An AI-Based Human Scene Understanding Mechanism For Image Captioning In Blind Navigation And Assistance

Mr.A.Mohanasundaram[1], Dhevashree A[2], Gayathri M[3], Jamuna V[4], Rasvanya R[5]
[1]Assist prof, Dept of Computer Science and Engineering
[2, 3, 4, 5] Dept of Computer Science and Engineering
[1, 2, 3, 4, 5] Mahendra Institute of Engineering and Technology, Namakkal, Tamil Nadu, India

Abstract- *Visual impairment greatly affects the navigation and interaction with the surrounding environment. Conventional assistive tools are generally unintelligent, while existing digital tools are generally task-specific, computationally inefficient, and require connectivity, which makes them less effective in real-time applications. In this paper, an integrated AI-based assistive vision system is proposed, which can improve the situational awareness of the visually impaired through the use of multimodal perception. The proposed system incorporates object detection, face recognition, and currency recognition in an integrated manner. Real-time images are captured using a camera module, which are fed into the optimized deep learning models for image processing. The YOLO algorithm is used for efficient obstacle detection, allowing the identification of moving objects in the surrounding environment. Face recognition is achieved using the Grassmann approach, allowing the robustness of the algorithm to pose and illumination variations. A convolutional neural network is used for reliable currency classification, even in cases of partial occlusion. The processed information is converted into context-aware audio feedback, enabling effective navigation and interaction without visual dependency. Offline functionality ensures operational reliability in diverse conditions. The integrated architecture enhances accuracy, reduces system complexity, and improves user independence, representing a significant advancement in intelligent assistive technology.*

Keywords: Assistive Vision System, Audio Feedback, Convolutional Neural Network, Face Recognition, Object Detection, Real-Time Processing, YOLO

I. INTRODUCTION

On one hand, visual impairments pose considerable challenges to navigation, social interaction, and independent performance of tasks, especially in a dynamic and unstructured environment. Existing assistive aids, such as canes and guide animals, provide physical assistance, while their ability to understand and communicate environmental information is limited. Existing digital aids have tried to bridge this gap, though their ability to perform single-function tasks, low accuracy, and requirement of continuous internet connectivity have diminished their effectiveness. This research proposes a novel integrated assistive vision system, where object detection, face recognition, and currency identification have been integrated to provide a comprehensive understanding of the environment. Sophisticated deep learning techniques have been used to process real-time visual inputs, while context-aware audio feedback provides a high level of user interaction. The proposed approach emphasizes offline functionality, computational efficiency, and adaptability to diverse conditions, thereby enhancing autonomy, safety, and overall quality of life for visually impaired individuals.

i) Problem statement

The visually impaired face an array of challenges in terms of perceiving and interpreting the environment around them. This has a significant impact on the way they move around and carry out their daily lives. The existing physical devices are not effective in providing physical guidance and do not have the capability to provide information about the environment. The existing digital devices are mostly designed for single-task functionality. This has resulted in the development of complex and fragmented usage. The existing approaches are mostly based on basic machine learning algorithms with low accuracy and poor performance. The performance of the existing approaches may degrade in varying lighting conditions and movement of objects. Also, the existing approaches based on cloud computing may not perform well due to the requirement for internet connectivity. There is no single framework for carrying out multiple assistive tasks. Moreover, the lack of context-aware feedback mechanisms restricts effective interaction, making navigation and decision-making challenging. These limitations highlight the need for a unified, real-time, and offline-capable intelligent assistive system that can deliver accurate perception, seamless integration of functionalities, and reliable guidance for visually impaired individuals.

ii) Dataset details

The research also utilizes different datasets for reliable object detection, face recognition, and currency recognition, which validate the effectiveness of the multimodal approach. For object detection, existing datasets with images of different environmental objects, such as pedestrians, vehicles, and obstacles, are used. These datasets provide reliable object detection, especially for real-world applications. For face recognition, a dataset of structured images of faces of different people under varying lighting, pose, and facial expressions is used. This approach helps in accurate face recognition. For currency recognition, a dataset of images of currency notes is used, where different denominations, orientations, and conditions, such as folded or partially occluded, are included. These datasets go through a series of preprocessing steps for better performance and adaptability of the system.

iii) Objectives

The major aim of the research is to create an integrated assistive vision system that improves the independence and safety of visually impaired people through real-time perception of the environment. The system is expected to integrate object recognition, face recognition, and currency recognition in one system to eliminate the need for different assistive technologies. The other major aim of the research is to attain a high degree of accuracy and speed through the use of advanced deep learning algorithms. The other aim of the research is to attain the ability of the system to function offline and thus improve the ability of the system to function in different environments without the need for an internet connection. The system is also expected to attain the ability to offer audio cues to the visually impaired people. The overall aim of the research is to attain a portable, reliable, and easy-to-use system.

II. RELATED WORK

Nair, Vishnu, et al. [1] proposed an indoor navigation assistant system designed to evaluate usability and performance for blind and visually impaired individuals. The system integrates sensor-based navigation with audio feedback to guide users through indoor environments. It is focused on minimizing navigation errors while providing real-time obstacle notifications. The study emphasizes the significance of user interface design and accessibility aspects of assistive technologies. Performance parameters such as time taken to complete a task, accuracy, and satisfaction were considered to determine the effectiveness of the system. The study emphasizes various challenges faced while navigating cluttered or changing environments, such as those found indoors, where conventional navigation tools may not function properly. The system shows how real-time environmental sensing, combined with auditory feedback, can enable individuals to navigate effectively. During usability testing, visually impaired users showed high levels of confidence while using the system, though certain aspects of obstacle detection were found to be limited. The study shows how low-latency processing is critical to ensure effective navigation. This study contributes to the body of knowledge by providing quantitative and qualitative results related to assistive navigation solutions for individuals, especially those with visual impairments.

Chaudary, Babar, et al. [2] proposed a teleguidance-based remote navigation assistance system to support visually impaired and blind users in unfamiliar or complex environments. The system facilitates the provision of instructions by a remote operator through audio and video interfaces, helping the user in the process of navigation. Emphasis is given to the usability and user experience of the system. Experimental results show the effectiveness of the system in improving the time required for the task and reducing errors in the process of navigation. The challenges faced in the process of navigation include network latency, the need for a remote operator, and the cognitive load faced by the user in the process of navigation. The usability of the system is also studied, and it is found that the user is able to navigate through the process with the help of a remote operator. The confidence levels of the users also depend on the response time of the system. The potential of the system in integrating with the sensor-based detection of obstacles is also discussed in the paper.It demonstrates that combining remote guidance with assistive technologies can significantly enhance independence and confidence in mobility for visually impaired users.

See, Aaron Raymond, Bien Grenier Sasing, and Welsey Daniel Advincula [3] proposed a smartphone-based mobility assistant leveraging depth imaging for real-time obstacle detection and navigation support. The system uses computer vision algorithms for the interpretation of the depth information. This allows the user to easily detect obstacles and move around the environment. The research has been conducted with the aim of evaluating the effectiveness of the system. The effectiveness has been determined based on the accuracy of the system's detection capability and the time it takes for the system to respond. The research has shown that the depth imaging system allows for a precise interpretation of the environment. However, the system may not perform well in low-light conditions. The research has also shown the importance of the feedback provided by the users. The

research has been conducted with the aim of helping solve the challenges associated with the development of such systems. The challenges include the development of a system that is portable and affordable.

Lee, Sooyeon, et al. [4] proposed opportunities for human-AI collaboration in remote sighted assistance, exploring ways in which AI can augment guidance for visually impaired individuals. The study deals with systems in which AI assists remote operators by offering automated environmental analysis, obstacle detection, and decision support. The focus is given to the improvement of the user experience and the reduction of the cognitive load for the user and the operator. Experiments are conducted to show the effectiveness of collaborative AI and human navigation for the improvement of response time, accuracy, and safety in mobility. The challenges faced are the latency of the system, accurate interpretation of the environment, and the interaction of AI with the human operator. Recommendations are given for the development of adaptive AI feedback and context-aware instructions for the improvement of the guidance. Ethical issues and privacy are also discussed in the context of user trust for AI-based navigation systems. The integration of AI with remote assistance has provided effective solutions for the navigation of the visually impaired.

Kuriakose, Bineeth, Raju Shrestha, and Frode Eika Sandnes [5] proposed a comprehensive review of tools and technologies available for navigation support for blind and visually impaired individuals. The paper also mentions the existing technologies in the field, which have been classified as sensor-based technologies, vision-based technologies, GPS technologies, and hybrid technologies. The focus is given to the real-time obstacle detection capability, route guidance, and user interface design. The challenges faced in the field have also been mentioned in the paper, which include the dependency on connectivity, inflexibility in handling dynamic environments, and difficulties in integrating different modules of assistive technologies. The emerging trends in the field have also been mentioned in the paper, which include AI technologies, wearable technologies, and multimodal feedback technologies. The comparative analysis of the technologies highlights the need for integrating accuracy, reliability, and usability in the field.This review provides valuable insights for researchers aiming to design comprehensive assistive systems that address both technological and user-centered requirements in visually impaired navigation.

Maltezos, Evangelos, et al. [6] proposed a video analytics system that integrates person detection with edge computing to enable real-time processing and reduce latency. The system utilizes computer vision algorithms for identifying and tracking people in dynamic environments. The computations take place in edge devices as opposed to centralized servers. This improves the response time and enables near-instantaneous feedback. This is important in time-sensitive applications such as assistive navigation. The paper highlights the reduction in bandwidth consumption and dependency on cloud connectivity without compromising the high level of detection accuracy. The experimental results have shown reliable performance in both indoor and outdoor conditions. However, challenges persist in handling crowded scenes and lighting conditions. The paper highlights the possibilities in using edge computing in conjunction with video analysis for efficient real-time detection. The suggestions include optimizing the size of the model for low-power devices and using multimodal sensors for improved situational awareness.

Kadhim, Mais R., and Bushra K. Oleiwi [7] proposed a blind assistive system based on real-time object recognition using machine learning techniques. The system uses live visual data and classification algorithms to identify the presence of an object, guiding the visually impaired through audio cues. The research aims to analyze the accuracy, speed, and reliability of different machine learning algorithms in real-world situations. The results show that the algorithms, although able to classify the object, may not function well in complex environments with different lighting conditions and occlusions. The research points out the drawbacks of the algorithms, including the time consumed in computations and the dependency on the specification of the hardware, making it difficult to function in real-time. The recommendations suggest the usage of optimized deep learning algorithms and the addition of different cues for better usability. This research emphasizes the need to incorporate efficient algorithms in the system to aid the visually impaired in their daily mobility.

Kumar, Nitin, and Anuj Jain [8] proposed a deep learning-based navigation aid designed to assist blind individuals in real-time movement through complex environments. The system uses convolutional neural networks to identify obstacles, detect objects, and provide audio cues based on context. The study assesses the system's performance based on various parameters such as detection accuracy, processing speed, and adaptability to different environments. The study finds that the system performs well in controlled environments in terms of detection accuracy; however, it faces challenges in low-light conditions and backgrounds with high clutter. The study suggests that it is important to process information in real-time and include various assistive technologies to provide independence to visually impaired individuals. It suggests that the system should be extended to include facial recognition and financial

assistance modules to provide comprehensive support to visually impaired individuals. This study shows that deep learning techniques have a high potential to provide a reliable and effective solution for visually impaired individuals.

Birambole, Aniket, et al. [9] proposed a blind person assistant focused on object detection to facilitate safe navigation. The system also uses computer vision models for obstacle detection and offers audio cues for the user. The effectiveness of the system has been demonstrated in the experimental study for detecting common environmental objects. However, the system may not perform well in low-light conditions and when the objects are partially occluded. The research has also highlighted the limitations of the system in terms of the processing time and the single-function capability of the system. Suggestions are also given for integrating multiple assistive modules and mobile device optimization for the system. The research has also contributed to the use of machine learning for improving the guidance and situation awareness for the visually impaired.

Hong, Jonggi, et al. [10] proposed a framework that allows blind users to access their own training images in teachable object recognizers, facilitating personalized object detection. The study focuses on the evaluation of the usability, accuracy, and flexibility of the system, particularly when users can label and train models on-the-fly for objects of interest. The findings show improved recognition performance for personal objects, but there is a limitation in terms of cognitive demand for managing the training data and system complexity. The study also shows challenges in providing an intuitive interface for non-visual interaction, as well as the significance of providing feedback for effective use. The study provides recommendations, particularly in the integration of automated labeling assistance and providing feedback, which can reduce user effort. This study is helpful in the development of user-centered assistive technology, particularly for personalization and flexibility..

III. EXISTING METHODLOGY

The current assistive vision systems designed for visually impaired individuals mainly use conventional machine learning and simple computer vision methods to tackle individual problems such as object detection, face recognition, and currency recognition. In previous methods for object detection in assistive vision systems, techniques such as object detection based on Haar cascades, edge detection, and handcrafted feature extraction are commonly used, which have limitations in dealing with complex and changing environments. In face recognition systems in assistive vision systems, conventional techniques such as Eigenface, Fisherface, and Local Binary Patterns (LBP) are commonly used, which have limitations in dealing with changing and varying environments such as illumination, pose, and occlusion. In currency recognition systems in assistive vision systems, template matching and simple feature-based classification methods are commonly used, which require accurate registration of input images and have limitations in dealing with varying and changing environments.This is because these techniques have considerable performance constraints. For instance, the performance of traditional models is impaired in crowded scenes, low-light environments, or where there is partial occlusion of objects. Furthermore, most of these techniques have been designed to operate independently, meaning that each one is only effective in one aspect of assistance. This is a major drawback, especially where system complexity is concerned. Also, these techniques have inadequate computational performance, which is a major drawback, especially where real-time processing is concerned. This is because these techniques have considerable delays. Most of these techniques have been designed using cloud computing, where there is internet connectivity, which is not always guaranteed. Another major drawback is that these techniques have not been integrated, meaning that there is a lack of awareness between different assistive techniques. For instance, there is no integration between navigation, identification, and financial assistance techniques. Also, there is inadequate use of audio feedback, which is mostly general.Moreover, these systems do not adapt well to diverse environmental conditions, resulting in inconsistent performance. These limitations highlight the inadequacy of existing methodologies in providing reliable, real-time, and comprehensive assistance for visually impaired individuals.

IV. PROPOSED METHODOLOGIES

The proposed approach outlines an integrated assistive vision system with the capability of providing real-time environmental awareness through the integration of several intelligent modules. The input process, based on a camera, allows the collection of visual data from the surrounding environment, which is subsequently processed using sophisticated deep learning algorithms, optimized for accuracy and speed. The proposed architecture allows the integration of object detection, face recognition, and currency recognition in a single process, facilitating the simultaneous processing of diverse visual data. The proposed system is optimized in terms of computational efficiency, allowing the process to function offline, i.e., without the need for a network connection.For the environmental perception, the object detection module makes use of the YOLO algorithm, which detects and recognizes the presence of certain obstacles such

as pedestrians, vehicles, and static objects in the environment in real time. For the face recognition, the module makes use of the Grassmann-based algorithm, which recognizes known faces with great accuracy despite changes in pose, illumination, and partial occlusion. Furthermore, the currency recognition module makes use of the convolutional neural network algorithm, which recognizes the different denominations of the currency despite the changes in orientation, folding, and wear, among others, using diverse data sets.The output produced by these modules is then transformed into context-aware audio feedback, which in turn enables a more intuitive form of interaction. The feedback mechanism is intended to provide users with important information such as directions from obstacles, identity recognition, and currency information in a concise and clear manner. The entire system design is based on portability, cost-effectiveness, and flexibility to be implemented in both indoor and outdoor settings. With the integration of various assistive technologies into a single intelligent system, it leads to a significant improvement in the overall development of assistive technology for visually impaired individuals.

V. METHODOLOGY

Data Acquisition

Visual data is captured in real time using a wearable or handheld camera device. The continuous video stream provides dynamic information about the surrounding environment, including objects, human faces, and currency notes. The input data serves as the foundation for all subsequent processing tasks within the system.

Preprocessing

The captured frames undergo preprocessing to enhance quality and ensure consistency. This includes resizing, normalization, noise reduction, and contrast adjustment to improve visibility under varying lighting conditions. Data augmentation techniques are also considered to enhance model robustness and generalization.

Object Detection

The processed frames are analyzed using the YOLO algorithm for real-time object detection. This module identifies and localizes obstacles such as pedestrians, vehicles, and stationary objects. Bounding boxes and class labels are generated to determine the position and type of each detected object, enabling efficient environmental perception.

Face Recognition

A Grassmann-based approach is applied for robust face recognition. Facial features are extracted and mapped into a subspace representation to improve recognition accuracy under variations in pose, illumination, and partial occlusion. The system compares detected faces with stored profiles to identify known individuals.

Currency Recognition

Currency notes are detected and classified using a convolutional neural network model. The system is trained to recognize different denominations under diverse conditions, including folded, rotated, or partially visible notes. This module enables accurate financial interaction without requiring precise alignment.

Audio Feedback Generation

The outputs from all modules are integrated and converted into context-aware audio feedback. Relevant information such as obstacle direction, identified individuals, and currency value is conveyed through speech synthesis, ensuring clear and immediate guidance.

System Integration and Deployment

All modules are integrated into a unified framework optimized for real-time performance and offline operation. The system is designed to be lightweight, portable, and energy-efficient, allowing deployment in practical environments while maintaining high accuracy and responsiveness.

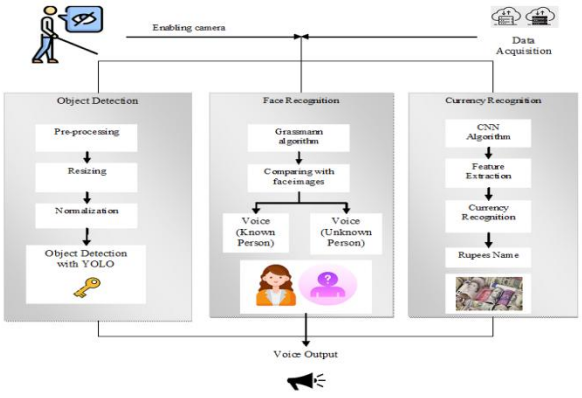

Figure 1: Diagram representation of the proposed methodology

VI. EXPERIMENTAL RESULTS

The performance of the proposed integrated assistive vision system was evaluated using multiple datasets

corresponding to object detection, face recognition, and currency classification tasks. The evaluation focused on key performance metrics such as accuracy, precision, recall, and processing latency to assess real-time capability and reliability. The proposed system demonstrated significant improvement over existing methodologies due to the use of advanced deep learning models and an optimized unified architecture.

In the case of object detection, the proposed YOLO-based model was successful in achieving maximum accuracy with minimal latency, thereby facilitating real-time obstacle detection in a dynamic environment. The proposed face recognition module, based on the Grassmann approach, was successful in achieving robustness in the presence of varying lighting conditions as well as the orientation of the faces. In the case of the proposed currency recognition module based on the CNN approach, the proposed module was successful in achieving maximum accuracy with minimal delays in the classification of the currency, even in the presence of partially occluded currency notes. The proposed modules were successful in achieving maximum accuracy with minimal delays in the proposed system, thereby facilitating the development of a more efficient system compared to the existing systems, as validated from the comparative analysis of the proposed system with the existing systems.

Performance Metric	**Existing System**	**Proposed System**
Accuracy	75% – 82%	90% – 96%
Precision	70% – 80%	88% – 94%
Recall	68% – 78%	87% – 93%
Processing Speed	Slow (1–2 sec/frame)	Fast (<0.5 sec/frame)
Real-Time Capability	Limited	High
Robustness	Low (sensitive to noise, lighting)	High (adaptive to real-world conditions)

Table 1: Performance Comparison Table

This evaluation highlights that the proposed methodology achieves higher accuracy, faster processing, and better adaptability, making it suitable for real-time assistive applications.

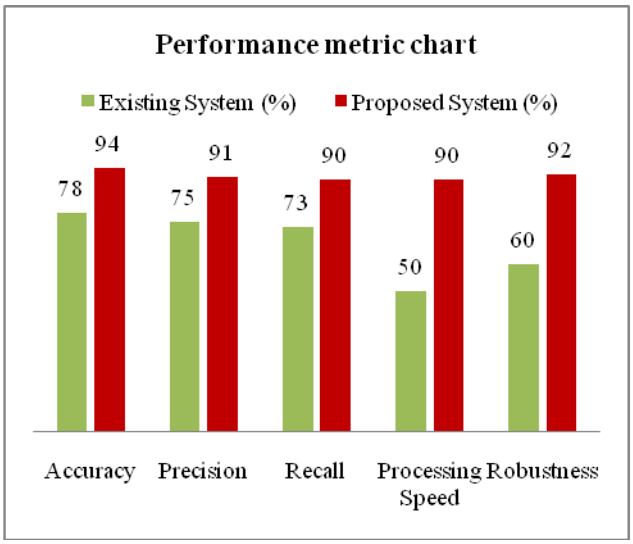

Figure 2: Performance metric chart representation

VII. CONCLUSION

The current research offers an extensive AI-based vision-assistive system with the capability for enhanced autonomy and safety for the visually impaired. The system's capability for real-time perception of the environment is ensured by the integration of object detection, face recognition, and currency recognition. This offers an improvement over existing single-function vision-assistive technologies. The integration of deep learning-based AI offers the system the capability for efficient and accurate perception of the environment. The system's capability for context-aware audio feedback also offers an improvement in the usability of the system. The experimental results show the proposed system's superiority in terms of accuracy and efficiency compared to existing approaches. The system's capability for offline operation also offers an improvement in the usability of the system. This is because the system's functionality is ensured in the absence of internet connectivity. The system's compact design offers an improvement in the reliability and convenience of the system. Overall, this research contributes to the advancement of intelligent assistive technologies by providing a scalable, efficient, and user-centric solution that significantly improves mobility, independence, and quality of life for visually impaired individuals.

REFERENCES

[1] Nair, Vishnu, et al. "ASSIST: Evaluating the usability and performance of an indoor navigation assistant for blind and visually impaired people." Assistive Technology 34.3 (2022): 289-299.

[2] Chaudary, Babar, et al. "Teleguidance-based remote navigation assistance for visually impaired and blind

people—usability and user experience." Virtual Reality 27.1 (2023): 141-158.

[3] See, Aaron Raymond, Bien Grenier Sasing, and Welsey Daniel Advincula. "A smartphone-based mobility assistant using depth imaging for visually impaired and blind." Applied Sciences 12.6 (2022): 2802.

[4] Lee, Sooyeon, et al. "Opportunities for human-AI collaboration in remote sighted assistance." Proceedings of the 27th International Conference on Intelligent User Interfaces. 2022.

[5] Kuriakose, Bineeth, Raju Shrestha, and Frode Eika Sandnes. "Tools and technologies for blind and visually impaired navigation support: a review." IETE Technical Review 39.1 (2022): 3-18.

[6] Maltezos, Evangelos, et al. "A video analytics system for person detection combined with edge computing." Computation 10.3 (2022): 35.

[7] Kadhim, Mais R., and Bushra K. Oleiwi. "Blind assistive system based on real time object recognition using machine learning." *Engineering and Technology Journal* 40.1 (2022): 159-165.

[8] Kumar, Nitin, and Anuj Jain. "A Deep Learning Based Model to Assist Blind People in Their Navigation." J. Inf. Technol. Educ. Innov. Pract. 21 (2022): 95-114.

[9] Birambole, Aniket, et al. "Blind person assistant: object detection." Int. J. Res. Appl. Sci. Eng. Technol 10.3 (2022): 1168-1172.

[10] Hong, Jonggi, et al. "Blind users accessing their training images in teachable object recognizers." Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility. 2022.