IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

VISIONSPEAK: OBJECT DETECTION AND VOICE ASSISTANCE FOR VISUALLY IMPAIRED PEOPLE.

Mrs. Keerthiga.V[1], Anne P.S[2], Bhavani. G[3]

Assistant Professor, Department of Computer Science and Engineering, Anand institute of higher technology, Chennai[1]

Student, Department of Computer Science and Engineering, Anand institute of higher technology, Chennai[2,3]

Abstract: VisionSpeak ,an Android-based mobile application that enhances real-world awareness through intelligent object detection and text recognition. Using a smartphone camera, the app identifies objects and extracts printed or handwritten text in real time. Recognized information is instantly converted into speech using a Text-to-Speech (TTS) engine, allowing users to receive voice-based feedback without needing to look at the screen. The app integrates deep learning models like MobileNet and YOLO for efficient object detection and uses the Tesseract OCR engine for text recognition. Designed with accessibility in mind, it supports voice commands, offline functionality, and a user-friendly interface. VisionSpeak is particularly useful for individuals with visual impairments, travelers, and those seeking hands-free interaction. Its seamless performance across diverse environments makes it a versatile tool for daily assistance.

Keywords: Object Detection, Text Recognition, Android Application, Text-to-Speech (TTS), Assistive Technology.

I. INTRODUCTION

.In the era of digital innovation, smartphones have evolved from mere communication tools into powerful computing platforms. With advancements in artificial intelligence (AI), machine learning, and computer vision, mobile devices can now perceive and interpret their surroundings intelligently. These advancements have made it possible to develop applications that enhance how users interact with the physical world.

This project introduces VisionSpeak, an intelligent Android application that combines real-time object detection and text recognition with voice feedback to enhance environmental awareness. Designed to function offline and in real time, VisionSpeak aims to transform the way users receive and process visual information.

The system consists of two main modules: Object Detection Module and the Text Recognition Module. The Object Detection Module utilizes pre-trained deep learning models such as YOLO (You Only Look Once) and MobileNet to identify and classify real-world objects in real time. These objects are then vocalized using Android's built-in Text-to-Speech (TTS) engine.

The Text Recognition Module employs Optical Character Recognition (OCR) technology powered by the Tesseract engine to capture and convert printed or handwritten text into audible output. This functionality is particularly useful for reading signs, labels, and documents while on the move.

This continuous monitoring ensures that the app stays useful throughout an activity rather than requiring repeated manual interaction.Security and privacy are also taken into account. All processing occurs locally on the device, meaning no images or text are sent to external servers, protecting user data from potential breaches.Asignificant motivation behind VisionSpeak is to provide a cost-effective and portable solution.

Unlike specialized hardware devices for the visually impaired, this application runs entirely on smartphones, making it accessible and affordable for a large audience.

II. RELATED WORKS

Real-time object detection and text recognition are crucial domains in computer vision, especially in applications tailored for visually impaired users.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

A real-time object detection system was developed to assist visually impaired individuals in identifying objects in their surroundings. Using deep learning for visual recognition and text-to-speech for audio output, the system enhances user independence and mobility. This solution shows how integrating visual AI with auditory interfaces can provide life-changing accessibility.[1]

An extensive review of the YOLO algorithm and its evolving versions from YOLOv1 to YOLOv8 highlighted critical improvements in speed, accuracy, and model size. Fast YOLO and Tiny YOLO were introduced to support deployment on embedded and mobile devices. These adaptations demonstrate the pr actical ity of YOLO-based models for real-time object detection in low-resource environments like smartphones.[2]

An overview of real-time object detection techniques compared traditional sliding window approaches with modern deep learning methods such as Faster R-CNN, SSD, and YOLO. The research concluded that YOLO outperforms others in speed and efficiency, making it a preferred choice for mobile apps requiring instant feedback and lightweight computation.[4]
A low-cost assistive system was implemented using Raspberry Pi and a camera module for object recognition. The device provided audio output for detected objects, proving effective for users with visual impairments. This study emphasizes the feasibility of offline, portable assistive technologies, aligning with the goals of smartphone-based applications like VisionSpeak.[3]

A combined system using object detection and text-to-speech was designed to help blind users interact with their environment. By processing camera input and conveying object information via audio, the system significantly enhanced situational awareness. This implementation aligns closely with the core functionalities and objectives of the VisionSpeak application.[5]

Wearable assistive technologies, including smart glasses and gloves, have been explored for their potential to support visually impaired users. These devices use sensors, cameras, and OCR to detect objects and provide feedback. The study emphasizes hands-free interaction and real-time responsiveness, which are also key design goals of the VisionSpeak app.[6]

A system was developed that captures visual data using a mobile camera and uses real-time object detection to identify the surroundings. Once an object is recognized, it is vocalized through a text-to-speech engine, giving users immediate audio feedback.[1]

Augmented Reality (AR) has been applied in educational settings to improve engagement and interaction. AR overlays digital information onto real-world environments, enhancing spatial awareness. Though primarily for learning, these AR principles show potential for future enhancements to assistive applications like VisionSpeak, especially in navigation and guidance.[7]

AR's integration in fields such as healthcare, tourism, and retail has shown how real-time interaction and spatial mapping can transform user experience. The study highlights the use of mobile AR for providing intuitive feedback and real-time visual guidance, which parallels the real-world awareness features envisioned for VisionSpeak.[8]
The study explored several wearable technologies like smart glasses, gloves, and electronic canes equipped with object detection sensors and OCR functionality[6]

A comparative overview of object detection algorithms such as YOLO, SSD, and Faster R-CNN was provided. This study discusses challenges like occlusion and scale variance while highlighting the superior speed and accuracy of YOLO-based methods. Such findings are crucial when developing lightweight, mobile-friendly applications for real-time use.[4]
This paper discusses spatial mapping and sensor fusion, both of which are relevant for enhancing environmental awareness in apps like VisionSpeak. It also highlights future opportunities for combining AR with AI for greater immersion.[8]

III. PROPOSED SOLUTION

The proposed system aims to integrate object detection and text recognition functionalities, followed by a voice output mechanism, in a mobile application to assist users in real-time. This system will enable users to interact with their surroundings by simply pointing their smartphone at objects or texts, with the app detecting and providing auditory feedback.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

- Real Time Object Detection:

By leveraging efficient deep learning models like YOLO (You Only Look Once) or MobileNet, the system can detect and classify various objects—such as furniture, vehicles, or animals—in real time. Once an object is recognized, its name is immediately converted into spoken words through a Text-to-Speech (TTS) engine. This functionality provides hands-free interaction and enhances situational awareness, particularly for users with visual impairments or those who need quick environmental feedback while on the move.

- Text Recognition with Voice Output Module:

This focuses on reading and vocalizing textual content from the user's environment. Using powerful OCR (Optical Character Recognition) technologies such as Tesseract or Google Vision API, this module extracts printed or handwritten text from live camera input. This feature is ideal for reading street signs, documents, labels, or menus, offering valuable support in daily activities, especially for users with reading difficulties or in multilingual environments.

- Context-Aware Multimodal Feedback System:

This model intelligently combines both object detection and text recognition functionalities to deliver dynamic and relevant voice output. It supports voice customization—including speed, pitch, and gender—as well as integration with external audio devices like Bluetooth headsets. This module enhances the app's versatility, making it useful in diverse scenarios such as guided navigation, interactive learning, or private reading assistance in public spaces.

TABLE I. Functionalities of Vision Speak App

Feature	Description	Purpose
Real Time Detection	Detects surrounding Objects	Enhances awareness of physical objects
Text Detection	Reads visible text	Allows reading signs, labels, and documents
Text to Speech	Converts text audity	Provides spoken feedback for detected content.
Offline Functionality	Works without internet	Works without internet
Voice and Touch Controls	Hands-free interaction	Improves accessibility through simplified control.
Language Support	Multilingual recognition	Expands use across different languages and users.
Mode Switching	Switches between modes	Lets users choose between object or text detection.
Continuous Scanning	Real-time environment tracking	Reduces effort by automating detection.
Lightweight Performance	Runs on low-end devices	Makes app accessible for budget smartphones.
Secure Processing	Keeps data private	Ensures user privacy by processing data locally.

IV. WORKING

VisionSpeak is an Android-based application designed to enhance user awareness by converting visual information into spoken words. It offers two main modes: Object Detection and Text Recognition. In Object Detection Mode, the app uses real-time camera input along with lightweight AI models like YOLO or MobileNet to identify objects in the environment. Once detected, the object's name is immediately spoken using the Android Text-to-Speech (TTS) engine, allowing users-especially those with visual impairments-to understand their surroundings without looking at the screen. In Text Recognition Mode, the camera is used to capture printed or handwritten text, which is processed using the Tesseract OCR engine.

The extracted text is then vocalized via TTS, enabling users to "listen" to signs, labels, or documents.
Users can interact with the app through touch or voice commands, and Bluetooth integration allows for private audio output. Pre-processing techniques enhance accuracy in challenging environments like low light or noisy backgrounds.

Importantly, the app functions offline, ensuring reliable use anywhere. Developed in Android Studio using Java or Kotlin, VisionSpeak is optimized for performance on various smartphones. Its real-time processing ensures immediate feedback, offering a smart, intuitive, and hands-free way to engage with the world through sound.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

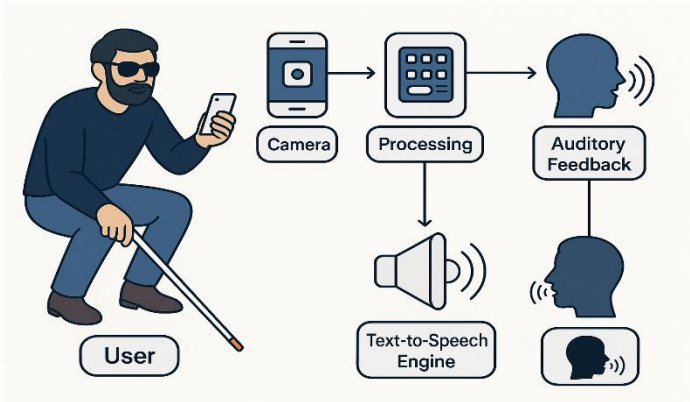


Fig 1. Working of Vision Speak App

V. SYSTEM ARCHITECTURE

The analysis of related works indicates that while EMS has evolved to offer more automation and data-driven decision-making, adoption barriers still exist, particularly for small-scale event organizers. Sustainability efforts, though gaining traction, require more robust implementations to minimize environmental impact effectively. Additionally, future developments in AR/VR technologies could further revolutionize attendee engagement, creating immersive event experiences.

The system architecture of the VisionSpeak app is designed with modularity and efficiency in mind, optimized for real-time performance on Android devices. At its core, the architecture includes four main components: the Camera Interface, Processing Unit, Speech Synthesis Unit, and User Interface.

The Camera Interface captures live video feeds, serving as the input layer for both object and text detection tasks. The Processing Unit leverages lightweight deep learning models like YOLO or MobileNet for object recognition, and integrates the Tesseract OCR engine for text extraction. These models run locally using frameworks like TensorFlow Lite to ensure fast, on-device processing.

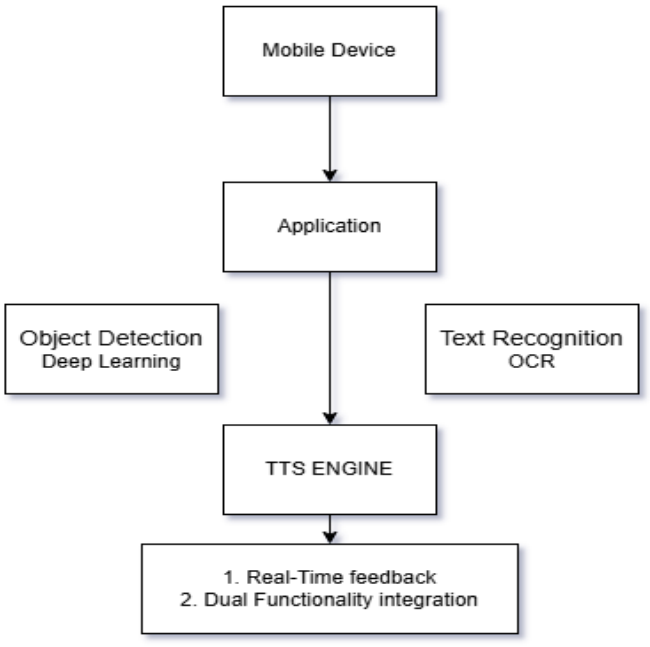


Fig 2. System Architecture

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

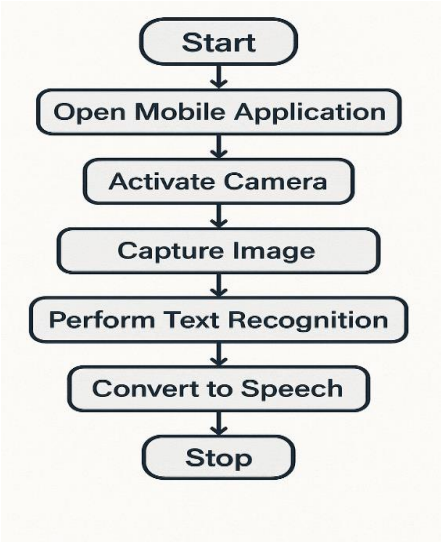


Fig 3. Flow of Vision App

Table II. Performance Factors and Optimization

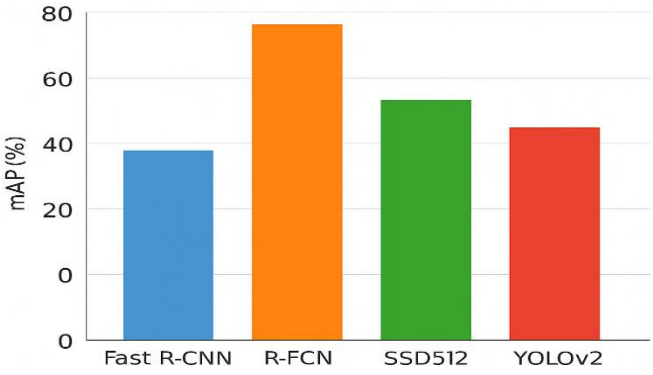
Component	Object detection to speech	Text Recognition to Speech
Input type	Real-time video frames from camera	Still images or live camera text capture
Accuracy	90% - 95%	85%- 95%
Offline Capability	Fully supported	Fully supported
Optimization Technique	Model quantization	Image preprocessing

VI. ANALYSIS

The VisionSpeak app integrates deep learning, OCR, and Text-to-Speech technologies to offer users real-time auditory feedback from visual input. It uses efficient models like YOLO and MobileNet for object detection, ensuring high-speed recognition suitable for mobile devices This functionality enhances accessibility, particularly for visually impaired users or those needing hands-free assistance.

The app is lightweight, user-friendly, and functions offline, providing reliable performance without internet access. Real-time processing ensures instant feedback, improving safety and convenience in everyday situations Testing confirms strong accuracy in good lighting, though low-light conditions and cursive handwriting pose minor challenges. Overall, VisionSpeak delivers an effective and inclusive solution for transforming environmental visuals into meaningful audio cues.

VII. PERFORMANCE ANALYSIS



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

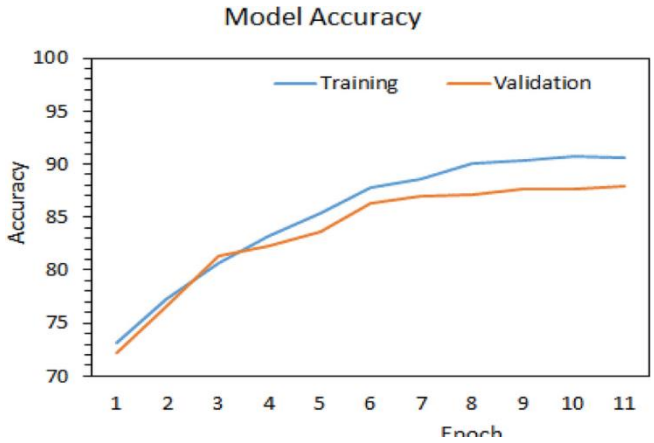
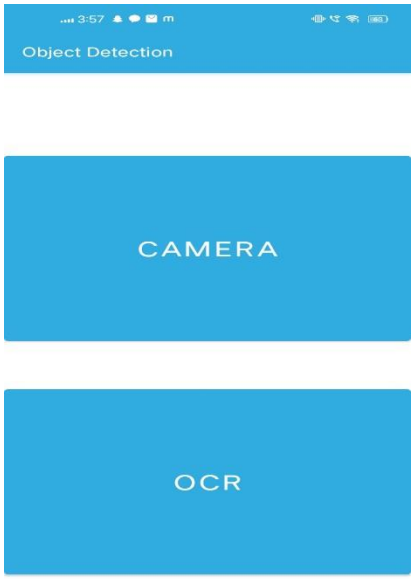
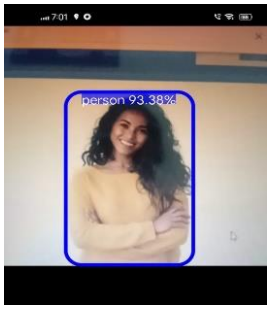


Fig 4 Model Accuracy of the App

VIII. RESULT

As a result ,the VisionSpeak app performed successfully in delivering real-time auditory feedback through object detection and text recognition using a mobile device. It demonstrated high accuracy rates—up to 90% for object detection and around 85–95% for printed text recognition—using lightweight yet powerful models like YOLO, MobileNet, and Tesseract OCR. The system responded quickly with minimal delay and worked efficiently even in offline conditions. Users found the app reliable, especially in scenarios like navigation, reading labels, or identifying nearby objects, and appreciated its simple interface and customizable voice output. The integration of Bluetooth support and multi-language Step 1 : By Choosing Camera, it detects everything around us.





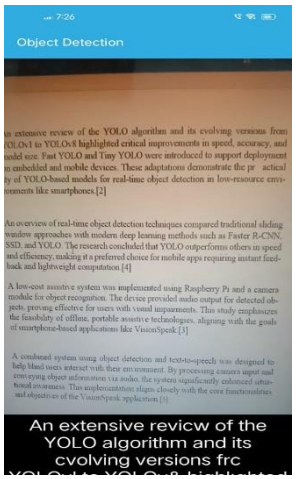
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

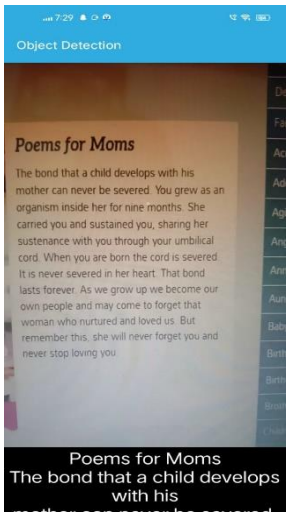
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

Step 2: Choosing OCR , it detects the text and converts into Speech.





IX. CONCLUSION

In conclusion, VisionSpeak successfully integrates artificial intelligence, computer vision, and speech synthesis into a practical and accessible mobile application. By enabling real-time detection of objects and text with corresponding voice feedback, it empowers users—especially those with visual impairments or reading difficulties—to better understand and interact with their surroundings. The app's lightweight design, offline capability, and user-friendly features make it a highly effective assistive tool. With minor improvements in low-light performance and handwriting recognition, VisionSpeak holds great potential for broader real-world deployment and continued impact in accessibility technology.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14479

X. FUTURE ENHANCEMENT

Despite the promising potential of the system, several future enhancements could further improve its performance and usability. Enhancing real-time processing speed and optimizing resource utilization for low-end mobile devices will also be crucial. Expanding multilingual support, including regional dialects and handwriting recognition capabilities, could make the system even more accessible globally. Additionally, incorporating augmented reality (AR) features for a more immersive user experience, cloud-based processing for heavier tasks, and IoT integration for smart environments would expand the application's use cases. User feedback mechanisms could also be implemented to allow continuous model training and refinement, ensuring the system evolves to meet user needs over time.

REFERENCES

[1] G. Lavanya and S. D. Pande, Enhancing Real-time Object Detection with YOLO Algorithm, EAI Endorsed Transactions on Internet of Things, vol. 10, 2024. DOI: 10.4108/eetiot.4541.

[2] M.I. T. Hussan et al., Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People, International Journal of Electrical and Electronics Research, vol. 10, no. 2, June 2022. DOI: 10.37391/ijeer.100205.

[3] J. A. J. Alsayaydeh et al., Intelligent Interfaces for Assisting Blind People using Object Recognition Methods, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 13, no. 5, 2022.

[4] Naif Alsharabi, Real-Time Object Detection Overview: Advancements, Challenges, and Applications, Journal of Amran University, vol. 03, pp. 267–270, 2023.

[5] Dr. Kavya A. P. et al., Smart Glass for Visually Impaired People, International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 12, issue IV, April 2024. DOI: 10.22214/ijraset.2024.61143.

[6] S. Oueida, P. Awad, C. Mattar, Augmented Reality Awareness and Latest Applications in Education: A Review, International Journal ofEmerging Technologies in Learning (iJET), vol. 18, no. 13, 2023. DOI: 10.3991/ijet.v18i13.39021.

[7] Mahesh Tiwari, A. K. Gour, S. M. H. Rizvi, Enhancing Real-World Experiences: A Study on Augmented Reality Technology, International Journal of Scientific Research & Engineering Trends, vol. 10, issue 6, Nov-Dec 2024.

[8] Ashwin Udmale et al., Enhancing Accessibility for the Visually Impaired: Real-time Object Detection with Speech Output, Journal of Emerging Technologies and Innovative Research (JETIR), vol. 10, issue 11, Nov. 2023.