box, and if an Anchor Box overlap the Ground Truth Box the greatest, the objectness prediction score is called to be one (1). The cost of the anchor box is zero for those anchor boxes whose overlap surpasses the fixed threshold. Just one anchor box is used to map each ground truth box. The calculation of the confidence loss is the only one made if the anchor box is not preferred and allocated to the BBox.

Gradual optimization is used to regress the Anchor Box to the Ground Truth Box, as seen in Fig. 4. Currently, coordinate parameters are described as follows:

$$bb_x = \partial(p_x) + d_x \quad (1)$$

$$bb_y = \partial(p_y) + d_y \quad (2)$$

$$bb_w = t_w f^{p_w} \quad (3)$$

$$bb_h = t_h f^{p_h} \quad (4)$$

Where, p_x, p_y, p_w, p_h are the prediction generated by YOLO. d_x, d_y are located at the anchor's grid cell's upper left corner. t_w, t_h are the anchor's width and height. The estimated bounding box are bb_x, bb_y, bb_w, bb_h. $\partial(p_0)$ is denote as bounding box confidence score.

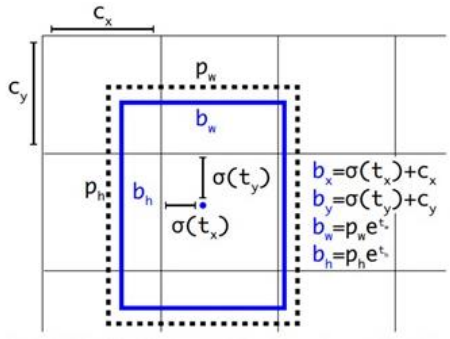

Figure 4. Anchor box regression.

2) Feature pyramid network (FPN)

Every point of the image receives three predictions from YOLOv3. A bounding box, an objectness score, and an 80 class score are all included in the prediction; as a result, we require S*S*[3*(4+1+80)] prediction. This method is related to FPN. As shown in Fig. 5, predictions are made at three distinct scales. At the last feature map layer, the initial prediction is made. After that, the feature map is upsampled by an aspect of 2. Model of YOLOv3 uses element-wise addition to combine the feature map with up-sampled features. The application of a convolutional layer yields second predictions. Performing the second prediction will produce a lot of semantic data.

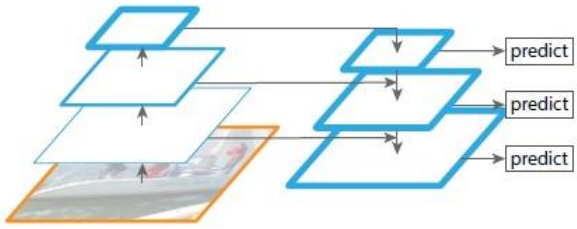

Figure 5. Feature pyramid network (FPN).

YOLOv3 has a mean average precision (mAP) and intersection over union (IOU) score that is swift and accurate. It operates

a lot more quickly than alternative detecting techniques.

IV. EXPERIMENTAL RESULTS

This section details the simulation outcomes for the YOLOv3 deep learning model's multiple object detection. On an Intel® Core i5 computer with 12 GB of RAM, a YOLOv3 model simulation was analyzed.

The publicly available videos and images have been provided as a test collection. The collection includes three different types of images, including day, night, and near infrared images. YOLOv3 object detections are displayed in Figs. 6, 7, 8, 9 and 10. The outcome in Fig. 6 demonstrates that the algorithm is capable of detecting objects of different dimensions in images taken from a variety of camera angles and distances. This characteristic results from the use of FPN in YOLOv3.

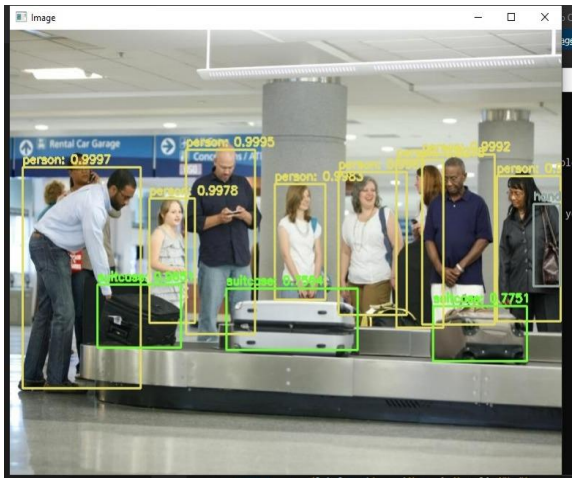

Figure 6. Object detection using YOLOv3 in scene 1.

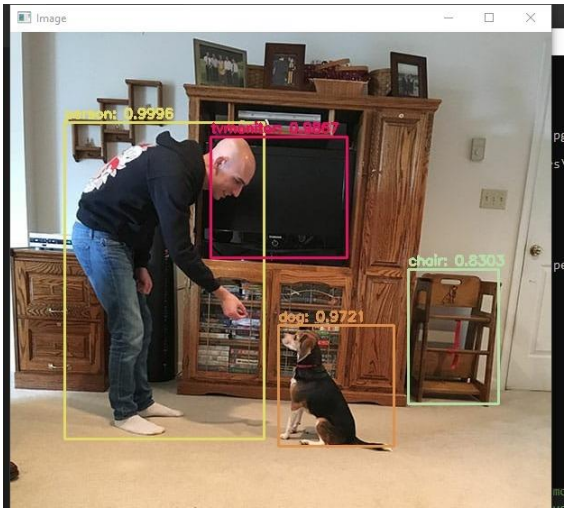

Figure 7. Object detection using YOLOv3 in scene 2.

Figure 8. Object detection using YOLOv3 in scene 3.

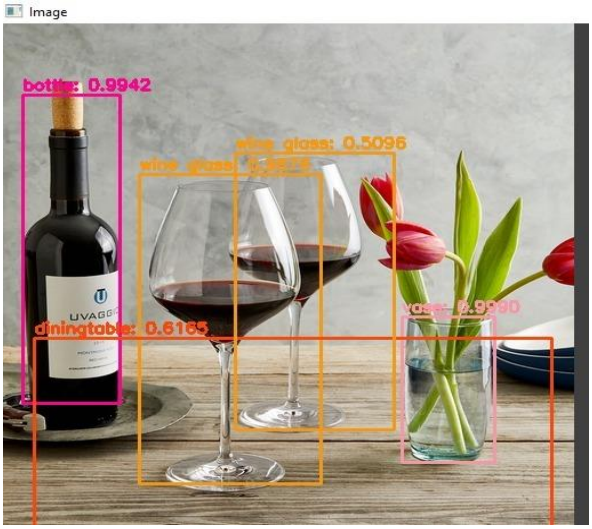

Figure 9. Object detection using YOLOv3 in scene 4.

Figure 10. Object detection using YOLOv3 in scene 5.

V. CONCLUSION

The majority of computer vision systems and robotic computer vision depend deeply on object localization and detection. Although substantial development has been made in recent years and certain current approaches are already included in numerous consumer gadgets (such as face identification for auto-focus in phones) or have been integrated into assisted driving systems, we are still a long way from obtaining human-level performance, particularly in the area of open-world learning. Deep learning-based detection of objects has become a popular area of study recently because of its strong capacity for learning and benefits in handling occlusion, size transformations, and backdrop shifts. This study presents a deep learning-based object detection system that addresses several sub-problems, like clutter, occlusion, and low resolution, using the new YOLOv3 model.

REFERENCES

[1] V. D. Nguyen et all., "Learning Framework for Robust Obstacle Detection, Recognition, and Tracking", IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 6, pp. 1633-1646, June 2017.

[2] Zahraa Kain et all, "Detecting Abnormal Events in University Areas", 2018 International conference on Computer and Applications(ICCA),pp. 260-264, 2018.

[3] P. Wang et all., "Detection of unwanted traffic congestion based on existing surveillance system using in freeway via a CNN-architecture trafficnet", IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, 2018, pp. 1134-1139.

[4] Q. Mu, Y. Wei, Y. Liu and Z. Li, "The Research of Target Tracking Algorithm Based on an Improved PCANet", 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 2018, pp. 195-199.

[5] H. C. Baykara et all., "Real-Time Detection, Tracking and Classification of Multiple Moving Objects in UAV Videos", 29th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Boston, MA, 2017, pp. 945-950.

[6] W. Wang, M. Shi and W. Li, "Object Tracking with Shallow Convolution Feature", 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 2017, pp. 97-100.

[7] K. Muhammad et all., "Convolutional Neural Networks Based Fire Detection in Surveillance Videos", IEEE Access, vol. 6, pp. 18174- 18183, 2018.

[8] D. E. Hernandez et all., "Cell Tracking with Deep Learning and the Viterbi Algorithm", International Conference on Manipulation, Automation and Robotics at Small Scales (MARSS), Nagoya, 2018, pp. 1-6.

[9] X. Qian et all., "An object tracking method using deep learning and adaptive particle filter for night fusion image", 2017 International Conference on Progress in Informatics and

Computing (PIC), Nanjing, 2017, pp. 138-142.

[10] Y. Yoon et all., "Online Multi-Object Tracking Using Selective Deep Appearance Matching", IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), Jeju, 2018, pp. 206-212.

[11] H. S. Bharadwaj, S. Biswas and K. R. Ramakrishnan. "A large scale dataset for classification of vehicles in urban traffic scenes", Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing,ACM,2016.

[12] Mohana et.al., "Performance Evaluattion of background modeling methods for object Detection and Tracking",International Conference on Inventive systems and Control (ICISC).

[13] G. Chandan et.al., "Real Time Object Detection and Tracking Using Deep Learning and OpenCV", International Conference on Inventive Research in Computing Applications (ICIRCA), 2018.

[14] Mohana et.al., "Elegant and efficient algorithms for real time object detection, counting and classification for video surveillance applications from single fixed camera" International Conference on Circuits, Controls, Communications and Computing (I4C),2016.

[15] Mohana et.al., "Simulation of Object Detection Algorithms for Video Survillance Applications", 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud),2018.

[16] A. Raghunandan et. al., "Object Detection Algorithms for Video Surveillance Applications," International Conference on Communication and Signal Processing (ICCSP), 2018.

[17] A. Mangawati et al., "Object Tracking Algorithms for Video Surveillance Applications," 2018 International Conference on Communication and Signal Processing (ICCSP), 2018.

[18] Mohana, et al., "Design and Implementation of Object Detection, Tracking, Counting and Classification Algorithms using Artificial Intelligence for Automated Video Surveillance Applications" Advanced Computing and Communication Society (ACCS)- 24th annual International Conference on Advanced Computing and Communications (ADCOM-2018), IIITB, Bangalore