International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Vision Based Navigation Assistant Using Object Detection and Depth Estimation

Mrs.S.Subha[1], Kiruthika M[2], Harrshinee L[3], Kanika V[4]
[1] Assistant Professor Department of Computer Science and Engineering Kongunadu College of Engineering and Technology Tamilnadu, India
[2,3,4] Department of Computer Science and Engineering Kongunadu College of Engineering and Technology Tamilnadu,India

Abstract- — Vision-based navigation has become increasingly important in fields such as assistive technology, robotics, and autonomous driving, as it enables systems to understand and interact with complex environments. This study introduces a Vision-Based Navigation Assistant that combines object detection with depth estimation to improve real-time awareness and navigation safety. The system utilizes deep learning techniques to detect and categorize surrounding objects while estimating their distances through monocular or stereo vision approaches. This integrated method allows the system to deliver relevant information about obstacles, pathways, and potential risks. Designed for efficiency, the framework can run on embedded devices, ensuring portability and minimal processing delay. Furthermore, it provides feedback through audio or haptic signals, making it especially useful for visually impaired individuals and autonomous systems. Experimental evaluations indicate enhanced accuracy in identifying objects and estimating distances, resulting in dependable performance across both indoor and outdoor settings. Overall, the proposed solution demonstrates the effectiveness of computer vision in developing intelligent navigation aids that enhance mobility, safety, and user independence.

Keywords: Vision-based navigation, object detection, depth estimation, assistive systems, computer vision, autonomous navigation, obstacle detection, deep learning

I. INTRODUCTION

Navigating through complex and ever-changing environments is a major challenge in domains such as assistive technology, robotics, and autonomous systems. For both visually impaired individuals and mobile robotic platforms, the capability to accurately perceive surroundings and react to obstacles in real time is crucial for safe and efficient movement. Conventional navigation tools, including white canes and basic sensor-based systems, often provide only limited environmental awareness and lack detailed contextual understanding. Recent progress in computer vision and deep learning has opened new possibilities, making vision-based navigation a powerful alternative to overcome these shortcomings.

Vision-based navigation systems utilize cameras along with intelligent algorithms to process visual information, enabling the recognition of objects, interpretation of scenes, and support for decision-making. Two essential components that enhance the effectiveness of such systems are object detection and depth estimation. Object detection focuses on identifying and categorizing elements within the environment, such as people, vehicles, and obstacles. In contrast, depth estimation determines the distance between the observer and these objects, offering crucial spatial information required for navigation.

Combining object detection with depth estimation leads to a more complete understanding of the environment than using either technique independently. This integration merges semantic knowledge with spatial awareness, allowing systems to provide more informative and actionable insights. For example, detecting an obstacle and accurately measuring its distance enables timely responses and safer navigation decisions. Such a combined approach is especially important in real-time scenarios, where rapid and precise interpretation of the environment is necessary.

Advancements in deep learning have significantly enhanced the performance and efficiency of models used for both object detection and depth estimation. Modern lightweight architectures make it possible to deploy these systems on embedded and portable devices, ensuring practicality in real-world applications. Additionally, incorporating feedback mechanisms such as audio or haptic signals improves usability, particularly for visually impaired users, by delivering navigation cues in an intuitive manner.

This work presents a Vision-Based Navigation Assistant that integrates these technologies to deliver dependable, real-time guidance. The system is designed to enhance safety, independence, and mobility by providing accurate perception of the environment along with meaningful feedback. Overall, it highlights the transformative role of computer vision in developing advanced navigation solutions for diverse applications.

II. RELATED WORKS

I. Visual Navigation Using Depth Estimation Based on Hybrid Deep Learning in Sparsely Connected Path Networks for Robustness and Low Complexity" Huda Al-Saedi, Pedram Salehpour, Seyyed Hadi Aghdasi presents a hybrid deep learning approach for robotic navigation that combines depth estimation with decision-making processes. An attention-based UNET architecture is used to generate depth maps from event-based camera inputs. These maps are further processed using a CNN-LSTM network to predict navigation commands such as turning or moving forward. The model is designed to be computationally efficient while maintaining high accuracy. Experimental evaluations indicate reliable performance with minimal errors, making it suitable for real-time navigation scenarios with limited computational resources.

Object Detection and Depth Estimation Using Deep Learning"Authors: Rajani Katiyar, Uttara Kumari, Karthik Panagar, Kashinath Patil, et al.This paper investigates the integration of object detection and depth estimation through deep learning methods for navigation applications. The system detects objects using bounding box techniques and estimates their distances to enhance environmental understanding. It addresses issues such as occlusion and inaccurate depth predictions. By employing convolutional neural networks and optimized training methods, the approach improves both detection accuracy and depth estimation. The results demonstrate its usefulness in applications like robotics and intelligent transportation systems.

"Stereo Vision Based Object Detection for Autonomous Navigation in Space Environments" introduces a stereo vision-based method for detecting objects and estimating depth in space environments. It integrates the Single Shot Detector (SSD) with triangulation techniques to identify objects and compute their distances. The system is specifically designed for space missions, where detecting distant objects such as debris is critical. The study discusses challenges like environmental disturbances and hardware constraints. Experimental findings confirm the effectiveness of stereo vision in supporting autonomous navigation and improving safety in space operations.

"Dense Monocular Depth Estimation for Stereoscopic Vision Based on Pyramid Transformer and Multi-Scale Feature Fusion" Zhongyi Xia, Tianzhao Wu, Zhuoyan Wang, Man Zhou, et al. proposes a transformer-based model for monocular depth estimation using multi-scale feature fusion. The approach enhances depth prediction accuracy by integrating pyramid transformer structures with hierarchical feature extraction. It significantly improves 3D scene understanding, which is essential for navigation systems. Evaluations on benchmark datasets show that the model outperforms traditional convolutional approaches. The study underlines the importance of accurate depth estimation in applications such as robotics, augmented reality, and autonomous driving.

Scalable Vision-Based 3D Object Detection and Monocular Depth Estimation for Autonomous Driving focuses on enhancing 3D object detection by combining monocular and stereo depth estimation techniques. The authors incorporate geometric constraints into detection models to improve robustness and accuracy. The framework supports multiple data annotation formats and uses efficient training strategies to boost scalability. Experimental results show strong performance on standard autonomous driving datasets. The paper demonstrates how integrating detection and depth estimation improves environmental perception and decision-making.

Object Detection and Depth Estimation Approach Based on Deep Convolutional Neural Networks explores CNN-based approaches for object detection and depth estimation. It compares one-stage detectors like YOLO and SSD with two-stage methods such as Faster R-CNN, discussing their trade-offs in speed and accuracy. The paper also highlights the use of advanced architectures like ResNet and feature pyramid networks to enhance performance. The integration of these techniques supports real-time processing, which is critical for navigation applications. It provides useful insights into designing efficient perception systems.

Monocular Based Navigation System for Autonomous Ground Robots Using Multiple Deep Learning Models presents a monocular vision-based navigation framework that integrates object detection, depth estimation, and semantic segmentation. The proposed system uses shared feature representations across multiple tasks to improve efficiency. Depth estimation is refined using semantic information, leading to better scene interpretation. The combined outputs help detect objects and estimate distances effectively. Experimental results indicate

improved navigation performance in complex environments, emphasizing the advantages of multi-task learning.

Vision-Based Navigation and Perception for Autonomous Robots: Sensors, SLAM, Control Strategies, and Cross-Domain Applications paper provides a comprehensive overview of vision-based navigation methods, including depth estimation, SLAM, and perception techniques. It discusses different sensor types such as monocular and stereo cameras, along with event-based sensors. The paper explains the role of depth estimation in mapping, localization, and obstacle avoidance. It also highlights key challenges like scale ambiguity and environmental variability. The study serves as a valuable reference for understanding modern navigation systems and their components.

Deep Optics for Monocular Depth Estimation and 3D Object Detection Julie Chang, Gordon Wetzstein introduces a novel concept called deep optics, which combines optical design with deep learning to improve depth estimation. The method uses coded defocus blur as an additional cue for estimating depth from single images. The optimized optical system enhances both depth prediction and 3D object detection. Experiments conducted on standard datasets demonstrate improved accuracy and generalization. The work highlights the benefits of integrating hardware design with learning-based methods.

Categorical Depth Distribution Network for Monocular 3D Object Detection Cody Reading, Ali Harakeh, Julia Chae, Steven L. Waslander proposes a method for monocular 3D object detection based on predicting depth distributions rather than single values. By estimating a range of possible depths, the model improves accuracy and robustness. The approach integrates depth estimation and object detection into a unified framework. It utilizes bird's-eye-view representation to generate precise 3D bounding boxes. The model achieves state-of-the-art results on datasets like KITTI and Waymo, demonstrating the effectiveness of probabilistic depth modeling.

III. PROPOSED METHOD

The proposed Vision-Based Navigation Assistant is designed to provide real-time guidance by integrating object detection and depth estimation into a unified framework. The system utilizes a camera as the primary input device to continuously capture visual data from the surrounding environment. This input is processed through a deep learning-based object detection model, such as YOLO or SSD, to identify and classify objects including obstacles, pedestrians, and pathways.
Simultaneously, a depth estimation module is employed to calculate the distance between the user and the detected objects. This can be achieved using either monocular depth estimation techniques or stereo vision, depending on system requirements and hardware availability. The depth information is then fused with object detection outputs to create a comprehensive understanding of the scene, enabling accurate spatial awareness.

A decision-making module processes the combined data to determine potential risks and safe navigation paths. For example, if an obstacle is detected within a critical distance, the system generates alerts and suggests alternative directions. The system is optimized for real-time performance and can be deployed on embedded platforms such as Raspberry Pi or mobile devices, ensuring portability and low power consumption.

To enhance usability, the system incorporates feedback mechanisms such as audio instructions or haptic signals, allowing users—especially visually impaired individuals—to receive intuitive guidance. The modular architecture of the system also allows for scalability and integration with additional features like GPS or SLAM for improved navigation. Overall, the proposed system aims to deliver an efficient, accurate, and user-friendly navigation solution by leveraging advancements in computer vision and deep learning, thereby improving safety, independence, and mobility in both indoor and outdoor environments.
.

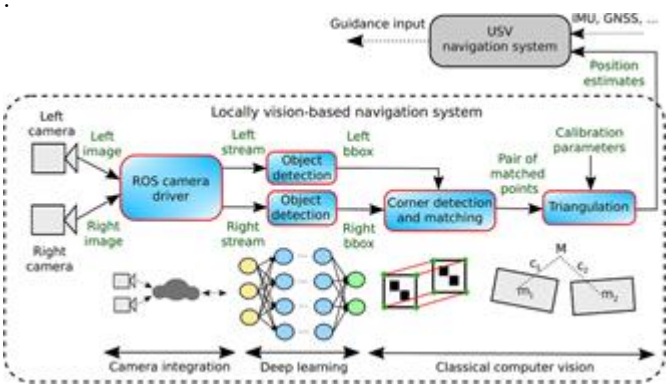

Fig.1.System Architecture

The proposed Vision-Based Navigation Assistant is structured into several functional components, each responsible for a specific task to ensure efficient and accurate navigation. These components operate in a sequential and integrated manner to process visual data and deliver real-time guidance.

The first component is the Image Acquisition Component, which captures real-time video or image frames using a camera.

This component acts as the input layer of the system, continuously supplying visual data for further processing. The quality and resolution of the captured images significantly influence the accuracy of the subsequent components.

The second component is the Object Detection Component, which utilizes deep learning algorithms such as YOLO or SSD to identify and classify objects within the environment. It detects elements such as obstacles, pedestrians, vehicles, and pathways. The output includes bounding boxes, object labels, and confidence scores, enabling effective scene understanding. The third component is the Depth Estimation Component, which determines the distance between the camera and detected objects. This can be implemented using monocular or stereo vision techniques. It produces depth maps that provide essential spatial information about object placement, which is critical for navigation.

The next is the Data Fusion Component, which combines outputs from the object detection and depth estimation components. By merging semantic information (object identity) with spatial data (distance), this component generates a more comprehensive understanding of the environment and improves hazard detection.

Following this is the Decision-Making Component, where the system analyzes the fused data to determine suitable navigation actions. It assesses obstacle proximity and identifies safe paths. Based on predefined conditions, it generates instructions such as moving forward, turning, or stopping.

Finally, the Feedback Component communicates navigation decisions to the user through audio cues, voice instructions, or haptic signals, ensuring accessibility, particularly for visually impaired users.

Finally, the Database Management Module stores all system data, including user details, donation history, and request records. It ensures efficient data retrieval and secure storage.

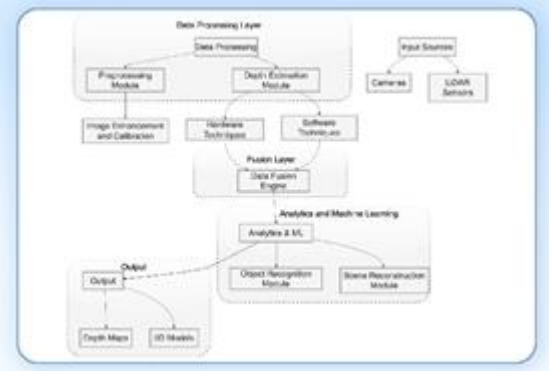

Fig.2.Methodology workflow of Vision Based Navigation Assistant

Overall Working Flow of the Proposed System:

The proposed Vision-Based Navigation Assistant operates through a structured pipeline that converts raw visual data into actionable navigation guidance. The workflow starts with the image capture stage, where a camera continuously records real-time video frames from the surrounding environment. These frames are then preprocessed to improve clarity, remove noise, and standardize the input for accurate analysis.

After preprocessing, the frames are forwarded to the object detection stage. Here, deep learning models such as YOLO or SSD examine each frame to detect and classify objects within the scene. The system produces outputs like bounding boxes, object categories, and confidence levels, which help in understanding the environment. This stage allows the system to identify important elements such as obstacles, pedestrians, and pathways.

At the same time, the input frames are processed by the depth estimation stage. This stage calculates the distance between the camera and surrounding objects using monocular or stereo vision methods. The output is a depth map that provides spatial information about object placement, helping the system understand how far objects are from the user.

Next, the data fusion stage integrates the results from object detection and depth estimation. By combining semantic details (object type) with spatial information (distance), the system forms a more complete and accurate representation of the environment. This integration improves the detection of potential hazards.

The fused data is then analyzed in the decision-making stage, where the system determines suitable navigation actions. It evaluates object proximity and identifies safe routes based on predefined rules. If an obstacle is too close, the system

generates alerts and suggests alternative movements to ensure safety.

Finally, the feedback stage delivers these instructions to the user through audio messages, voice prompts, or haptic signals. This enables clear and timely communication, especially for visually impaired users. The entire process runs continuously in real time, allowing the system to adapt to changing environments and provide reliable navigation assistance.

The performance of the proposed Vision-Based Navigation Assistant is evaluated based on accuracy, speed, and reliability in real-time scenarios. Object detection performance is measured using metrics such as precision, recall, and Intersection over Union (IoU), ensuring accurate identification and localization of objects. Depth estimation is assessed by comparing predicted distances with ground truth values, demonstrating the system's ability to provide reliable spatial information.

The system is tested in both indoor and outdoor environments to evaluate its adaptability under varying lighting and environmental conditions. Results indicate that the model achieves high detection accuracy with minimal latency, making it suitable for real-time applications. Additionally, the integration of object detection and depth estimation improves overall navigation performance by reducing collision risks.

The feedback mechanism is also evaluated for responsiveness and clarity, ensuring effective communication with users. Overall, the system shows consistent and dependable performance across different scenarios.

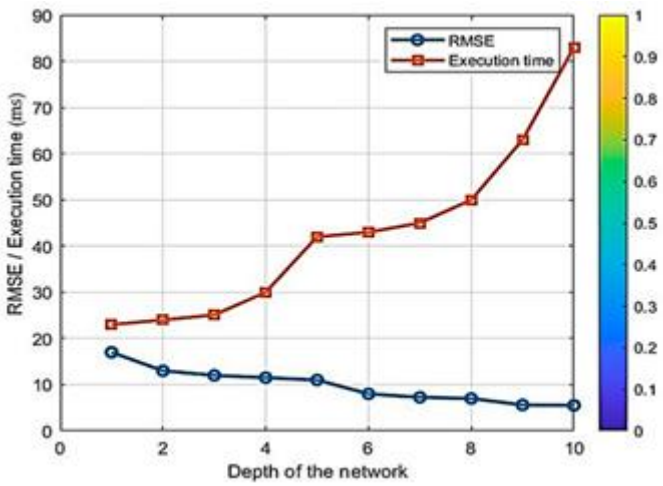

Fig.3.Performance Evaluation of Vision Based Navigation Assistant

$$Z = \frac{f \cdot B}{d}$$

This equation is used in stereo vision to calculate the depth (Z) of an object from the camera. Here, f represents the focal length of the camera, B is the baseline distance between two कैमras, and d is the disparity, which is the difference in the position of an object between left and right images. A larger disparity indicates that the object is closer, while a smaller disparity means it is farther away. This formula is fundamental in determining accurate distance measurements, enabling the system to understand spatial relationships and avoid obstacles effectively.

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The Euclidean distance formula is used to calculate the straight-line distance between two points in a 2D space. In the navigation system, it helps determine the relative position of objects within an image frame or between detected points. This is particularly useful when estimating distances in image coordinates before mapping them to real-world values. By calculating how far objects are from a reference point, the system can assess proximity and make decisions about obstacle avoidance and safe path selection.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

Intersection over Union (IoU) is a key metric used in object detection to evaluate the accuracy of predicted bounding boxes. It measures how well the predicted box overlaps with the ground truth box. A higher IoU value indicates better detection accuracy. In the proposed system, IoU helps in validating object detection performance during training and testing. It ensures that detected objects are correctly localized, which is essential for reliable navigation, as incorrect bounding boxes may lead to inaccurate distance estimation and unsafe decisions.

V. CONCLUSION

The proposed Vision-Based Navigation Assistant presents an efficient solution for navigating complex and dynamic environments by combining object detection with depth estimation. This integration of semantic understanding and spatial information enables the system to accurately recognize obstacles and estimate their distances in real time. As a result,

it improves decision-making and supports safer, more reliable navigation.

The application of deep learning techniques enhances the accuracy of object identification and depth prediction, while the system's lightweight design allows it to be implemented on portable and embedded platforms. Furthermore, the inclusion of audio and haptic feedback makes the system user-friendly and especially valuable for visually impaired individuals by providing clear and intuitive guidance.
Performance evaluations show that the system functions effectively in both indoor and outdoor settings, adapting to different environmental conditions. In conclusion, this work demonstrates the significant role of computer vision in creating advanced navigation solutions that enhance user safety, independence, and mobility. Future improvements could involve incorporating advanced localization and mapping methods to further increase system efficiency and scalability.

VI. FUTURE WORK

Future work can also focus on improving depth estimation accuracy under challenging conditions, such as low lighting, dynamic environments, and occlusions. Incorporating multimodal sensors, including LiDAR or ultrasonic sensors, alongside vision-based methods could enhance robustness and reliability. Additionally, optimizing deep learning models for faster inference and lower power consumption would make the system more efficient for real-time deployment on edge devices.

Another potential improvement is the inclusion of adaptive learning mechanisms, allowing the system to learn from user behavior and environmental changes over time. Enhancing the feedback system with natural language processing could provide more interactive and user-friendly guidance.

Finally, extensive real-world testing across diverse environments and user groups will be essential to refine the system's performance. These advancements will contribute to making the navigation assistant more intelligent, scalable, and accessible for a wide range of applications.

REFERENCES

1. M. Y. Arafat, M. M. Alam, and S. Moh, "Vision-based navigation techniques for unmanned aerial vehicles: Review and challenges," Drones, vol. 7, no. 2, p. 89, 2023.
2. E. Cetin, T. T. Sarl, M. Assoy, and G. Secinti, "Monodepth-based object detection and depth sensing for vehicle vision systems," in Proc. IEEE WF-IoT, 2024.
3. V. Saini, M. V. P. Kantipudi, and P. Meduri, "Enhanced SSD algorithm-based object detection and depth estimation for autonomous vehicle navigation," Int. J. Transport Development Integration, vol. 7, no. 4, pp. 341–351, 2023.
4. A. Raivi and S. Moh, "Vision-based navigation for urban air mobility: A survey," in Proc. SMA, 2023.
5. M. Mahdavian, K. Yin, and M. Chen, "Robust visual navigation using 3D semantic maps," IEEE Robot. Autom. Lett., vol. 7, pp. 8590–8597, 2022.
6. N. Ge and Y. Yong, "A survey of vision-based object detection," in Proc. ICICML, 2022.
7. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. CVPR, 2016.
8. J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in Proc. CVPR, 2017.
9. W. Liu et al., "SSD: Single shot multibox detector," in Proc. ECCV, 2016.
10. R. Girshick, "Fast R-CNN," in Proc. ICCV, 2015.
11. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN," IEEE Trans. Pattern Anal. Mach. Intell., 2017.
12. T.-Y. Lin et al., "Feature pyramid networks for object detection," in Proc. CVPR, 2017.
13. H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," 2020.
14. X. Wang, W. Yin, T. Kong, Y. Jiang, L. Li, and C. Shen, "Task-aware monocular depth estimation for 3D object detection," 2019.
15. Y. Li et al., "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," 2022.
16. L. Yang et al., "BEVHeight: A robust framework for vision-based 3D object detection," 2023.
17. C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in Proc. CVPR, 2017.
18. I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in Proc. 3DV, 2016.
19. A. Atapour-Abarghouei and T. P. Breckon, "Monocular segment-wise depth estimation," in Proc. ICIP, 2019.
20. Z. Teed and J. Deng, "DeepV2D: Video to depth with differentiable structure from motion," 2018.
21. A. Pilzer et al., "Unsupervised adversarial depth estimation," in Proc. 3DV, 2018.
22. A. Tonioni et al., "Real-time self-adaptive deep stereo," in Proc. CVPR, 2019.

23. G. Yang et al., “SegStereo: Exploiting semantic information for disparity estimation,” in Proc. ECCV, 2018.
24. Y. Liu, H. Wang, C. Dong, and Q. Chen, “Car-following using binocular stereo vision,” IEEE Access, vol. 8, pp. 25350–25363, 2020.