

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Real-time object recognition and audio feedback system for visually impaired individuals

Gökmen Katipoğlu^{1*}, Semih Utku²

¹ Kafkas University, Engineering Faculty/Computer Engineering, Kars, Türkiye

² Dokuz Eylül University, Engineering Faculty/Computer Engineering, İzmir, Türkiye

ARTICLE INFO

* Correspondence: gokmenkatipoglu@kafkas.edu.tr

DOI: <https://doi.org/10.46793/ET26.D08K>

PAGES: D69 – D74

ABSTRACT

This study presents a real-time object recognition system specifically designed to support visually impaired individuals by improving their independence and environmental awareness. Unlike general-purpose object detection systems, the proposed approach is designed with accessibility as a primary focus. It emphasizes intuitive interaction and clear auditory feedback to better support the needs of visually impaired users. The system combines computer vision, natural language generation, and speech synthesis within a mobile application. The system then provides immediate audio descriptions of nearby objects and their positions in a clear and understandable way. The architecture consists of a React Native-based mobile frontend and a Python-based backend deployed on cloud infrastructure. A pretrained YOLOv8 object detection model is utilized to identify objects in real time. Spatial analysis is performed to determine the relative position of objects within the scene. Experimental results demonstrate that the application achieves reliable object detection with low latency. Furthermore, the user-centered design, including bilingual support and touch-based interaction, enhances usability and accessibility. Overall, the proposed solution contributes to the field of assistive technologies by offering an efficient, cost-effective, and portable tool. It improves environmental awareness and supports independent mobility for visually impaired individuals.

KEYWORDS

YOLOv8, Computer Vision, Assistive Technology, Visual Impairment.

1. INTRODUCTION

According to global disability statistics, millions of visually impaired individuals encounter difficulties when interacting with their environment. Tasks such as identifying objects, recognizing obstacles, and understanding spatial relationships require visual perception and often limit independent mobility. Traditional assistive tools such as canes and guide dogs provide navigation support but cannot offer detailed contextual information regarding surrounding objects [1].

Recent advances in artificial intelligence and computer vision have enabled the development of intelligent assistive systems capable of interpreting visual information and presenting it through alternative sensory channels. Object detection models such as YOLO have demonstrated remarkable performance in real-time image understanding tasks [2][3].

This paper proposes a mobile-based assistive application specifically developed for visually impaired users. The system captures images through a smartphone camera, detects objects, determines their spatial positions, and generates spoken descriptions that help users understand their surroundings. Similar studies have shown that combining object detection and audio feedback can significantly improve environmental awareness and independent mobility for visually impaired individuals [4][5]. The motivation behind this project lies in the desire to improve the quality of life and autonomy for visually impaired individuals. Existing tools often fall short in providing comprehensive assistance, making it necessary to innovate a solution that can meet both navigation and object identification needs. By building a user-centered tool that leverages modern AI techniques, this project aims to fill these gaps and support users in their daily lives.

The importance of inclusive design and accessibility for all is also a driving factor, as it encourages the development of technology that addresses real-world needs. With a successful outcome, this project could inspire future applications that broaden access to vital technology for other differently-abled communities.

The main contributions of this study are:

- i. Development of an accessibility-oriented mobile application.
- ii. Integration of object detection, spatial reasoning, and audio feedback.
- iii. Context-aware scene description generation.
- iv. Real-time cloud-based deployment architecture.
- v. Evaluation under realistic environmental conditions.

2. RELATED WORKS

Recent advances in computer vision and deep learning have significantly improved the performance of object detection systems [6][7]. Convolutional neural networks and modern object detection architectures have enabled real-time recognition of objects in complex environments, making them suitable for various assistive applications. In particular, the YOLO family of object detectors has gained considerable attention due to its ability to achieve high detection accuracy while maintaining low computational latency [2][3].

Object detection technologies have been successfully applied in various domains, including autonomous navigation, traffic sign recognition, surveillance systems, and healthcare applications [6][7]. The combination of computer vision and mobile computing has further enabled the development of lightweight and portable solutions capable of operating on smartphones and embedded devices [8].

Several studies have focused on developing assistive technologies for visually impaired individuals. Tapu et al. proposed a wearable system that combines object recognition and face detection techniques to provide audio-based environmental awareness for visually impaired users [1]. Several other studies have proposed image-processing-based assistive systems to support object recognition and environmental awareness for visually impaired users [4][9]. Although these solutions demonstrated promising results, many of them rely on specialized hardware or focus on limited use cases.

Recent developments in mobile computing have made it possible to deploy advanced object detection algorithms on portable devices. Such systems improve user independence by converting visual information into audio feedback and enabling real-time understanding of surrounding environments [4][5]. However, many existing approaches provide only object labels without contextual information regarding object locations and spatial relationships.

Unlike previous studies, the proposed system integrates real-time object detection, spatial reasoning, natural language generation, and speech synthesis within a single mobile application. By providing contextual audio descriptions of detected objects and their relative positions, the system aims to improve environmental awareness and independent mobility for visually impaired users.

3. SYSTEM ARCHITECTURE

The proposed architecture consists of four primary components:

1. Mobile Frontend
2. Object Detection Module
3. Scene Description Generator
4. Audio Feedback Module

The frontend is implemented using React Native and Expo, allowing users to capture images and receive spoken feedback. The backend is developed in Python and deployed using Docker containers on Hugging Face cloud infrastructure. The workflow of the system is as follows:

- i. User captures an image.
- ii. Image is sent to the backend API.
- iii. YOLOv8 detects objects.
- iv. Spatial positions are calculated.
- v. Natural language descriptions are generated.
- vi. Text is converted into speech.
- vii. Audio feedback is returned to the user.

Figure 1 presents the class diagram of the proposed system and illustrates the relationships among its core software components. The architecture is organized around the AppController module, which coordinates the interaction between im-

age acquisition, object detection, and audio feedback processes. The CameraHandler component is responsible for capturing images from the mobile device, while the ImageProcessor module performs object detection using the YOLOv8 framework. The detected objects and their spatial information are then forwarded to the feedback generation module. The AudioFeedback class implements the FeedbackProvider interface and delivers spoken descriptions to the user. This modular structure improves maintainability, scalability, and integration between the system components.

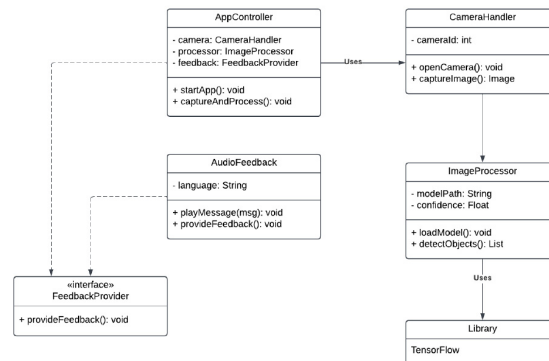


Figure 1: UML class diagram of the system

4. METHODOLOGY

4.1. Dataset and Object Detection

The proposed system was developed using a filtered subset of the Open Images Dataset V7, which contains a large collection of annotated images for object detection tasks. Only environmentally meaningful object categories were selected in order to improve usability and eliminate unnecessary information from the generated descriptions [10]. A pretrained YOLOv8 model provided by Ultralytics was employed for object detection due to its ability to achieve a balance between detection accuracy and real-time performance [11]. The use of a pretrained model eliminated the need for computationally expensive retraining while maintaining reliable detection performance. After image acquisition, the model identifies objects within the scene and produces bounding boxes, class labels, and confidence scores. Examples of object detection outputs generated by the proposed system are presented in Figures 3.

4.2. System Workflow

Detected objects are assigned to predefined spatial regions such as top-left, center, and bottom-right. This information provides users with orientation context and improves scene understanding. A custom description generation module creates human-readable scene summaries by combining detected object names with spatial information. Redundant object-part detections are removed to improve clarity. Generated descriptions are converted into speech using a Text-to-Speech engine and delivered to the user through the mobile application. The overall workflow of the proposed system is illustrated in Figure 2.

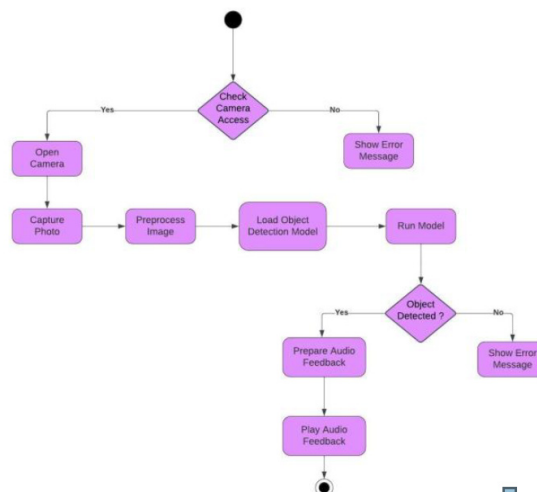


Figure 2: Workflow of the proposed object detection and audio feedback system

4.3. Detection Results

To demonstrate the effectiveness of the proposed approach, several sample images were processed using the object detection pipeline. The YOLOv8 model successfully identified multiple objects and generated meaningful descriptions for audio feedback. Figures 3 (a and b) demonstrate representative outputs obtained from the object detection module. The detected objects and their spatial locations are transformed into descriptive sentences and then synthesized into speech.

Recent studies have highlighted the effectiveness of combining computer vision and audio feedback technologies for assistive applications targeting visually impaired users [5].

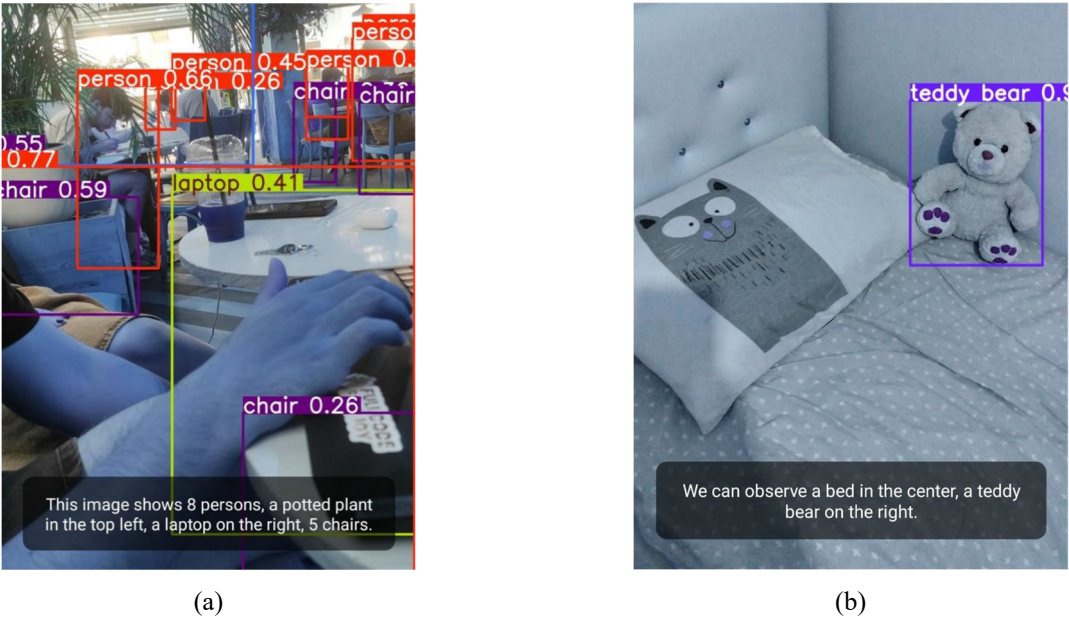


Figure 3: (a) Example of object detection results obtained using the YOLOv8 model
(b) Object detection and localization results in a complex scene

5. EXPERIMENTAL EVALUATION

To evaluate the effectiveness and practical usability of the proposed system, a series of experiments were conducted using real-world images captured under different environmental conditions. More than fifty images were processed through the complete system pipeline, including image acquisition, object detection, spatial analysis, description generation, and audio feedback. The test dataset contained indoor, outdoor, and urban scenes with varying levels of complexity, lighting conditions, and object density.

The evaluation focused on three main aspects: object detection accuracy, audio description quality, and overall user experience. Object detection accuracy was assessed by manually comparing the detected objects against ground-truth labels present in each scene. The number of correctly detected objects, missed detections, and false positives were examined to evaluate the reliability of the YOLOv8-based detection module.

In addition to detection performance, the quality of the generated audio descriptions was evaluated. The spoken outputs were reviewed with respect to clarity, grammatical correctness, timing, and contextual relevance. Particular attention was given to whether the generated descriptions accurately reflected both the detected objects and their relative positions within the scene. The description generation module successfully transformed structured detection outputs into concise and understandable sentences suitable for auditory presentation.

Figures 3 a and b illustrate representative examples of object detection and localization results produced by the proposed system. The results demonstrate that the model is capable of detecting multiple objects simultaneously while preserving spatial information required for generating contextual descriptions. Even in scenes containing several objects, the system was able to provide meaningful summaries without overwhelming the user with excessive information.

The responsiveness of the system was also evaluated. Experimental observations showed that the cloud-based architecture and optimized object detection pipeline enabled near real-time processing. The communication between the mobile application and the backend server remained stable throughout the experiments, allowing users to receive audio feedback shortly after image capture. This low-latency behavior is particularly important for assistive applications where timely information can directly affect user experience and environmental awareness. A summary of the experimental setup and observations is presented in Table 1. The results indicate that the proposed system can operate reliably under different environmental conditions while maintaining responsive audio feedback and effective object recognition performance.

Table 1: Summary of Experimental Evaluation

Evaluation Metric	Observation
Test Environments	Indoor, Outdoor, Urban
Object Detection Capability	Successful detection of commonly encountered objects
Audio Description Quality	Clear, understandable, and context-aware feedback

Evaluation Metric	Observation
System Responsiveness	Near real-time performance
Mobile–Cloud Communication	Stable during all experiments
Scalability	Modular architecture supports future extensions
Overall Performance	Suitable for assistive real-world applications

Overall, the experimental results indicate that the proposed system successfully detects commonly encountered objects and generates understandable audio descriptions under real-world conditions. The combination of object detection, spatial reasoning, and speech synthesis provides a practical assistive solution that can enhance environmental awareness and support independent mobility for visually impaired individuals. Decimal numbers

6. CONCLUSION

This paper presented a real-time object recognition and audio feedback system developed to support visually impaired individuals in understanding their surroundings. The proposed solution combines computer vision, object detection, spatial analysis, natural language generation, and speech synthesis within a mobile application framework. By transforming visual information into contextual audio descriptions, the system aims to improve environmental awareness and support independent mobility.

The developed architecture integrates a React Native-based mobile application with a cloud-based backend and a pre-trained YOLOv8 object detection model. Experimental evaluations conducted under different environmental conditions demonstrated that the system can successfully detect commonly encountered objects, determine their relative positions, and generate clear audio descriptions with low latency. The results indicate that the proposed approach is capable of providing meaningful and timely information to users while maintaining real-time performance.

In addition to its detection capabilities, the proposed system offers a practical and cost-effective assistive solution that operates using only a standard smartphone, eliminating the need for specialized hardware. The modular architecture also facilitates scalability and enables the integration of additional functionalities in future versions.

Future work will focus on expanding multilingual support, incorporating voice-command interaction, and improving scene understanding capabilities through more advanced object detection and vision-language models. Furthermore, future studies may include comprehensive user evaluations involving visually impaired participants to assess usability, accessibility, and real-world effectiveness in greater detail.

REFERENCES

1. A. Tapu, B. Mocanu and T. Zaharia, “Wearable Assistive Devices for Visually Impaired: A State of the Art Survey”, *Pattern Recognition Letters*, Vol. 137, pp. 37–52, (2020)
2. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 779–788, (2016)
3. Sohan, M., Sai Ram, T., & Rami Reddy, C. V. (2024). A Review on YOLOv8 and Its Advancements. In *International Conference on Data Intelligence and Cognitive Informatics* (pp. 529–545). Springer. https://doi.org/10.1007/978-981-99-7962-2_39
4. Manirajee, L., Shariff, S. H., Rashid, S. M., & Syar, S. (2024). Assistive Technology for Visually Impaired Individuals: A Systematic Literature Review (SLR). *International Journal of Academic Research in Business and Social Sciences*, 14(2), 596–611. <https://doi.org/10.6007/IJARBS/v14-i2/20827>
5. Zafar, S., Asif, M., Ahmad, M. B., Ghazal, T. M., Faiz, T., Ahmad, M., & Khan, M. A. (2022). Assistive Devices Analysis for Visually Impaired Persons: A Review on Taxonomy. *IEEE Access*, 10, 13354–13366. <https://doi.org/10.1109/ACCESS.2022.3146728>
6. Y. LeCun, Y. Bengio and G. Hinton, “Deep Learning”, *Nature* 521, pp. 436–444, (2015)
7. Walle, H., De Runz, C., Serres, B., & Venturini, G. (2022). A Survey on Recent Advances in AI and Vision-Based Methods for Helping and Guiding Visually Impaired People. *Applied Sciences*, 12(5), 2308. <https://doi.org/10.3390/app12052308>
8. K. M. R. Leiva, M. A. C. Gato, and J.J.S. Olmedo, “Estimation of Spatio-Temporal Parameters of Gait and Posture of Visually Impaired People Using Wearable Sensors”, *Sensors*, Vol. 23(12), pp. 1–28, <https://doi.org/10.3390/s23125564>, (2023)
9. Kumar, N., & Jain, A. (2022). A Deep Learning Based Model to Assist Blind People in Their Navigation. *Journal of Information Technology Education: Innovations in Practice*, 21, 95–114. <https://doi.org/10.28945/5006>

10. A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov and V. Ferrari, “The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale”, *International Journal of Computer Vision*, Vol. 128(4), pp. 1956–1981, (2020)
11. G. Jocher, A. Chaurasia and J. Qiu, “YOLO by Ultralytics”, <https://github.com/ultralytics/ultralytics>, (2024)
12. Samavati, F. C., Abadi, A. R. G., & Samavati, F. C. (2025). Assistive Technologies for the Visually Impaired: A Review. *Cureus Journals*, 2(1). <https://doi.org/10.7759/s44389-025-06891-1>
13. S. Bagwan and S. Sankpal, “VisualPal: A Mobile Application for Object Recognition and Voice Assistance for Visually Impaired People”, *International Journal of Engineering Research and Applications*, Vol. 5(4), pp. 30–34, (2015)
14. S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, “Path Aggregation Network for Instance Segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44(2), pp. 1025–1038, (2018)
15. M. A. Almarshad, S. Al-Ahmadi, and M. S. Islam “Adoption of Transformer Neural Network to Improve the Diagnostic Performance of Oximetry for Obstructive Sleep Apnea”, *Sensors*, Vol. 23(18), pp. 1–27, <https://doi.org/10.3390/s23187924>, (2023)