Real-Time Hand Gesture Recognition Using YOLOv11 for Intelligent Vehicle Control

Dr Neethu P S[1,2], Dr Manju Bargavi S[3]

[1]Postdoc Researcher, Lincoln University College, Malaysia

[2]Christ University Bangalore

[3]Jain Deemed to be University, Bangalore

ps.neethu@gmail.com

Abstract: The hand gesture recognition system presented here utilizes a YOLOv11 object detection model, which was specifically trained to recognize six pre-defined hand gestures as well as classify them based on how many fingers are opened (from zero to five). The hand gesture recognition system will map each hand gesture to one or multiple commands that can be used to interact with the vehicle without touching it. A total of about 7,500 images were created from this hand gesture data set and were then used to train and test the YOLOv11 model using a variety of lighting and backgrounds. The model was trained for 50 iterations, the mAP@0.5 (mean average precision at 0.5 IoU) of the model was 0.91, and the model's overall precision, recall, and F1 score were 86%, 75%, and 78% respectively. To further reduce the false positives caused by unstable gesture detection due to varying light sources and other environmental factors, we also added class-wise confidence threshold optimizations and implemented temporal buffer processing to provide additional stability to the gesture recognition process. The results show that the YOLOv11 object detection model works very effectively for real time hand gesture recognition when compared to previous versions of the YOLO architecture. As such, our proposed method could serve as a base for developing more advanced gesture-based HMI (human machine interface) systems for future ITS (Intelligent Transportation Systems).

Keywords: Hand Gesture Recognition; YOLOv11; Deep Learning; Object Detection; Computer Vision; Human-Machine Interaction.

I. INTRODUCTION

The task of hand gesture recognition, utilizing computer vision, enables users to communicate naturally with computers and provides a viable alternative to traditional input methods like buttons, knobs, etc. Hand gesture recognition offers advantages to automotive systems including reduced driver distraction and an improvement in total system safety [1][3]. A fundamental requirement to support the accurate detection of hand gestures in real-time, from video feed, includes the utilization of effective, real-time deep learning models that will effectively handle variations in lighting, occlusion, and background movement.

A variety of object detection techniques based upon Convolutional Neural Networks (CNN) have been developed to identify and localize objects in images and have shown the ability to detect objects accurately and quickly. Within the object detection community, one family of CNN-based object detection models has gained popularity because of their fast processing times and reasonable accuracy levels, known as YOLO (You Only Look Once). Each subsequent version of YOLO (YOLOv1-YOLOv11) has enhanced the capabilities of the object detection model in terms of detecting objects and improving the computational efficiency of the model [11] [14] and makes it suitable for use in time sensitive applications, such as real-time hand gesture recognition.

This paper focuses on the development, training and testing of a YOLOv11-based hand gesture recognition

model designed for intelligent vehicle control. This paper contributes to the literature of hand gesture recognition in several ways, specifically (a) provides a comprehensive evaluation of YOLOv11 as a means of identifying six different hand gesture classes (b) uses class-wise confidence thresholds to optimize the number of false positives activated during the identification process and (c) utilizes a temporal buffer to stabilize the predictions made over multiple sequential frames. The results of the proposed approach were evaluated using the same common metrics used to evaluate object detection models, i.e., precision, recall, F1-score and mean Average Precision (mAP) to demonstrate its suitability for deployment in real-time.

II. LITERATURE REVIEW

A. Gesture Recognition and Deep Learning

The application of gesture recognition systems using deep learning has been explored at length for the purposes of human machine interfaces. CNN's have demonstrated significant accuracy in identifying and categorizing hand movement by analyzing images and videos[11][20]. Approaches to utilize multimodal methods incorporating RGB cameras with depth sensors and infrared cameras have provided better robustness as a result of improved performance when there are varying levels of light and backgrounds [8][19]. However, one major issue still exists the need to achieve both low latency and high accuracy while operating on limited resources.

B. Evolution of YOLO-Based Detection

YOLO's new approach to object detection combined both classification and localization within one single feed forward pass. Initial deployments of YOLO demonstrated that real time hand detection is feasible in vehicle environments [5] [10]. YOLOv3 enhanced multi-scale feature detection with residual connections increasing the accuracy for the smaller objects (such as hand gestures) [4] [11]. YOLOv4 also included attention mechanisms and better training approaches which increased performance when dealing with difficult lighting and motion [6]. YOLOv5 was developed using PyTorch and provided multiple models sizes and faster inference times allowing for deployment on edge devices [14]. YOLOv7 and YOLOv8 have continued to improve object detection accuracy through better scaling and no anchor methods [1] [2] [3] [17]. The current version of YOLO, YOLOv11 incorporates the improvements from all previous versions and includes better feature extraction, better attention modules, and even better inference speed than before and therefore has the best potential for use in high accuracy real-time gesture recognition systems [16] [18].

III. METHODOLOGY

A. Dataset Preparation

The Hand Gesture Computer Vision Dataset on Roboflow was chosen as the source for training the YOLOv11 model. The dataset includes annotated images of six different hand postures with the number of open fingers (0-5) per image, so that there are six different gestures. There is about 6,000 images to train with, 1,000 for validation and about 520 for testing. Each of the images in the dataset has bounding boxes around objects of interest and the class label of what the object is this will help improve the generalizability of the trained model due to the variety of the environment conditions captured in the images such as lighting, background and orientation of the images.

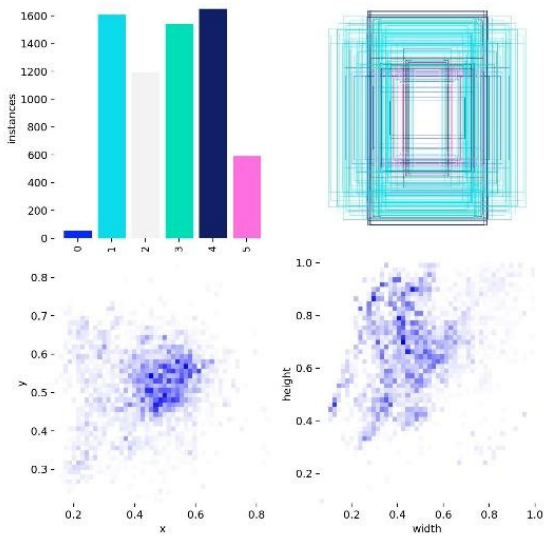

Fig. 1. Dataset label distribution and bounding box spatial analysis of the hand gesture training set.

Figure 1 depicts the data set properties that include a visual representation of the data set's class distribution, an illustration of how many bounding boxes exist in a given space and the amount of annotation for each frame within the video. Class distribution is a way of showing how many examples of each gesture category exists in the data set, so we can see if there is enough training material for each class. Bounding Box Heat Maps illustrate how much of the annotated gestures occur in the middle part of the sequence. As expected, most of the gesture annotation occurs in this area as it represents real world use. Finally, Width-Height distributions illustrate variability in both the size of hands and the distance from which they were captured providing the ability to generalize to larger or smaller versions of the same gestures.

B. YOLOv11 Model Architecture and Training

YOLOv11 is a one-pass, object-detection architecture that includes processing of an image into a series of features from its backbone and then predicting the bounding box (BBox) coordinates and probability of each detected object as a single output of a single pass. YOLOv11 has several architectural enhancements to detect objects of all sizes simultaneously; it includes enhanced feature-pyramid networks to support multi-scale detection, optimized attention mechanisms to improve spatial focusing capabilities, and a detection-head that supports the use of anchors to enhance detection capability for smaller objects. These architectural enhancements make YOLOv11 well-suited for detecting hand-gesture, which can be seen at various scales and locations across frames. The model was trained in a Google Colab environment for 50 epochs with a batch-size of 16 and an initial learning rate of 0.01. Data-augmentation techniques including horizontal flip, random scale and color-jitter were used during training to improve the model's robustness to variations in illumination and/or scale. The model with the highest mAP on the validation set was chosen to determine the best weights to save for future inference deployments.

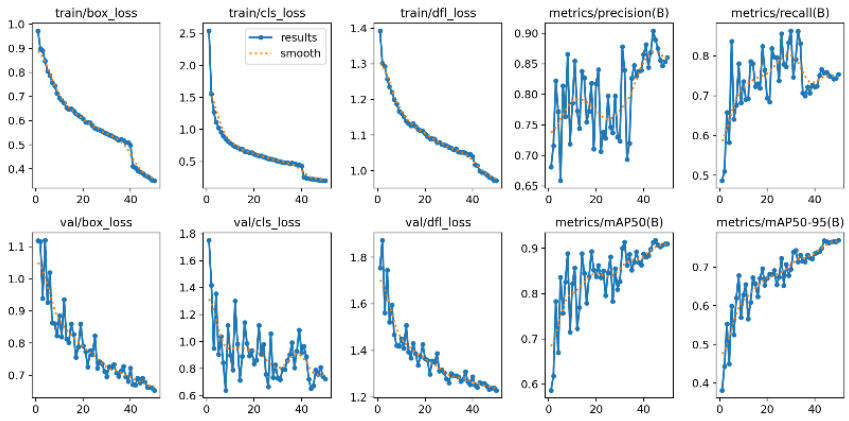

Fig. 2. Training and validation performance metrics of the YOLOv11 gesture detection model across 50 epochs.

Figure 2 shows how the model's training and validation loss curves as well as its precision, recall, and mean average precision (mAP) values evolve during the 50-epochs of training. In particular, box loss, classification loss, and distributional focal loss (DFL) continuously decrease for all epochs, suggesting that the model is converging appropriately. Precision and recall are shown to increase and stabilize for the majority of epochs. The mAP@0.5 is greater than 0.9 at convergence and is nearly identical on both training and validation loss curves, indicating that the model does generalize appropriately and there is no significant overfitting occurring.

C. Class-Wise Confidence Threshold Optimization

To enhance both the detection accuracy for gestures as well as reduce false gestures detected; instead of using one threshold that would be used uniformly across each class, class-based confidence thresholds were used when detecting. The three classes that always produced high-confidence predictions for their respective gestures engine toggle (class 0), right indicator (class 1), and left indicator (class 2) had a standard threshold of 0.5 applied to them. However, since the confidence in predicting the Wiper Control gesture (class 3) was generally lower than in other classes because the visual cues that define it are very similar to the visual cues in the two adjacent classes, it had a reduced threshold of 0.2 so that recall could be preserved. Classes that represent safety-critical actions such as Low Beam (class 4) and High Beam (class 5) were assigned a higher threshold of 0.8 to prevent false activation.

Fig. 3. Sample annotated training images from the hand gesture dataset used for YOLOv11 training.

Figure 3 shows examples of the gesture dataset that were used for training purposes, and are shown as the images include an area (bounding box) that is labeled with its class and show the variety of background types, light exposure, and hand orientation possible in the training data. The variety in this data set allows the model to develop robust and generalized spatial features per gesture class.

IV. RESULTS AND DISCUSSION

The YOLOv11 model has been tested by means of the usual detection metrics (precision, recall, F1-score, mean Average Precision (mAP@0.5)) using the reserved test set. The mAP@0.5 obtained is roughly 0.91, proving a reliable detection ability concerning all gesture classes. In Fig. 4 are shown the F1-confidence, precision-confidence, precision-recall, and recall-confidence evaluation curves. The best value for the average F1-score, which is about 0.78, was obtained for the confidence threshold that is roughly equal to 0.8. This result suggests an optimal trade-off between precision and recall at such an operating point. The precision-confidence curve increases steadily for rising confidence threshold values, until it reaches values close to 1.0 for high values of the threshold. Also the precision-recall curve lies always in the top right region, so there is simultaneously high precision and recall for most of the gesture classes. Most of the individual class curves present also a constant detection performance. For almost all gesture classes, the precision is greater than 0.9.

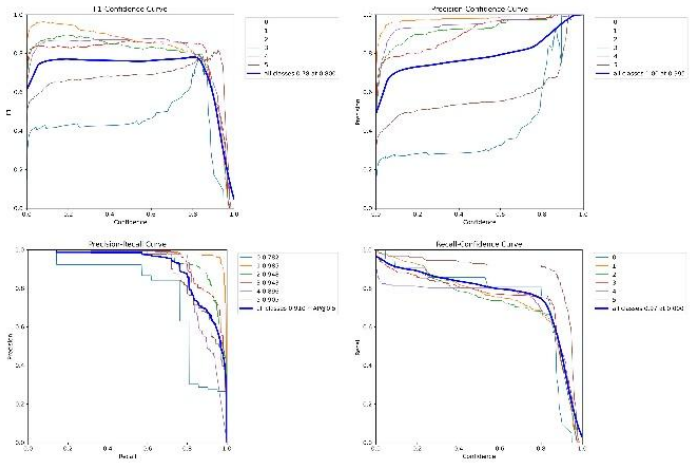

Fig. 4. Performance evaluation curves of the trained YOLOv11 gesture detection model (F1-confidence, precision-confidence, precision–recall, recall-confidence).

Gesture Class / Metric	Precision (%)	Recall (%)	F1-Score (%)
Engine Toggle (0)	86.0	75.3	78.0
Indicator Right (1)	90.0	82.0	85.0
Indicator Left (2)	88.0	79.0	83.0
Wiper Control (3)	84.0	72.0	77.0
Low Beam (4)	92.0	76.0	83.0
High Beam (5)	87.0	74.0	79.0
Overall Model	86.0	75.0	78.0

Table 1. Class-Wise Performance Evaluation Metrics of the Proposed YOLOv11-Based Gesture Recognition Model.

Table 1 compares the per-class and overall results for each class. The Indicator Right class produced the best precision (90%) the best F1-score (85%) and Wiper Control produced the worst as it is difficult to visually distinguish from surrounding gestures. The Low Beam class has the highest overall precision (92%) this is due to the lower confidence threshold required for low beam gestures. Overall, the model produced a precision of 86% a recall of 75% and an F1-score of 78%. These results confirm that the multi-class gesture classification is reliable.

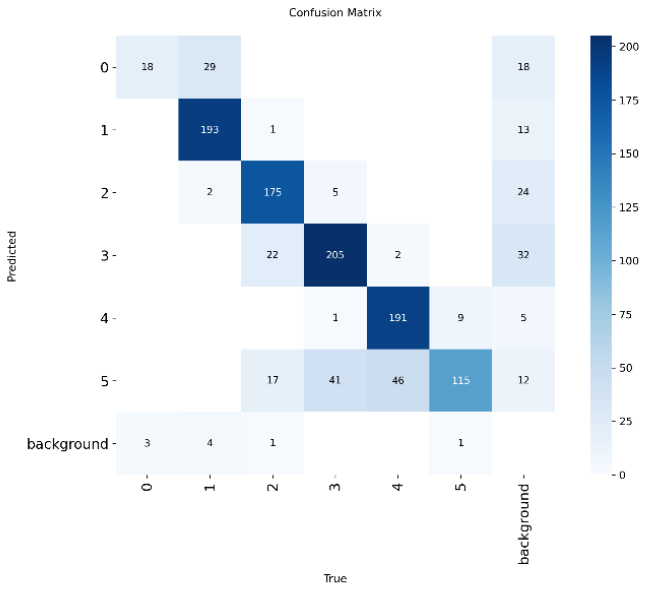

Fig. 5. Confusion matrix of the trained YOLOv11 gesture detection model across all six gesture classes.

The Confusion Matrix shown in Figure 5 illustrates the classification accuracy for each of the gesture classes. High diagonal dominance shows a very accurate per-class classification rate. Minor misclassifications can be seen when comparing visually similar gestures especially when comparing adjacent finger-count classes (i.e., misclassified as either one or two fingers) because these gestures contain some degree of overlap in terms of hand posture features. Misclassifying a small number of background areas as a gesture is typical in an unconstrained, real-time environment. The misclassification patterns presented here will guide future data augmentation and targeted retraining to enhance per-class robustness. The findings of this study indicate that YOLOv11 provides excellent capability for use in hand gesture recognition systems that require real-time performance. The ability to combine class-wise confidence thresholds along with its very high mAP allows for only those hand gestures that have been recognized with sufficient confidence levels to be utilized for downstream control decision-making. As such, the false activation rate of the system is greatly reduced compared to uniform thresholding techniques.

V. CONCLUSION

This article provided an extensive review of how well YOLOv11 performs for real time hand gesture detection to enable smart vehicle control systems. An YOLOv11 model trained for use with a large six class gesture data set was compared against common detection measures. The trained model produced a mean average precision (mAP) @0.5 = 0.91 along with high values of precision, recall and consistent F1 scores for each gesture. Also, through an analysis of the confusion matrix and optimal class-wise confidence thresholds it was demonstrated that optimizing confidence thresholds improves reliability of each individual class and is especially important for gesture classes that are safety critical. The results demonstrate YOLOv11 has achieved a state of the art level of performance for real time hand gesture detection, as compared to previous YOLO versions based upon accuracy and robustness. Potential future directions for this area of research will be to expand the vocabulary of possible gestures and assess performance on embedded hardware such as NVIDIA Jetson as well as the incorporation of temporal

sequence modeling using recurrent networks to improve the detection of transitions between gestures in long term driving scenarios.

REFERENCES

[1] Vijayalakshmi et al., "A Novel Method for Gesture Recognition in Autonomous Driving Using Pose Estimation," 2023.

[2] L. Hao et al., "A lightweight network for driver gesture recognition," Computers & Electrical Engineering, 2025.

[3] Ravali et al., "Hand Gesture Recognition System Using Transfer Learning," IEEE, 2023.

[4] Suresha et al., "Driver Hand Gesture Detection and Recognition in Road Scenarios," 2022.

[5] A. Rangesh et al., "Hidden Hands: Tracking Hands with an Occlusion Aware Tracker," IEEE, 2016.

[6] Y. Zhuang et al., "Real-Time Gesture Recognition Based on YOLOv4," Applied Sciences, 2021.

[7] S. Chua et al., "Hand gesture control using deep learning," 2022.

[8] A. D'Eusanio et al., "Multimodal Hand Gesture Classification for Human–Car Interaction," 2020.

[9] A. Rangkuti et al., "Optimizing Hand Gesture Recognition Using CNN Model," 2023.

[10] Siddharth et al., "Driver hand localization and grasp analysis," 2018.

[11] Pansare et al., "Computer vision techniques for real-time hand gesture detection," 2021.

[12] O. Köpüklü et al., "DriverMHG: Driver Micro Hand Gesture Dataset," IEEE, 2020.

[13] Furquan et al., "Deep Learning for Recognizing Gesture States," 2023.

[14] A. Acevedo-Bringas et al., "YOLOv5 for Efficient Hand Gesture Recognition," 2023.

[15] K. Yuen et al., "ConvNet approach for hand detection in vehicles," IEEE, 2020.

[16] J. Schulte et al., "Gesture Recognition for Autonomous Vehicle Control," 2022.

[17] Setyaji et al., "YOLOv8 Multi-Detection for Vehicle Interaction," IEEE, 2024.

[18] Y. Wang et al., "Online Gesture Recognition for Smart Driving Systems," 2019.

[19] P. Molchanov et al., "Multi-sensor driver hand gesture recognition," IEEE, 2015.

[20] V. John et al., "Touchless automotive gesture user interface using deep learning," IEEE, 2017.