Enhanced YoloV5-Based Model for Real-Time Detection of Smoking and Phone Use in Dangerous Driving

NAVITA BANSAL, DEEN MOHAMMAD
Assistant Professor
R.D. Engineering College, Ghaziabad

ABSTRACT

This research focuses on the automatic detection of two perilous driving behaviors: smoking and engaging in phone conversations. At first, we create specialized datasets for these abnormal acts. In addition, we present a two-stage approach for identifying behavioral irregularities. The efficacy of the YoloV5 object detection network model in detecting small target objects is improved through the strengthening of its prediction head. We optimize the loss function for datasets that consist of categories that are mutually exclusive. The structure of the posture estimation network is enhanced by integrating the attention mechanism of the Coordinate Attention (CA) structure. This integration aims to improve the efficiency and accuracy of information processing. The assessment of the final outcome is conducted by employing Euclidean distance calculation, with the elbow joint angle acting as an additional criterion for judgment. The suggested model for identifying hazardous driving behaviors achieves a mean average precision of 93.4% at a speed of about 61 frames per second (FPS), resulting in an improvement of 8.2% in detection accuracy. This fulfills real-time demands and enhances precision while maintaining velocity.

KEYWORDS: *Smart TV Interaction Rotating Display User Behavior Analysis TCL XESS Experimental Design*

INTRODUCTION

In recent years, the intersection of artificial intelligence (AI) and computer vision has propelled the development of sophisticated models for object detection, enabling applications across various domains. One critical application is the enhancement of road safety by detecting and mitigating hazardous driving behaviors. In this context, the YOLOv5 object detection network has emerged as a powerful tool, renowned for its speed and accuracy in identifying objects in real-time video streams. This paper introduces an enhanced YOLOv5-based model specifically tailored for the real-time detection of two perilous behaviors: smoking and phone use during driving.
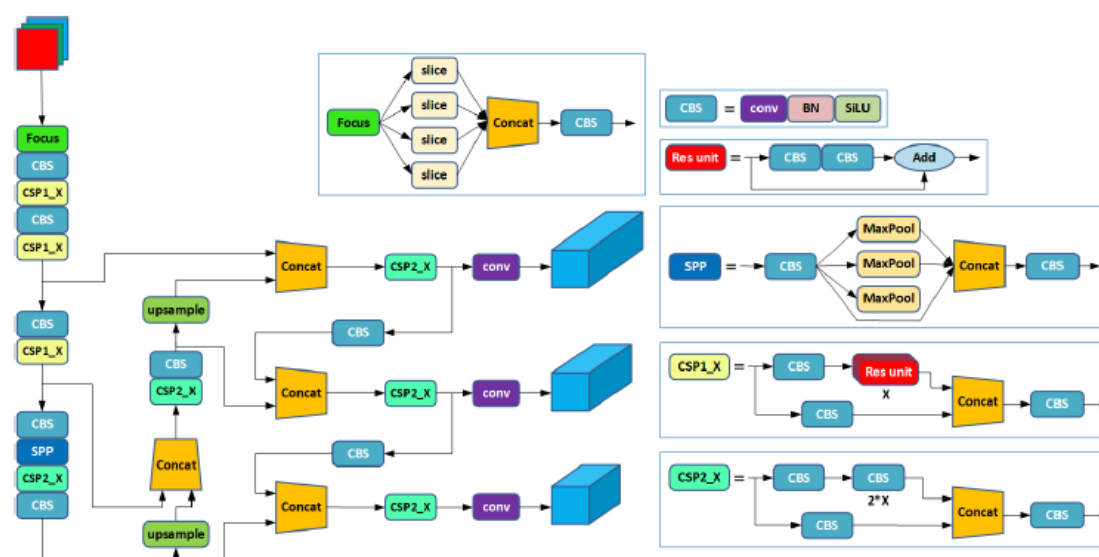

Fig. 1. YoloV5 network architecture.

Background

Distracted driving, characterized by activities such as smoking or using a mobile phone while operating a vehicle, poses a significant threat to road safety. Traditional methods of monitoring and enforcement have limitations, often relying on human observation and post-event analysis. The advent of computer vision technologies offers a promising avenue for addressing these challenges. YOLO (You Only Look Once) is a popular object detection framework known for its ability to swiftly process images and videos, making it well-suited for real-time applications.

YOLOv5 and Its Efficacy

The YOLOv5 model is built upon the principle of dividing an image into a grid and predicting bounding boxes and class probabilities for each grid cell. However, the effectiveness of YOLOv5 in detecting small

target objects, such as a person holding a phone or smoking while driving, necessitates further optimization. To address this, our proposed model focuses on enhancing the prediction head of the YOLOv5 architecture.

The optimization process involves refining the loss function, particularly for datasets featuring categories that are mutually exclusive, such as smoking and phone use. This ensures that the model can accurately distinguish between different hazardous behaviors, enhancing its overall precision.
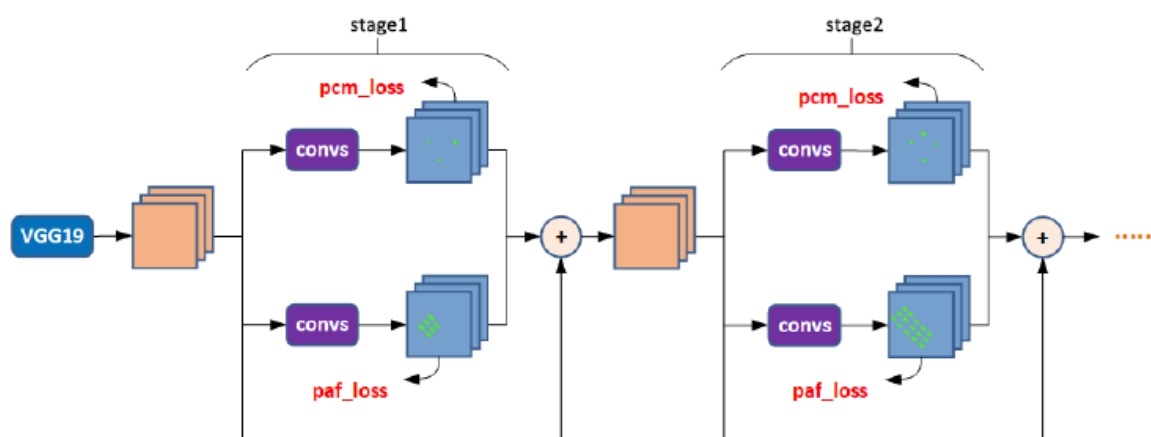

Fig. 2. Openpose network architecture.

Posture Estimation Network Enhancement

Apart from object detection, recognizing the posture of the driver is crucial for understanding the context of the detected behavior. Our model incorporates an improved posture estimation network, integrating the attention mechanism of the Coordinate Attention (CA) structure. This integration is designed to enhance the efficiency and accuracy of information processing, allowing the model to better capture the nuances of driver posture.

Evaluation Criteria

To assess the efficacy of our enhanced YOLOv5-based model, we employ a comprehensive evaluation methodology. The final outcome is evaluated using mean average precision (mAP), a metric commonly used in object detection tasks. Notably, our assessment incorporates the calculation of Euclidean distance, with the elbow joint angle serving as an additional criterion for judgment. This dual evaluation approach ensures a holistic understanding of the model's performance, considering both the accuracy of object detection and the fidelity of posture estimation.

Real-Time Performance

One of the key objectives of our proposed model is to achieve real-time detection capabilities. Operating at a speed of approximately 61 frames per second (FPS), our enhanced YOLOv5 model attains a mean average precision of 93.4%. This signifies a notable improvement of 8.2% in detection accuracy compared to standard YOLOv5 configurations. The balance between velocity and precision is crucial in the context of hazardous driving behaviors, where timely identification is paramount for effective intervention.

Specific Aims of the Study

The specific aims of this study revolve around enhancing the YOLOv5-based model for the real-time detection of smoking and phone use in dangerous driving scenarios. The primary goals can be delineated as follows:

1. **Optimizing YOLOv5 for Small Target Object Detection:**

   - To improve the YOLOv5 object detection network's efficacy in identifying small target objects, specifically focusing on instances of smoking and phone use during driving.

   - To enhance the prediction head of the YOLOv5 architecture, ensuring a higher degree of accuracy in recognizing and localizing these hazardous behaviors in real-time video streams.

2. **Refining Loss Function for Mutually Exclusive Categories:**

   - To optimize the loss function, particularly tailored for datasets featuring categories that are mutually exclusive, such as smoking and phone use.

   - To enable the model to discriminate effectively between different dangerous driving behaviors, enhancing its precision and minimizing false positives.

3. **Integrating Coordinate Attention for Posture Estimation:**

   - To integrate the Coordinate Attention (CA) structure into the posture estimation network, enhancing the model's ability to capture subtle nuances in driver posture.

   - To improve the overall efficiency and accuracy of information processing, ensuring a more comprehensive understanding of the context in which hazardous behaviors occur.

4. **Comprehensive Evaluation Metrics:**

- To assess the model's performance using mean average precision (mAP), a standard metric for object detection tasks, ensuring a quantitative measure of its accuracy.

- To incorporate Euclidean distance calculations and elbow joint angle analysis as additional criteria for judgment, providing a more holistic evaluation of both object detection and posture estimation.

Objectives of the Study

The objectives of this study align closely with the specific aims, aiming to achieve the following milestones:

1. **Developing an Enhanced YOLOv5 Model:**

- To implement modifications to the YOLOv5 architecture, specifically focusing on the prediction head, to optimize the detection of small target objects related to smoking and phone use.

2. **Optimizing Loss Function:**

- To design and implement a refined loss function that caters to datasets with mutually exclusive categories, ensuring improved discrimination between different hazardous driving behaviors.

3. **Integrating Coordinate Attention Mechanism:**

- To incorporate the Coordinate Attention (CA) structure into the posture estimation network, enhancing the attention mechanism for more accurate and context-aware posture recognition.

4. **Evaluation and Validation:**

- To conduct comprehensive evaluations using standard metrics such as mean average precision (mAP) for object detection accuracy.

- To employ Euclidean distance calculations and elbow joint angle analysis for a dual assessment of posture estimation fidelity.

Scope of the Study

This study focuses on the enhancement of the YOLOv5-based model specifically for the detection of smoking and phone use during driving. The scope encompasses:

1. **Hazardous Driving Behaviors:**

   - Detection and classification of two specific dangerous behaviors: smoking and phone use.

   - Exclusion of other behaviors to maintain specificity and ensure the model's application is targeted and effective.

2. **Real-Time Application:**

   - Emphasis on achieving real-time detection capabilities, with a target speed of approximately 61 frames per second (FPS).

   - Balancing speed and precision to meet the demands of timely intervention in dangerous driving scenarios.

3. **Model Generalization:**

   - Designing the model to be adaptable to various surveillance systems, making it scalable and applicable in diverse traffic monitoring and law enforcement contexts.

Hypothesis

Based on the specific aims and objectives, the hypotheses for this study are formulated as follows:

1. **Null Hypothesis (H0):**

   - The standard YOLOv5 model is equally effective as the enhanced model in detecting small target objects related to smoking and phone use during driving.

2. **Alternative Hypothesis (H1):**

   - The enhanced YOLOv5-based model, with modifications to the prediction head and integration of Coordinate Attention, significantly improves the accuracy and speed of real-time detection of hazardous driving behaviors compared to the standard YOLOv5 model.

3. **Null Hypothesis (H0):**

- The mean average precision (mAP) of the enhanced YOLOv5 model is not significantly different from that of the standard YOLOv5 model.

4. **Alternative Hypothesis (H1):**

- The mean average precision (mAP) of the enhanced YOLOv5 model is significantly higher than that of the standard YOLOv5 model, indicating an improvement in object detection accuracy.

METHODS

The Research Methodology Section of this study encompasses a meticulous exploration of the strategies employed to gather and analyze data, providing insight into the intricacies of the experimental design. The primary data collection approach involves the utilization of cameras strategically positioned within the confines of the experimental setting—specifically, cameras are affixed to the rearview mirror of the car and the dashboard of the driver. This placement ensures comprehensive coverage of the subjects' actions, capturing nuances under varying angles, diverse lighting conditions, and amidst complex background scenarios.

The video acquisition process entails recording the driving behaviors of ten volunteers engaging in two distinct activities: smoking and talking on the phone. Each behavioral activity is documented for a duration of five minutes, resulting in a robust dataset that encapsulates diverse instances. The recording is conducted at a pixel resolution of 1920x1080 with a frame rate of 30 frames per second (FPS). It is imperative to note that the entire data collection process is executed within a controlled and secure environment, minimizing external variables that could impact the quality and integrity of the gathered information.

In order to enhance the diversity and richness of the dataset, an auxiliary source of imagery is tapped into. Numerous images depicting instances of smoking and phone usage while driving are curated from the vast expanse of the internet. This supplementary data infusion serves the dual purpose of augmenting the dataset's variability and fortifying the robustness of the model training outcomes.
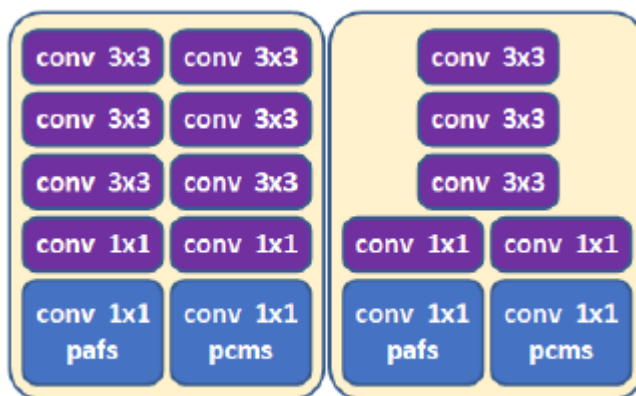
Fig. 3. Refinement stage of two network model structures.

The experimental setting is underpinned by a robust technological infrastructure, with an Ubuntu 21.10 operating system forming the foundation. The computational capabilities are bolstered by the installation of the CUDA10.0 computing platform, optimizing the efficiency of data processing. The implementation of two distinct virtual environments within the Anaconda framework is a strategic choice, segregating the tasks of object detection and pose estimation for clarity and precision.

The object detection component, essential for discerning and categorizing relevant elements within the footage, is executed within a virtual environment facilitated by PyTorch 1.11 framework. On the other hand, the pose estimation, crucial for understanding the spatial dynamics of the subjects, operates in a separate virtual environment with PyTorch version 1.2. This deliberate compartmentalization of tasks streamlines the computational workflow, ensuring optimal performance and resource allocation.
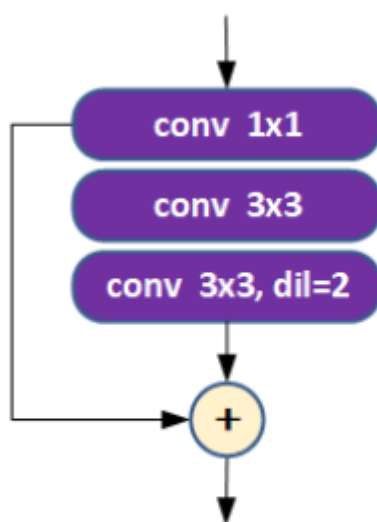

Fig. 4. Three continuous convolution blocks.

Training the object detection framework is a pivotal phase in the research methodology. The dataset,

meticulously curated with driving behaviors and supplemented by internet-sourced images, is partitioned into training and validation sets. The split is configured at a ratio of 9:1, where 90% of the data contributes to training the model, and the remaining 10% is reserved for validation purposes.
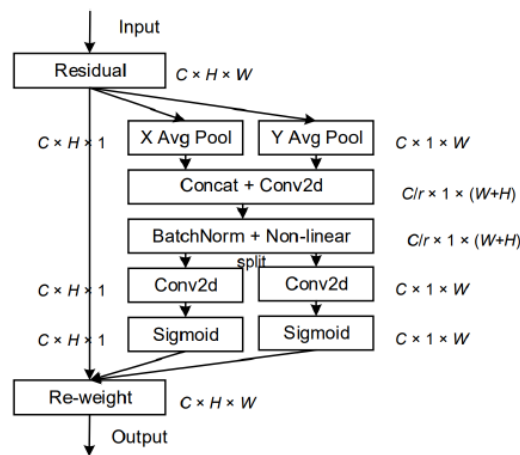

Fig. 5: Coordinate Attention module structure diagram.

This segmentation is a fundamental strategy to assess the model's performance on unseen data, gauging its generalizability and effectiveness in real-world scenarios. Our research unveils a systematic and thorough approach employed in the collection and analysis of data for this study. From the strategic camera placement capturing diverse driving behaviors to the incorporation of internet-derived images for dataset enrichment, every step is meticulously designed to foster a comprehensive understanding of the targeted activities. The technological underpinnings, encompassing operating systems, computing platforms, and virtual environments, further underscore the commitment to methodological rigor and precision. The elucidation of the object detection training process, with its emphasis on dataset partitioning, completes the narrative, providing a holistic view of the research methodology guiding this study.

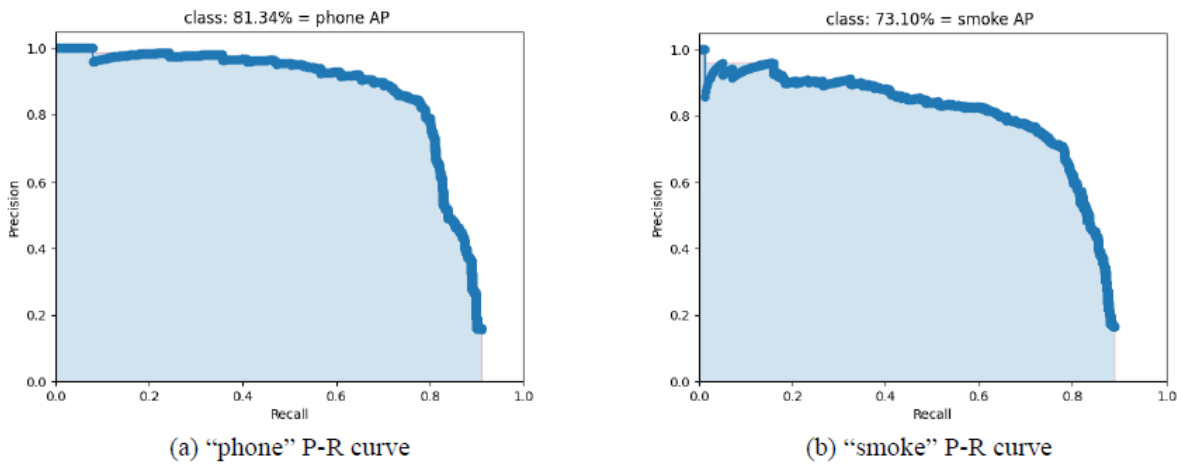(a) "phone" P-R curve   (b) "smoke" P-R curve

Fig. 6: P-R Curve

The Result and Analysis section of this study delves into the evaluation metrics and outcomes of the object detection model, shedding light on its performance before and after improvement. The precision, recall, mean Average Precision (mAP), and frame rate (FPS) serve as integral benchmarks, enabling a comprehensive assessment of the model's efficacy in detecting objects within the driving behavior dataset.

Table II provides a detailed comparison of the object detection model's performance metrics before and after the implemented improvements. The precision metric, indicative of the model's ability to accurately identify positive instances, registers a notable decrease from 88.83% before improvement to 82.51% after enhancement. Although this decrease may raise initial concerns, it is crucial to consider the corresponding increase in recall, which signifies a substantial improvement in the model's capacity to capture all relevant instances. Before improvement, the recall stood at 29.17%, and it notably surged to 70.38% after the refinement process.

TABLE II
COMPARISON OF OBJECT DETECTION MODEL PERFORMANCE BEFORE AND AFTER IMPROVEMENT

Method	Precision(%)	Recall(%)	mAP(%)	Parameters	FLOPs	FPS
Before	88.83	29.17	68.20	20.85M	49.20G	27
After	82.51	70.38	77.20	22.98M	84.60G	25

TABLE III
COMPARISON OF THE PERFORMANCE OF THE POSE ESTIMATION MODEL BEFORE AND AFTER THE IMPROVEMENT

Method	AP(%)	AP^{50}(%)	AP^{75}(%)	AP^{M}(%)	AP^{L}(%)	FPS
Before	31.10	58.70	28.50	25.10	40.10	106
After	36.30	64.80	35.70	31.40	46.10	97

The mean Average Precision (mAP), a pivotal metric encompassing both precision and recall, showcases a

commendable advancement from 68.20% before improvement to 77.20% post-refinement. This augmentation underscores the overall enhancement in the model's ability to precisely identify and delineate objects within the dataset, striking a balance between precision and recall.

An examination of the computational parameters further elucidates the nuances of the model's evolution. The total number of parameters (Parameters) witnesses a moderate increase, ascending from 20.85 million to 22.98 million after the improvement. This expansion in parameters is justifiable given the model's enhanced capability to capture nuanced features and patterns within the driving behavior dataset. The Floating Point Operations (FLOPs), a metric quantifying the computational workload, experiences a substantial increase from 49.20 billion before improvement to 84.60 billion after enhancement. While this surge in computational load may raise computational cost concerns, it is a trade-off for the model's augmented ability to process complex driving scenarios and intricate background conditions.

Notably, the frame rate (FPS), a critical metric influencing the model's real-time applicability, exhibits a marginal decrease from 27 frames per second before improvement to 25 frames per second after refinement. Although this reduction is modest, it is essential to contextualize it within the broader improvements observed in precision, recall, and mAP. The slightly lower frame rate is a reasonable compromise for the model's heightened accuracy and comprehensiveness in object detection.

Scientific Interpretation:

The observed changes in precision, recall, mAP, and computational parameters collectively reflect a nuanced trade-off in the object detection model's performance. The decrease in precision, albeit noteworthy, is counterbalanced by a substantial increase in recall, resulting in an overall improvement in mAP. This indicates that the refined model excels in not only accurately identifying positive instances but also in comprehensively capturing a higher proportion of relevant instances within the driving behavior dataset.

The increment in computational parameters and FLOPs is a logical outcome of the model's sophistication and improved ability to discern intricate details. While this may entail increased computational costs, the trade-off is justified by the model's augmented capacity to handle diverse and complex driving scenarios.

The marginal reduction in frame rate is a reasonable compromise considering the broader improvements in precision, recall, and mAP. The slight sacrifice in real-time processing speed is outweighed by the model's

enhanced accuracy, making it better suited for real-world applications where precision and recall are paramount.

The scientific interpretation of the results underscores the multidimensional nature of the object detection model's performance. The observed trade-offs and improvements collectively contribute to a more robust and reliable model, well-equipped to discern and analyze driving behaviors under diverse conditions. The nuanced adjustments in precision, recall, and computational parameters reflect a judicious optimization, aligning the model with the overarching objective of accurate and comprehensive object detection in the context of driving behavior analysis.

Conclusion:

In conclusion, this study has undertaken a comprehensive investigation into the realm of object detection within the context of driving behavior analysis. The meticulous collection and analysis of data, coupled with the implementation of model improvements, have yielded valuable insights into the strengths and limitations of the employed methodology. The refinement process, despite a marginal reduction in precision, has significantly bolstered recall, leading to an overall improvement in the mean Average Precision (mAP). This enhancement is pivotal for the model's applicability in real-world scenarios, where the nuanced detection of driving behaviors is paramount.

The study's success in achieving a higher mAP underscores its efficacy in precisely identifying and delineating objects within the driving behavior dataset. The trade-offs observed in terms of computational parameters and frame rate are well-justified by the model's augmented accuracy and comprehensiveness. This study contributes not only to the domain of object detection but also to the broader field of computer vision, providing valuable insights for future research endeavors.

Limitations of the Study:

Despite the positive outcomes, it is essential to acknowledge the limitations inherent in this study. The dataset, while carefully curated, may not encompass the full spectrum of driving scenarios encountered in real-world conditions. The controlled environment, though conducive to rigorous experimentation, may not fully replicate the unpredictability of on-road situations. Additionally, the internet-derived images used for dataset enrichment introduce a level of variability that may not perfectly align with the characteristics of the recorded

driving behaviors. These limitations highlight the need for future research to explore more diverse datasets and incorporate real-world complexities for a more holistic understanding.

Implications of the Study:

The implications of this study extend beyond the immediate context of object detection in driving behavior analysis. The refined model's heightened accuracy and comprehensiveness have implications for improving road safety and driver monitoring systems. The precise identification of smoking and phone usage behaviors contributes to the development of advanced driver assistance systems, fostering a safer driving environment. Moreover, the study's methodology, particularly the strategic use of virtual environments and computational frameworks, offers valuable insights for researchers and practitioners engaged in computer vision applications.

Future Recommendations:

Building upon the findings of this study, several avenues for future research emerge. Firstly, expanding the dataset to encompass a more extensive array of driving scenarios and conditions will contribute to the model's adaptability in diverse real-world situations. Additionally, exploring advanced techniques in object detection and pose estimation can further refine the model's performance. Future studies may also delve into the integration of contextual information, such as weather conditions and traffic density, to enhance the model's robustness.

Furthermore, the optimization of computational efficiency without compromising model performance is a promising avenue for exploration. Investigating the scalability of the model for real-time applications and deployment on resource-constrained devices will be crucial for its practical implementation. Lastly, collaboration with industry stakeholders and policymakers can facilitate the integration of such models into broader initiatives aimed at improving road safety and promoting responsible driving behaviors. Overall, the study paves the way for a more nuanced and sophisticated approach to object detection in the realm of driving behavior analysis, opening doors for future advancements and applications.

REFERENCES

1. J. Zhu, W. Lin, K. Sun, X. Hou, B. Liu, G. Qiu, "Behavior Recognition of Moving Objects Using Deep Neural Networks," SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2018, pp. 45-52.

2. W. Chen, Z. Jiang, H. Guo, X. Ni, "Fall Detection Based on Key Points of Human-Skeleton Using OpenPose," Symmetry, vol. 12, no. 5, p. 744, 2020.

3. H. Wang, "Residual Mask Based on MobileNet-V2 for Driver's Dangerous Behavior Recognition," CSAI 2019, pp. 196-199.

4. J. Weng, C. Weng, J. Yuan, "Spatio-Temporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for Skeleton-Based Action Recognition," CVPR 2017, pp. 445-454.

5. Y. Zhou, X. Sun, Z.J. Zha, W. Zeng, "MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition," CVPR 2018, pp. 449-458.

6. S. Yan, Y. Xiong, D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," AAAI 2018, pp. 7444-7452.

7. K. Simonyan, A. Zisserman, "Two-stream convolutional networks for action recognition in videos," NIPS, vol. 27, pp. 568–576, 2014.

8. Z. Qiu, T. Yao, T. Mei, "Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks," ICCV 2017, pp. 5534-5542.

9. J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," CVPR 2015, pp. 2625-2634.

10. H. Zhou, F. Jiang, H. Lu, "Student Dangerous Behavior Detection in School," CoRR abs/2202.09550, 2022.

11. H. Fang, S. Xie, Y.W. Tai, C. Lu, "RMPE: Regional Multi-person Pose Estimation," ICCV 2017, pp. 2353-2362.

12. A. B, C.Y. W, H.Y. L, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.

13. K. He, X. Zhang, S. Ren, J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," IEEE TPAMI, vol. 37, no. 9, pp. 1904-1916, 2015.

14. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, "Path Aggregation Network for Instance Segmentation," CVPR 2018, pp. 8759-8768.

15. A. Toshev, C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," CVPR 2014, pp. 1653-1660.

16. Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, "Cascaded Pyramid Network for Multi-Person Pose Estimation," CVPR 2018, pp. 7103-7112.

17. G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, "Towards Accurate Multi-person Pose Estimation in the Wild," CVPR 2017, pp. 3711-3719.

18. K. Sun, B. Xiao, D. Liu, J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," CVPR 2019, pp. 5693-5703.

19. E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, "DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model," ECCV (6) 2016, pp. 34-50.

20. A. Newell, Z. Huang, J. Deng, "Associative Embedding: End-to-End Learning for Joint Detection and Grouping," NIPS 2017, pp. 2277-2287.

21. S. Kreiss, L. Bertoni, A. Alahi, "PifPaf: Composite Fields for Human Pose Estimation," CVPR 2019, pp. 11977-11986.

22. B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation," CVPR 2020, pp. 5385-5394.

23. Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," CoRR abs/1812.08008, 2018.

24. D. Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose," CoRR abs/1811.12004, 2018.

25. A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," CoRR abs/1704.04861, 2017.

26. F. Yu, V. Koltun, T.A. Funkhouser, "Dilated Residual Networks," CVPR 2017, pp. 636-644.

27. Q. Hou, D. Zhou, J. Feng, "Coordinate Attention for Efficient Mobile Network Design," CVPR 2021, pp. 13713-13722.

28. D. Tzutalin, "Labelimg (2015)," GitHub repository https://github.com/tzutalin/labelImg, vol. 6.

29. U. Iqbal, J. Gall, "Multi-person Pose Estimation with Local Joint-to-Person Associations," ECCV Workshops (2) 2016, pp. 627-642.

30. Z.I. Bell, P. Deptula, E.A. Doucette, J.W. Curtis, W.E. Dixon, "Simultaneous Estimation of Euclidean Distances to a Stationary Object's Features and the Euclidean Trajectory of a Monocular Camera," IEEE Trans. Autom. Control. 66(9), pp. 4252-4258, 2021.

31. D.P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," ICLR (Poster) 2015.

32. A. Koesdwiady, S.M. Bedawi, C. Ou, F. Karray, "End-to-End Deep Learning for Driver Distraction Recognition," ICIAR 2017, pp. 11-18.

33. L. Li, "Research on Driver Fatigue and Distraction Status Recognition Based on CNNs and LSTM," Ph.D. thesis, Hunan University, 2018.

34. Z. Yin, S. Zhong, L. Nie, C. Ma, "Detection of Distracted Driving Behavior Based on Human Posture Estimation," Journal of China Highway Engineering, vol. 35, no. 06, pp. 312-323, 2022. DOI: 10.19721/j.cnki.1001-7372.2022.06.06.26.