Vehicle and Object Detection Using YOLOv3 Algorithm

G. Raja Gopal, Dr. N. Veeranjaneyulu

Department of Information Technology, Vignan's Foundation for Science, Technology & Research, Guntur, Andhra Pradesh, India.
veeru2006n@gmail.com

Abstract

YOLOv3, the third edition of the YOLO family, performs well on object detection, but using it for real-time vehicle and object detection on unmanned vehicles with limited computing capacity remains a very challenging task YOLOv3 has a high computational complexity. The main objective is to develop network architecture for vehicle and object detection based on YOLOv3. The total work will be broken down into three phases. Firstly, to reduce the model size and computing complexity, we introduce L1 regularization to the batch normalization layer, which allows us to recognize and remove distracting channels and layers. Secondly, to reduce the missed detection in crowded scenes and locate targets better, the Merge Soft-NMS which merges the bounding boxes with high overlap is designed based on Soft-NMS. Thirdly, considering the obvious aspect ratio of vehicle and objects, the anchor boxes which are designed based on multi-class is redesigned for better vehicle and object matching and localization in YOLOv3. In the experiment, compared with SINGLE SHOT DETECTION (SSD) and YOLOv3 which performs well on detection accuracy and speed is effective and compact for vehicle and object detection.

.

Keywords: *object detection, model compression, NMS, Single Shot Detection (SSD).*

1. Introduction

Automated driving is developed with the promise of preventing accidents, reducing emissions, transporting the mobility-impaired, and reducing driving related stress. The autonomy system of self-driving cars is generally divided into the perception system, the decision-making system, and the control system. Traffic safety is a very important issue for human life. Hundreds of people die every day due to traffic accidents. In addition to the loss of life, there are also some problems caused by not following the traffic rules. The perception system is commonly responsible for tasks such as obstacles detection and tracking, object and vehicle detection, traffic signs detection, road scene recognition, among others. With the rapid development in deep learning in recent years, many achievements have been made in object detection using convolutional neural networks compared to traditional methods based on the hand-engineered feature. As achieving a good balance between accuracy and speed, YOLO series detectors are probably the most widely used in a variety of fields. YOLOv3 is the third version of the

YOLO series, which adopts some novel strategies to improve YOLOs in speed and accuracy significantly.

With the rapid development of economy and the gradual people's quality of life is improving more and more families own private cars, which leads to the rapid growth of urban traffic vehicles, and traffic management is faced with great challenges. Therefore, rapid detection and identification of vehicles in traffic images have become an important task in urban traffic management and a research focus in the field of computer vision.

2. Literature Review

[1] The approach used in this research detects vehicles in real time using in tunnel imagery and estimates the distance between them to help drivers focus safety. The method generates vehicle learners to detect cars using YOLOv3-based deep learning methods using tunnel photos taken in multiple time zones, lighting, and conditions. It detects automobiles using the vehicles detector.[2] The YOLOv3 is proposed in this research for vehicle and object detection. To simplify the network design, YOLOv3 employs channel and layer pruning. Merge Soft-NMS performance improves, and anchor boxes improve detection accuracy. The trials show that it is capable of achieving higher detection accuracy, has fewer parameters, and runs faster.[3] The approach Contextual-YOLOv3 is used to detect the small objects in this paper. This enhances small target identification accuracy while retaining recognition speed. Small object detection algorithms that have been proposed in the past have a better balance of speed and accuracy. To replace the original classification probability, our approach creates a YOLOv3 contextual probability. Meanwhile, instead of using a non-maximum suppression algorithm (NMS) approach, utilise a context-based filtering algorithm for optimal window selection.[4] Based on YOLOv3, this research presents an enhanced approach of object detection in remote sensing images. The enhanced technique improves the mAP value of small object detection in remote sensing images without introducing any model parameters, lowers the LAMR value of small object detection in remote sensing images, and eliminates detection accuracy when compared to the original method.[5] The strategy presented in this paper uses an existing outstanding neural network to fuse and develop a new one. This network can use the benefits of RGB camera and radar point cloud to create a multi-sensor fusion-based 3D vehicle object detection system. When compared to the single use of RGB photos or 3D point cloud data, the accuracy of 3D targets may be greatly enhanced in general scenarios, while processing speed can also be improved.[6] The YOLOv3 model is utilised in this study for vehicle object detection on the vehicle and non-vehicle data sets as well as photographs. Using the depth convolution feature, the YOLOv3 model effectively removes the reliance of standard vehicle object identification models on manual feature selection. Label smoothing is used to reduce the model size, while k-means++ clustering is used to prevent the gradient disappearance problem and increases training accuracy. The detection model can improve its detection of vehicle objects as result of this experiment.[7] The YOLOv3 model is trained with the GTX950 GPU in this study. It's

also compatible with the ZCU102 FGPA card, and the GPU and FGPA card's working speeds are compared. According to the observed results, the FPGA card produces the best performance. The chance of recognizing things can be improved by increasing the dataset size. The model can be run faster by altering the model input size.

3. Methods and Implementation

A. CHANNEL AND LAYER PRUNING: When experts design network architectures designed by experts for object detection, it doesn't guarantee that all component play an important role in an actual deployment scenario. Model compression is an essential component of deploying a deep object detector in computationally resource constrained environments. The channel and layer pruning is performed on convolutional layers to obtain a compact and effective network architecture for object detection. The channel and layer pruning in YOLOv3 and YOLOv3 are shown in following figure.

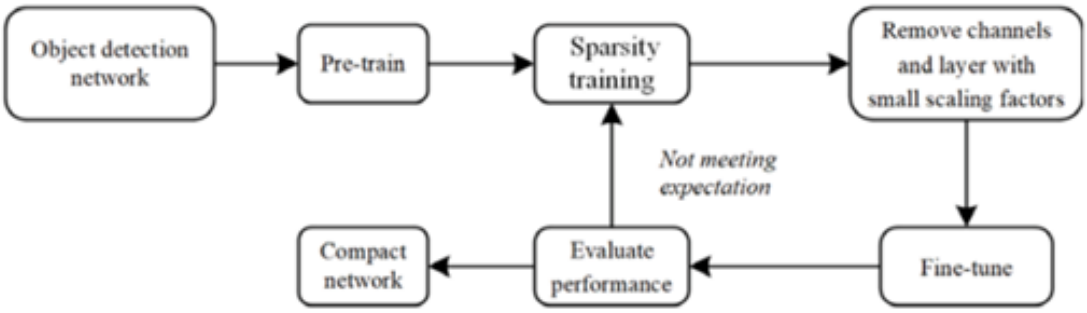

Fig 1: The procedure of a detection model being sparse trained and pruned to a compact model.

In order to speed convergence, the network is pre-trained before sparsity training. During sparsity training, unimportant channels are identified and removed from the convolutional layer. Most modern CNNs use batch normalization to achieve faster convergence and better generalization.

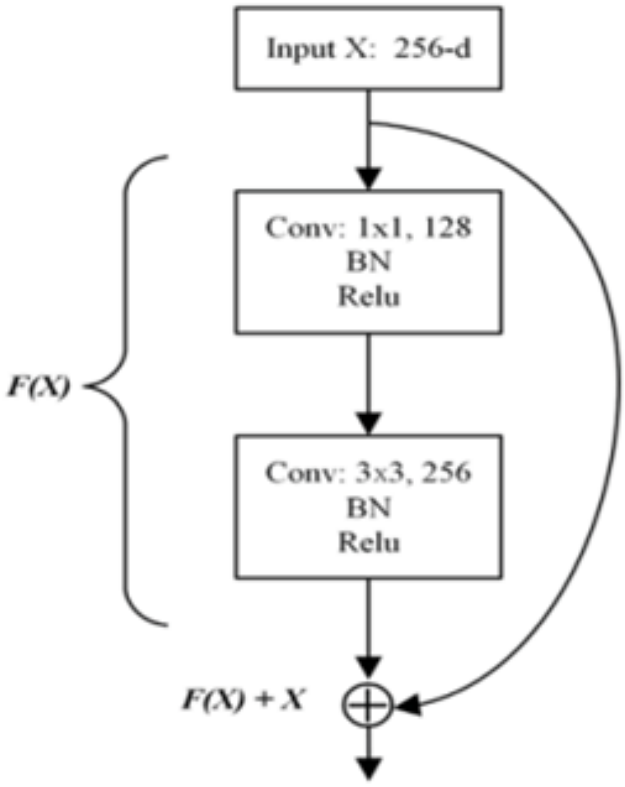

Fig 2: A residual block in YOLOv3

The YOLOv3 network uses 23 residual blocks to address the gradient disappearance issue caused by increasing depth in the deep neural networks. As shown in above figure, we input a 256-dimensional feature map that contains two convolution layers along with BN layer and Relu activation. The whole process is represented by F(X), where F(x) simply adds input X as input to the next convolutional layer.

Each residual block of the YOLOv3 consists of two convolution layers and an addition operation, each of which requires more parameters and computing operations, as well as run time memory. Some residual blocks, however, do not have an impact on the overall performance of the architecture. By analyzing the overall distribution of parameters in the second BN layer of all residual blocks after sparsity training, it appears that parameters in some residual blocks are almost zero. Consequently, the output of these residual blocks F(X) + X is approximately equal to X. The model identification performance was aided by these residual blocks to a lesser extent.

To identify unsuccessful residual blocks, all the residual blocks are sorted by the absolute parameter mean values of the second BN layer in ascending order. The amount of residual blocks to be prune is determined by the pruned model's performance. In the YOLOv3 network, eight residual blocks are deleted, resulting in the removal of 24 layers. The model becomes slimmer after the channel pruning phase, while the model becomes shorter after the layer pruning step.

A. MERGE SOFT-NMS: Non-Maximum Suppression (NMS) is currently used in most object detection systems to eliminate the redundant bounding box. The NMS sorts all bounding boxes for each class depending on confidence score, as seen in pseudo code given below. The bounding box M with the highest score is chosen, and any other bounding boxes that overlap heavily with box M are suppressed.

With just one life of code, the Soft-NMS employs the penalty function to decay the detection scores of all other bounding boxes that overlap. The figure 3 shows the penalty function used in Soft-NMS.

When the overlap is modest, the continuous penalty, also known as the Gaussian penalty function, decays the score gradually, while the output penalty increases considerably when the overlap is high. The bounding box score that does not overlap with M would not be penalised.

Before the post NMS, the bounding boxes that overlapped a lot were usually various predictions of same target. The score of these bounding boxes would be deleted in Soft-NMS due to the large overlap, but the location information is valuable and should not be disregarded. Setting a threshold in Merge Soft-NMS to merge the bounding boxes that overlap IOU with M exceeding it. The score of bounding boxes is utilised as the factor to average the coordinates of them, as seen in the pseudo code of merge Soft-NMS. The merged forecast is more reliable than the single box M prediction with the highest score.

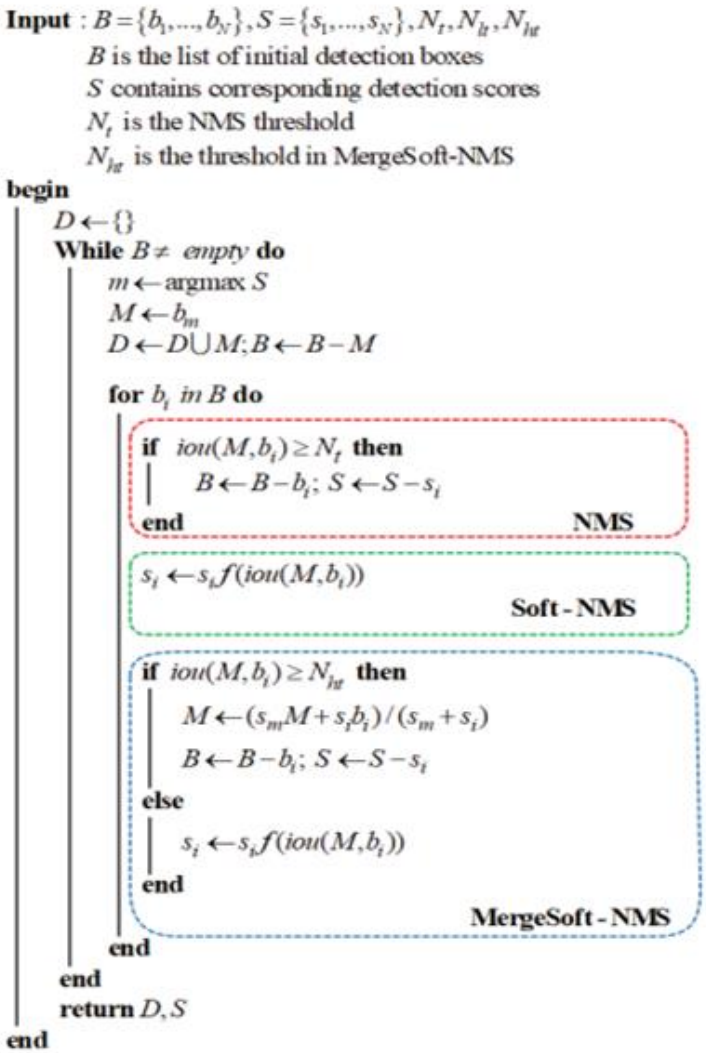

Fig 3: The pseudo code of 3NMS algorithms

C. IMPROVEMENT OF ANCHOR-BOX: To obtain the previous anchor boxes, YOLOv3 uses k-means clustering on the training datasets. The starting centres of the k-means clustering are chosen at random, which is likely to result in a local optimum. Unlike k-means, which selects beginning centres at random, k-means++ assures a better initialization of the centres, ensuring that the initial centres are widely apart as feasible and improving clustering quality. The rest of the k-means++ clustering is identical to the k-means clustering.

We use k-means++ clustering on classes of vehicle and object in vehicle, non vehicle data set and photos to build anchor boxes with a better representation for vehicle and object and make the model easier to learn.

4. Proposed System

In this proposed system YOLOv3 detect vehicles and objects and compare with Single Shot Detection (SSD). In terms of accuracy and speed. By comparing these two algorithms YOLOv3 gives the high accuracy with quick detection.

A. You Only Look Once Version 3 (YOLOV3): YOLO is a method that detects items in an image and their positions all at once using artificial neural networks. It divides the image into cells and produces output vectors that attempt to estimate confidence score values and bounding boxes for each cell's items. It based on the Darknet frame work, which was created in C/Cuda. It is fundamentally inspired by the Google Net structure, which comprises of 24 convolutions and 2 dense layers, and hence offers a very high performance.

In April 2018, the third version of YOLO was published. The primary motivation for this version is to determine bounding boxes for various sizes. The number of fundamental layer has been increased to 53. The YOLOV3 architecture is depicted in the diagram below.

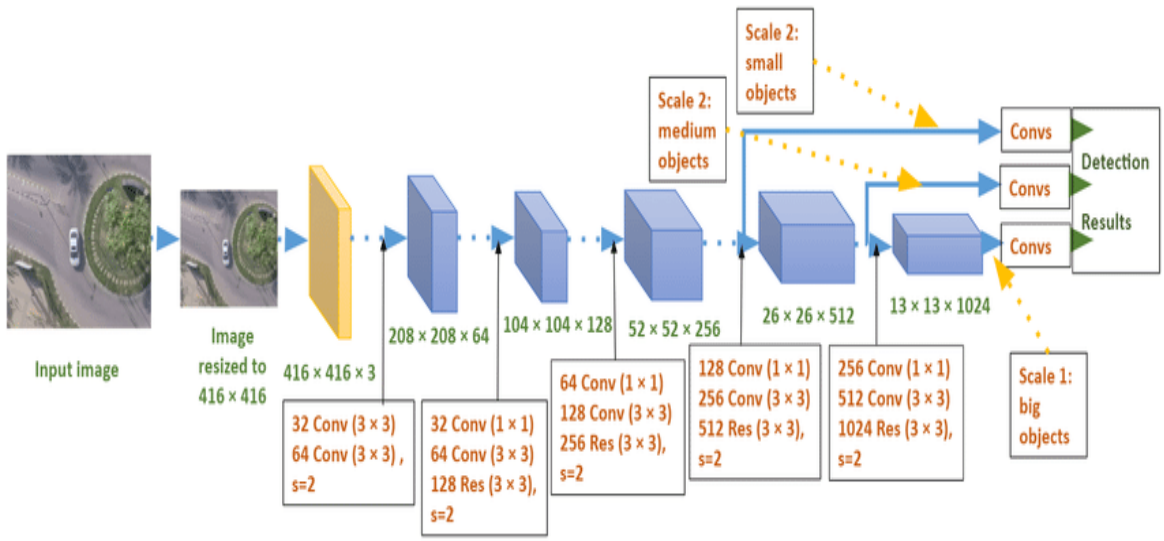

Fig 4: YOLOv3 Architecture

B. Single Shot Detection(SSD): Single Shot Detector like YOLO takes only one shot to detect multiple objects present in an image usig multibox. High detection accuracy in SSD is achieved by using multiple boxes or filters with different sizes, and aspect ratio for object detection. It also applies these filters to multiple feature maps from the later stages of a network. This helps perform detection at multiple scales. SSD has a base VGG – 16 network followed by multibox convolutional layer. Base neural network extract features.

VGG – 16 base network for SSD is standard CNN architecture for high quality image classification but without the final classification layers VGG – 16 is used for feature extraxtion. Input to SSD is an input image with ground truth bounding boxes for each object in the images. During training time the default boxes are matched over aspect ratio, location and scale to the ground truth boxes.

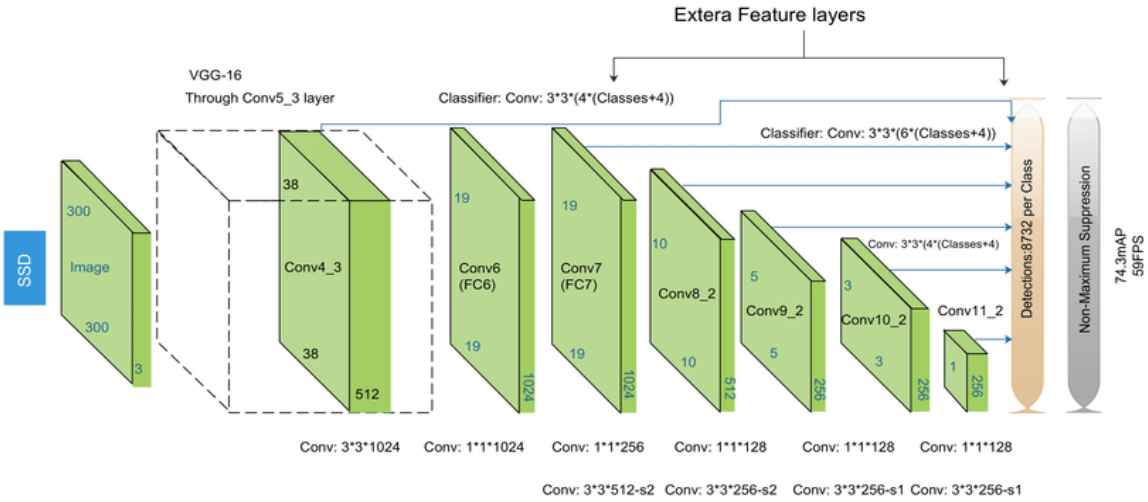

Fig 5: Single Shot Detection (SSD) Architecture

5. Experimental Results

First and foremost, the essential libraries and the environment must be established before the YOLOV3 model can be trained on the dataset. Google provided open source vehicle, object and label photos, which were downloaded from the open images dataset website. Thanks to the GPU hardware, the training process on the YOLOv3 model is completed rapidly with the help of the installation and download images.

Weights and configuration files are obtained as a result of the training. The GPU is then used to test images from the learned model. The accuracy and frames per second (FPS) values obtained from the test images are recorded, and the same model is ready to be used on the vehicle and image dataset.

The YOLOv3 is compared to the Single Shot Detection (SSD) in this experiment. To develop a more accurate detection model for vehicles and objects. When these two models are compared, YOLOv3 is the better detection model, with a detection accuracy of above 95%.

While the single shot detection model only gets to about 75% of the time.

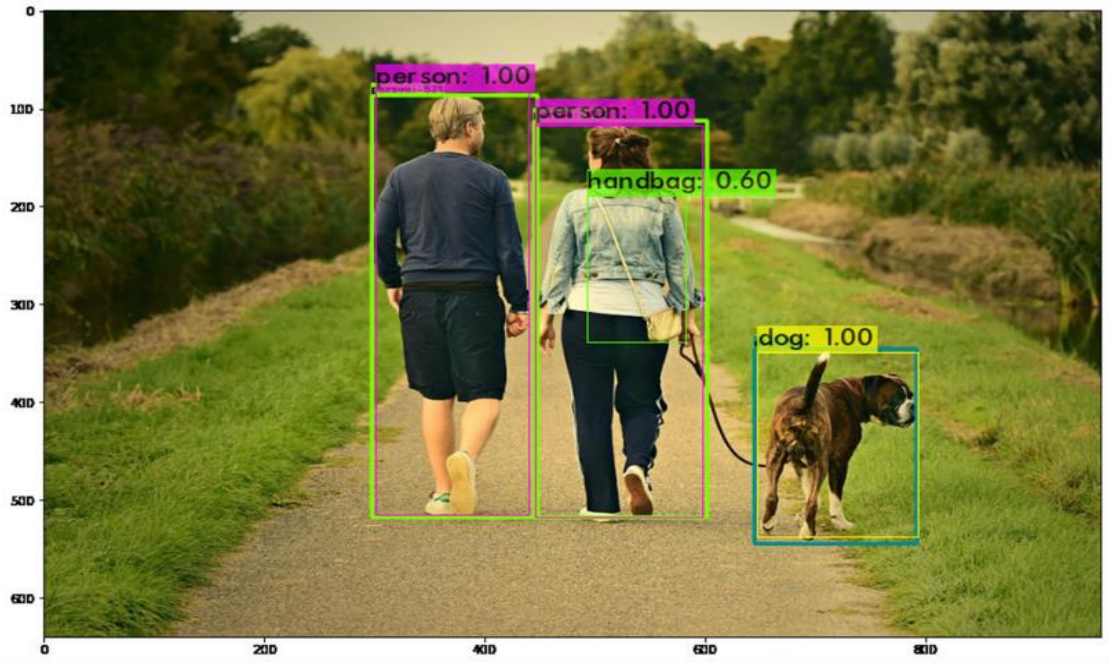

Fig 6: YOLOV3 DETECTED IMAGE

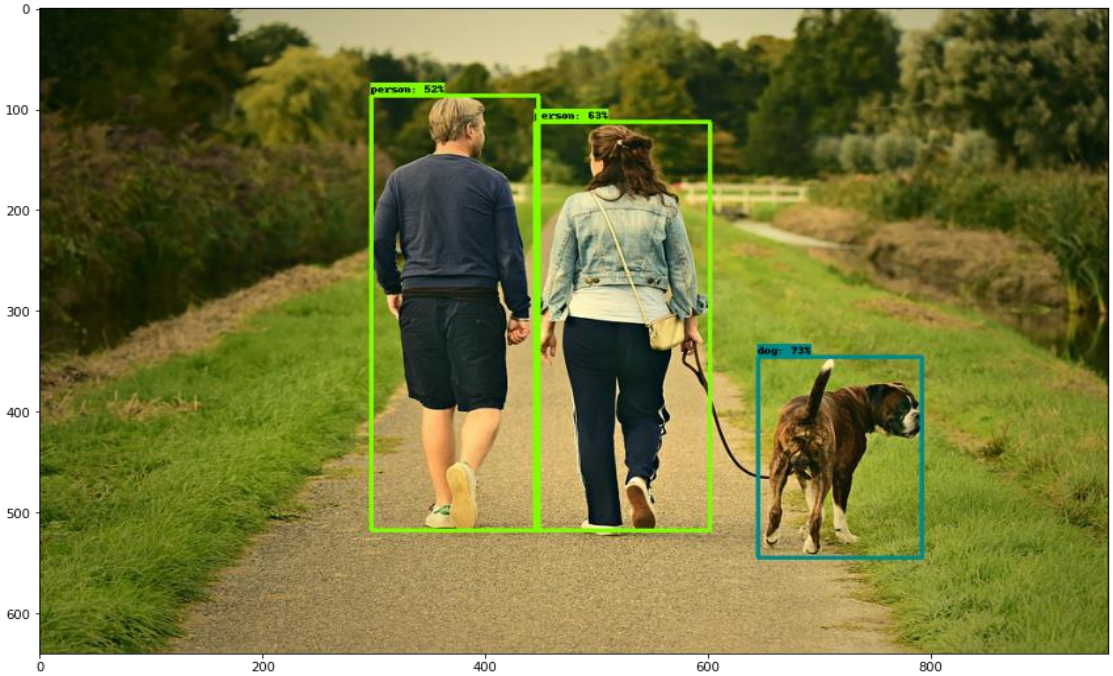

Fig 7: SINGLE SHOT DETECTED IMAGE

6. Comparative Analysis

The following table contracts the experimental results among YOLOv3 and SSD methods in terms of accuracy and detection time. These two methods were tested on images. We can see from below table, the accuracy of YOLOv3 is above 90%, the accuracy of SSD is less than 75%. It is shown that the accuracy of the YOLOv3 is higher and more accurate than SSD model. And the detection of YOLOv3 method is 0.16s, the

detection time of SSD method is 0.2s. It shows that YOLOv3 has a shorter detection time compared with SSD model.

Model	Datasets	Accuracy	Test Time/s
YOLOv3	Images	>90%	0.16
SSD	Images	<75%	0.2

7. Conclusion

In this experiment YOLOv3 model is used for vehicle and object detection on the images and vehicle data set. YOLOv3 model effectively avoids the dependence of traditional vehicle and object detection model on manual feature selection by using convolutional neural network. To reducing the model complexity and model size is done by L1 regularization. To reduce the missed detection in crowded scenes, Merge Soft-NMS is used. Through experiment, YOLOv3 detection Model can achieve better vehicle and object detection compared to SSD. It can be concluded from the experimental results YOLOv3 is better for vehicle and object detection.

References

[1] *J. B. Kim, "A Study on the Development of the Driver's Intensive Warning System during Tunnel Driving Based on Real-Time Vehicle Detection and Distance Estimation," TENSYMP 2021 - 2021 IEEE Reg. 10 Symp., pp. 1–4, 2021, doi: 10.1109/TENSYMP52854.2021.9550929.*

[2] N. Zhang and J. Fan, "A lightweight object detection algorithm based on YOLOv3 for vehicle and pedestrian detection," Proc. IEEE Asia-Pacific Conf. Image Process. Electron. Comput. IPEC 2021, pp. 742–745, 2021, doi: 10.1109/IPEC51340.2021.9421214.

[3] H. W. Luo, C. S. Zhang, F. C. Pan, and X. M. Ju, "Contextual-YOLOV3: Implement better small object detection based deep learning," Proc. - 2019 Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2019, pp. 134–141, 2019, doi: 10.1109/MLBDBI48998.2019.00032.

[4] K. Wu, C. Bai, D. Wang, Z. Liu, T. Huang, and H. Zheng, "Improved Object Detection Algorithm of YOLOv3 Remote Sensing Image," IEEE Access, vol. 9, pp. 113889–113900, 2021, doi: 10.1109/ACCESS.2021.3103522.

[5] Z. T. Li, M. Yan, W. Jiang, and P. Xu, "Vehicle object detection based on rgb-camera and radar sensor fusion," Proc. - Int. Jt. Conf. Information, Media, Eng. IJCIME 2019, pp. 164–169, 2019, doi: 10.1109/IJCIME49369.2019.00041.

[6] X. Sun, Q. Huang, Y. Li, and Y. Huang, "An improved vehicle detection algorithm based on yolov3," Proc. - 2019 IEEE Intl Conf Parallel Distrib. Process. with Appl. Big Data Cloud Comput. Sustain. Comput. Commun. Soc. Comput. Networking, ISPA/BDCloud/SustainCom/SocialCom 2019, pp. 1445–1450, 2019, doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00208.

[7] F. Esen, A. Degirmenci, and O. Karal, "Implementation of the Object Detection Algorithm (YOLOV3) on FPGA," Proc. - 2021 Innov. Intell. Syst. Appl. Conf.

ASYU 2021, pp. 3–6, 2021, doi: 10.1109/ASYU52992.2021.9599073.

[8] Bodapati, Jyostna Devi, U. Srilakshmi, and N. Veeranjaneyulu. "FERNet: A Deep CNN Architecture for Facial Expression Recognition in the Wild." Journal of The Institution of Engineers (India): Series B (2021): 1-10.

[9] Bodapati, Jyostna Devi, and Naralasetti Veeranjaneyulu. "Feature extraction and classification using deep convolutional neural networks." Journal of Cyber Security and Mobility (2019): 261-276.

[10] Bodapati, Jyostna Devi, Nagur Shareef Shaik, and Veeranjaneyulu Naralasetti. "Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification." Journal of Ambient Intelligence and Humanized Computing 12.10 (2021): 9825-9839.

[11] Bodapati, Jyostna Devi, Nagur Shareef Shaik, and Veeranjaneyulu Naralasetti. "Deep convolution feature aggregation: an application to diabetic retinopathy severity level prediction." Signal, Image and Video Processing 15.5 (2021): 923-930.

[12] Bodapati, Jyostna Devi, and Naralasetti Veeranjaneyulu. "Abnormal network traffic detection using support vector data description." Proceedings of the 5th international conference on frontiers in intelligent computing: Theory and applications. Springer, Singapore, 2017.

[13] Bodapati, Jyostna Devi, et al. "Joint training of two-channel deep neural network for brain tumor classification." Signal, Image and Video Processing 15.4 (2021): 753-760.

[14] Bodapati, Jyostna Devi, et al. "Msenet: multi-modal squeeze-and-excitation network for brain tumor severity prediction." International Journal of Pattern Recognition and Artificial Intelligence 35.07 (2021): 2157005.