DOI. No. 10.13052/2794-7254.001

Advancements in Deep-Learning-Based Object Detection in Challenging Environments

Abhishek Mukhopadhyay and Pradipta Biswas*

Abstract: This article focuses on object detection in challenging environments, where objects of interest need to be detected in images captured under unconstrained conditions. These environments can include outdoor scenes with varying lighting, weather conditions, and background clutter. Object detection in such scenarios is crucial for applications like autonomous driving, surveillance, and industrial production inspection. The paper explores existing techniques for object detection in challenging environments and proposes novel solutions to enhance its performance. One key aspect emphasized in the paper is the need for relevant datasets to validate models under unconstrained scenarios. While several datasets already exist, the research identifies gaps in specific situations and addresses them by proposing new datasets, such as the Indian Lane Dataset for autonomous vehicles. The proposed object detection framework leverages spatial information to detect objects in challenging environments, and its performance is evaluated against benchmark datasets and the newly proposed datasets. A real-world application demonstrates significant improvements in object detection performance in natural environments.

Keywords: Object detection, convolutional neural network, lane detection, novel traffic participants, automatic taxi.

1. Introduction

Object detection is fundamental research in computer vision to perceive the environment and localize objects. Modern object detection applications have been prevalent since the early 1960s. Its first applications were in office automation-related tasks, such as character pattern recognition systems and assembly & verification processes in the semiconductor industry. These applications directly contributed to European countries' economic development [1, 2]. Despite advancements, object detection remains challenging among all computer visual tasks. Analysing a scene and recognizing all the object's constituents is a daunting task [3]. Object detection is challenging for several reasons, like variations in object appearance, lighting conditions, scale, occlusions, and cluttered backgrounds. Zou et al. [4] provide a comprehensive overview of the history, challenges, and recent advances in object detection. The authors discuss the limitations of traditional object detection approaches, such as sliding window and feature-based methods, and the emergence of deep-learning-based methods. Liu et al. [5] provide an in-depth review of deep-learning-based object detection methods. The authors discuss the challenges of object detection, such as scale variation, occlusion, and multi-object detection, and how deep-learning-based methods address these challenges. Objects can be partially or fully occluded by other objects, making them difficult to detect. This is particularly true in industrial production or road participants detection, where checking the orientation of components becomes challenging due to illuminations or small objects being difficult to detect due to the high density of road participants. Variability in shape, size, and texture makes it challenging to detect them using a fixed set of criteria. The cluttered background is another challenging situation. Objects can be camouflaged by the background, making them difficult to distinguish. The requirement of large-scale datasets, robustness, adaptability to different environments and situations, and optimization towards speed and accuracy are the criterion for acceptance of any object detection model for real-time use. Traditional object detection techniques have several limitations, including limited flexibility, difficulty handling complex scenarios, and computational cost. Traditional object detection techniques can be computationally expensive, especially when processing large images or video streams. The performance of the traditional object detection model reached saturation point in 2010 [4]. The rebirth of the convolutional neural network changed the scenario [6]. Deep learning approaches can learn robust and high-level features from the image specific to the object class being detected, allowing them to adapt to new object classes or variations in object appearance. The progress in the computational system makes them faster and more efficient than traditional models. Overall, all these models broadly can be categorized into four classes, (i) Two-stage models, (ii) Single stage models, (iii) Transformer based models, and (iv) Segmentation based models. Table 1 explains the advantages and disadvantages of all four types of models. In recent time, researchers [4] suggested that semantic segmentation can improve object detection because of its more precise object localization, better feature representation, improved context modelling and so on.

This article describes one such novel hybrid semantic segmentation models for addressing object detection in challenging environments. It shows the application in two areas important for

Center for Product Design and Manufacturing, Indian Institute of Science, Bangalore, India
E-mail: abhishekmukh@iisc.ac.in; pradipta@iisc.ac.in
*Corresponding Author
Manuscript received 08 August 2023, accepted 02 January 2024, and ready for publication 21 March 2024.

autonomous vehicle. Finally, it describes one such case study to show how the model can be useful in real-world applications.

2. Architecture of Semantic Segmentation Model

A model is developed combining a dilated convolution branch in parallel to the encoder-decoder branch, inspired by the work of Badrinarayanan et al. [7]. The encoder part of the model utilizes the first three convolutional blocks of the Visual Geometry Group (VGG) 16 network [8] to extract image features. In the decoder part, each layer upsamples the feature maps corresponding to its encoder counterpart using memorized max pooling indices. These sparse feature maps are then convolved with decoder filters to produce dense feature maps. However, while predicting lane markings, segmentation methods based on the encoder-decoder architecture may struggle to preserve global context, smoothness, and continuity in the presence of occlusions and other road objects [9]. To address this, dilated convolutional layers are incorporated in parallel to the encoder-decoder branch. These layers enrich the feature map by leveraging the low-level shape features of lanes. The proposed dilated convolutional network consists of 5 convolutional layers that apply 3×3 convolutions with different dilation rates. A 1×1 convolution layer is used at the end to ensure the same number of channels as the input. Ultimately, a hybrid structure is constructed by employing a weighted summation of the outputs from the encoder-decoder branch and the dilated convolution network. The weights α and β are used to obtain the weighted sum of the two branches, representing the confidence scores for each model's prediction. Another branch, comprising convolutional and fully connected layers, is trained separately to predict α and β based on the input image. After feature fusion, a fully convolutional layer is employed for classifying objects of interest and background pixels. The block diagram of proposed architecture is depicted in Figure 1. In the following two sections, we have explained working of this model for two solve two different problems, i.e., lane detection from unconstrained road environments and irregular road object detection.

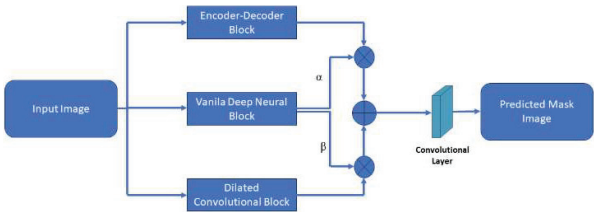

Figure 1.
The proposed semantic segmentation model.

2.1. Lane Detection from Unstructured Road Conditions

Lane detection is apparently a solved problem. However, there are a plethora of challenges in implementing it for real-life applications. The challenges in this problem include handling different road conditions, such as curved and straight roads, shadows, occlusions, and varying lighting conditions. These difficulties are addressed by creating an unstructured, challenging lane dataset. In this context, unstructured road scenarios exhibit ill-defined lane markings, ambiguous road boundaries, a diverse range of traffic participants, and variations in ambient conditions. The India Driving Dataset (IDD) [10] fulfils the requirements for an unstructured road scenario. The Indian Lane Dataset (ILD) was prepared from the India Driving Dataset (IDD) Segmentation (IDD 20k Part II) dataset [10] by making modifications. Computer Vision Annotation Tools (CVAT) [11] were employed to annotate the lanes manually. Lane markings were labeled and assigned indices that increased

Table 1.

Comparing performance of different object detection models on IDD dataset				
Models	Single-stage Object Detection	Two-stage Object Detection	Transformer-based Object Detection	Semantic Segmentation-based Object Detection
Advantages	– Simpler architecture – Faster inference speed – Fewer hyperparameters – Better for real-time applications	– Higher accuracy – Robust to occlusions – Handling small objects – Strong performance on large-scale datasets	– Better handling of global context – Adaptive receptive fields – Enhanced attention mechanisms – Can capture long-range dependencies	– Precise object boundaries – Accurate pixel-level labelling – Good for highly textured objects – Good for segmenting instances of same class
Disadvantages	– Lower accuracy – Limited handling of occlusions – Difficulty with small objects – May miss objects at different scales	– Slower inference speed – More complex architecture – More hyperparameters – Sensitivity to initialization	– Higher computational cost – Longer training time – Memory-intensive – Large model size	– Segmentation-based methods involve dense prediction, leading to slower inference – Challenging for objects with fine details – Semantic segmentation may struggle to detect small or heavily occluded objects – Requires additional memory to store per-pixel labels

Table 2.

Comparing performance of different lane detection models on ILD dataset by using IoU score						
Model	Low Light	Shadow	Curve	Highway	Normal	mIoU
DLF model	0.046	0.062	0.082	0.117	0.063	0.072
LaneNet	0.028	0.075	0.098	0.136	0.077	0.082
SCNN	0.022	0.130	0.153	0.187	0.126	0.124
RESA	**0.050**	0.192	0.269	0.332	0.228	0.214
Our model	0.044	**0.248**	**0.433**	**0.535**	**0.281**	**0.308**

Figure 2.

Performance of the model in detecting lane. Cloclwise: Input image, ground truth label, predcited mask image, superimposed predicted mask on original image.

from left to right. In cases where lane markings were not visible or absent, road boundaries and road dividers were annotated as lanes. The IDD dataset was analyzed and categorized into subgroups based on lighting and road conditions, such as no lane, low light, shadow, highway, curve, and normal. Finally, a labelled dataset of 614 images were created and named as Indian Lane Dataset (ILD). The proposed segmentation model is structured for single class segmentation, i.e., 1 for lane and 0 for non-lane pixels.

The proposed model was evaluated individually on five categories (low light, shadow, curve, highway, and normal) in ILD test images and was compared with four other models, including Discriminative Loss Function (DLF) based model [12], RESA [9], SCNN [13] and LaneNet [17]. A detailed comparison chart explaining performance between the proposed model and other models is reported in Table 2. It was found that irrespective of the road scenario, the proposed model achieved 42% improvement over RESA [9]. However, Intersection over Union (IOU) score was lower than RESA by 1% in low-light conditions. In this context, IoU measures the overlap between predicted and ground truth bounding boxes or segmentation masks.

2.2. Detection of Non-Conventional Traffic Participants

Another case study explores the detection of novel road participants, including pedestrians, autorickshaw, different type of trucks, motorcycles and riders in the challenging Indian road

Table 3.

Comparing performance of different object detection models on IDD dataset			
Models	F1 Score	Mean IoU	Latency (FPS)
YOLOv3	0.538	0.356	46.380
YOLOv4	0.403	0.267	30.616
YOLOv5	0.5341	0.355	29.858
YOLOv6	0.468	0.325	45.601
YOLOv7	0.557	0.393	60.259
RetinaNet	0.410	0.290	6.180
Mask RCNN	0.291	0.226	0.602
DeTR	0.356	0.234	0.355
UNet	0.134	0.070	6.250
I-ROD	0.478	0.363	11.638

environment. The unique characteristics of Indian roads, such as congestion, a diverse range of vehicles and transportation methods, and variations in object appearances, pose challenges for accurate detection. Initially, three state-of-the-art object detection models (Mask R-CNN, RetinaNet, YOLOv3) were compared, with YOLOv3 demonstrating superior performance in terms of both accuracy and latency [15]. The proposed segmentation model is modified to address the multi-class semantic segmentation problem.

The proposed model aims to detect a diverse range of road users in unconstrained road environments like in India. It should be trained and evaluated in this environment to make the model robust and accurate. India Driving Dataset [10] fulfills all these criteria as it covers a diversity of vehicles and pedestrians, ambient conditions, and so on. The proposed model is evaluated against nine other models, including two-stage models (Mask RCNN), single-stage models (YOLOv3, YOLOv4, YOLOv5, YOLOv6, YOLOv7, Retinanet), a transformer-based model (DETR), and a segmentation-based model (UNet). The evaluation metric used for comparison is the average Intersection Over Union (IOU) score. Table 3 provides a summary of the results, including mean IOU, F1 score, and latency. All models are tested on the same system (NVIDIA GeForce RTX 2070 GPU) to ensure consistent reporting of accuracy and latency. While YOLOv7 demonstrated better accuracy compared to proposed model, it is worth noting that researchers [14] argue in favor of pixel-wise predictions over bounding box-based predictions. Pixel-wise predictions assess the probability of an object's presence at each pixel, enabling more precise localization and segmentation, even in dense and cluttered scenes. They are also more effective at handling non-rigid objects and complex object shapes. In light of these arguments, this study aims to compare YOLOv7 and proposed model on a pixel-wise basis. To obtain pixel-wise segmented results from YOLOv7, the predicted bounding boxes (indicated by the green box in the YOLOv7 output image in Figure 3) are considered and converted into a segmented image. These segmented images are then compared with ground truth label images. The study reveals that the accuracy (IoU) drops from 0.55 to 0.39 when compared IoU of 0.45 proposed segmentation model. Figure 3 provides a detailed explanation of the comparison.

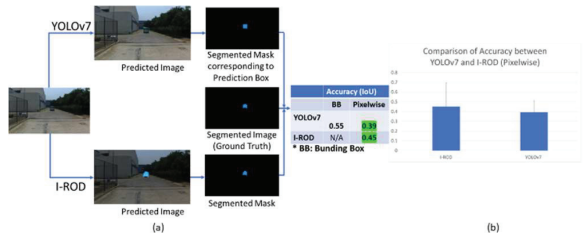

Figure 3.
A Comparison study between YOLOv7 and proposed model to show superiority of semantic segmentation models. This figure is best viewed in electronic form.

3. Application: Automatic Taxiing of Aircraft

The proposed lane and object detection model was successfully applied in the development of a fully autonomous system for aircraft taxiing. This system relied on the lane detection model as a backbone for lane navigation, while the object detection model played a crucial role in collision avoidance. Liu et al. [18] pursued a methodology akin to ours, framing the issue within a controlled simulation environment. This system use a lane and obstacle detection algorithm benchmarked against state-of-the-art models before implementation in our system. The experimental arrangement features a model replicating a taxiway, incorporating a light and lane indicator, as shown in supplementary video.[1] The green light guides the path toward the runway or hangar, while the red light serves as a stopping boundary for the aircraft. For this study, the TurtleBot3 serves as a substitute for the actual aircraft. This programmable Robot Operating System (ROS)-based mobile robot was chosen due to its ease of control via ROS and its capacity for sensor attachment, such as a camera. A Microsoft Livecam, affixed to the TurtleBot3, provides a vertical field of view (FOV) of 37.73 degrees, a horizontal FOV of 57.88 degrees, and covers distances up to 35 cm in the frontal direction.

For lane navigation, the system employed lane detection followed by a lane navigation algorithm to ensure that the aircraft stays on the center line and halts before entering the runway. The lane detection model provided points that were used in the control generation algorithm to generate inputs for autonomous navigation of the aircraft. The algorithm utilized an image mask containing only the detected lane points obtained from the lane detection algorithm. The 'steerBias' parameter indicated the distance between the camera's center and the middle lane in the x direction. If the center lane was positioned to the left of the camera center, the error was considered positive; otherwise, it was negative.

To prevent collisions, the object detection model was fine-tuned using an airport dataset before integrating it into the proposed system. The dataset consisted of six classes of objects, including airplanes, buses, cars, other vehicles, persons, and trucks. The "other vehicles" category encompassed dollies, pushback tugs, and tractors. The overall accuracy (mean IoU) of the object detection model was reported as 0.262, with a processing speed of 8.742 frames per second. Tables 4 and 5 provide a summary of the model's performance in the system.

Table 4.

Accuracy analysis of the object detection model for Automatic taxiing (class wise)			
Classes	IoU	Precision	F1 Score
Airplane	0.306	0.694	0.455
Bus	0.356	0.427	0.422
Car	0.277	0.539	0.405
Other Vehicles	0.078	0.413	0.156
Person	0.238	0.515	0.399
Truck	0.270	0.934	0.436

Table 5.

Accuracy analysis of the object detection model for Automatic taxiing (overall)	
Observed Units	Hybrid Models
mIoU	0.262
Precision	0.587
Recall	0.298
F1 Score	0.395
Latency (in milliseconds)	114.348

4. Discussions and Conclusions

This article makes significant contributions to the field of object detection research, focusing on both novelty and practical applications. A novel approach for lane detection in unconstrained environments is proposed, demonstrating real-time performance and outperforming other models. A new lane dataset is introduced and compared with existing models, showcasing the proposed model's superior accuracy and robustness across unseen environments, making it suitable for various applications. Further details of this work can be found in [16]. The article also addresses the limitations of state-of-the-art bounding box-based models in precise object localization, particularly in dense and crowded environments. To overcome this challenge, an in-depth study is conducted, leading to the proposal of a new object detection model with pixel-wise localization. This model exhibits improved performance in critical scenarios where precision is crucial. Moreover, the article presents a novel automated system for aircraft taxiing that integrates the lane and object detection algorithms to provide collision avoidance and real-time assistance. The navigation and collision avoidance system demonstrate efficacy in different lighting conditions and complex scenarios.[2] To summarize, this article contributes novel approaches to lane detection and object detection, along with the development of an automated system for taxiing aircraft, showcasing the practical applications of the proposed algorithms in real-time situations.

[1] https://www.youtube.com/watch?v=A_qG_a5w7lc.

[2] https://www.youtube.com/watch?v=A_qG_a5w7lc.

References

[1] L. G. Roberts. Pattern recognition with an adaptive network. In in: Proc. IRE International Convention Record, pages 66–70, 1960.

[2] James T Tippett, David A Borkowitz, Lewis C Clapp, Charles J Koester, and Alexander Vanderburgh Jr. Optical and electro-optical information processing. Technical report, Massachusetts Inst of Tech Cambridge, 1965.

[3] Richard Szeliski. Computer vision: algorithms and applications. Springer Nature, 2022.

[4] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. Proceedings of the IEEE, 2023.

[5] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. International journal of computer vision, 128:261–318, 2020.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017.

[7] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495, 2017.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[9] Tu Zheng, Hao Fang, Yi Zhang, Wenjian Tang, Zheng Yang, Haifeng Liu, and Deng Cai. Resa: Recurrent feature-shift aggregator for lane detection. arXiv preprint arXiv:2008.13719, 2020.

[10] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1743–1751. IEEE, 2019.

[11] Intel Community. Computer vision annotation tool (cvat). https://github.com/openvinotoolkit/cvat/.

[12] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. arXiv preprint arXiv:1708.02551, 2017.

[13] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[14] Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. In Computer vision and pattern recognition, volume 1804, pages 1–6. Springer Berlin/Heidelberg, Germany, 2018.

[15] Abhishek Mukhopadhyay, Imon Mukherjee, and Pradipta Biswas. Comparing cnns for non-conventional traffic participants. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings, pages 171–175, 2019.

[16] Abhishek Mukhopadhyay, L. R. D. Murthy, Imon Mukherjee, and Pradipta Biswas. "A hybrid lane detection model for wild road conditions." IEEE Transactions on Artificial Intelligence, 1–10, 2022.

[17] Z. Wang, W. Ren, and Q. Qiu, "Lanenet: Real-time lane detection networks for autonomous driving," arXiv preprint arXiv:1807.01726, 2018.

[18] C. Liu and S. Ferrari, "Vision-guided planning and control for autonomous taxiing via convolutional neural networks," in AIAA Scitech 2019 Forum, p. 0928, 2019.

Biographies

Abhishek Mukhopadhyay is working as Post Doctoral Researcher in I3D lab at Centre for Product Design and Manufacturing, Indian Institute of Science Bangalore. He completed his PhD in Computer Science and Engineering from IIIT Kalyani. He works in the area of Computer Vision, Object Detection, Machine Learning, Virtual Reality, HCI. He worked with Robert Bosch, Wipro, Faurecia, British Telecom in his research tenure. He achieved one best paper award (47th WWRF meeting, Bristol, 2022), ACM global fellowship to present his work in AutomotiveUI 2019 Doctoral Colloquium in the Netherlands, student travel grant in WWRF 48 in Abu Dhabi. Before joining PhD curriculum, he worked as assistant professor for 6 years in multiple Institutions in Kolkata, India.

Pradipta Biswas is an Associate Professor at the Centre for Product Design and Manufacturing and associate faculty at the Robert Bosch Centre for Cyber Physical Systems of Indian Institute of Science. He is a vice chairman of ITU Study Group 9 and also a Co-Chair of the IRG AVA and Focus Group on Smart TV at International Telecommunication Union. His research focuses on user modelling and multimodal human-machine interaction for aviation and automotive environments and for assistive technology. Earlier, he was a Senior Research Associate at Engineering Department, Research Fellow at Wolfson College and Research Associate at Trinity Hall of University of Cambridge. He completed Ph.D. in Computer Science at the Rainbow Group of University of Cambridge Computer Laboratory and Trinity College in 2010 and was awarded a Gates-Cambridge Scholarship in 2006.