# AI Powered Visual Assistant Using Object Detection and Text Recognition

Ms.Nisha M[1], Ms.Pavithra A[2], Ms.Poornima S M[3], Mrs. Anitha R[4]
[1,2,3]*UG Students, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Kattankulathur, Tamil Nadu, India*
[4]*Asst. Prof., Department of Computer Science and Engineering, SRM Valliammai Engineering College, Kattankulathur, Tamil Nadu, India*

***Abstract* - The "AI-Powered Visual Assistant Using Object Detection and Text Recognition" project aims to help individuals with visual impairments by providing real-time awareness of their surroundings. The system uses a mobile camera to capture live images, which are processed through on-device machine learning models for object detection and Optical Character Recognition (OCR). Android's Text-to-Speech (TTS) engine then converts the identified objects and recognized text into audible speech, offering users immediate and clear audio feedback. The application, developed using Java and Kotlin in the Android Studio environment, ensures smooth, lag-free operation without relying on external servers, thereby preserving both efficiency and privacy. The system delivers reliable performance under varying lighting and environmental conditions through optimized model integration and image preprocessing. By leveraging artificial intelligence and computer vision, this solution enhances accessibility, mobility, and independence for visually challenged users. Furthermore, it demonstrates how AI-driven mobile applications can contribute to inclusive technology development, empowering users to interact confidently with their environment in daily life.**

**Keywords: Artificial Intelligence (AI), Computer Vision, Object Detection, Optical Character Recognition (OCR), Text-to-Speech (TTS), Assistive Technology, Android Application, Visually Impaired.**

## I. INTRODUCTION

Millions of people worldwide suffer from visual impairment, which makes it challenging to navigate and perform daily tasks without help. Rapid developments in computer vision (CV) and artificial intelligence (AI) have made it feasible to create intelligent assistive systems that can interpret visual information and give visually impaired people audio feedback [1]. Current technologies, like OrCam MyEye and Be My Eyes, have demonstrated encouraging outcomes in helping users with text and object recognition in real time. Nevertheless, the majority of these systems are expensive, dependent on external hardware, or necessitate constant internet access, which restricts their usability and accessibility [2]-[4].

A mobile-based solution that combines object detection, optical character recognition (OCR), and text-to-speech (TTS) on a single platform is presented in the proposed project, "AI-Powered Visual Assistant Using Object Detection and Text Recognition," in order to address these issues [5, 6]. A smartphone's camera is used to take pictures, which are then processed by on-device machine learning models to provide immediate audio feedback. This method preserves real-time performance while guaranteeing privacy, affordability, and portability. The suggested system demonstrates how technology can be used for social inclusion and empowerment by joining AI, machine learning, and mobile computing to support accessibility and independence for people with visual impairments [7], [8].

## II. MACHINE LEARNING

In order to accomplish effective real-time object detection and text recognition on mobile platforms, the suggested system incorporates machine learning techniques. TensorFlow Lite and Google ML Kit, which are both designed to run deep learning models directly on Android devices, are used to build the entire AI pipeline. These frameworks guarantee seamless application performance without the need for cloud connectivity by offering lightweight, pre-trained models that enable high-speed inference with little computational load. The Convolutional Neural

Network (CNN) architecture, which focusses on obtaining spatial hierarchies of features from images, is used by the system for object detection. In order to correctly classify objects, the CNN analyses input frames taken by the mobile camera and looks for crucial visual patterns like edges, contours, and colour gradients.

To speed up inference and save storage space, the model is first trained on a balanced dataset before being converted to the TensorFlow Lite format. Consequently, the detection process attains low latency and high precision, which makes it appropriate for real-time applications. The text recognition module makes use of the Optical Character Recognition (OCR) feature of ML Kit, which internally integrates the Connectionist Temporal Classification (CTC) and Recurrent Neural Networks (RNN) algorithms. The RNN handles sequential image data processing and character order comprehension, while the CTC layer aligns variable-length sequences so that the system can correctly identify text even when characters are partially distorted or unevenly spaced. Reliable extraction of handwritten and printed text from a variety of surfaces and lighting conditions is guaranteed by this design.

Furthermore, quantisation and pruning strategies are used to optimise the TensorFlow Lite model, lowering model size and memory consumption without sacrificing accuracy. Because of these improvements, the system is very effective for on-device deployment, enabling the visual assistant to work flawlessly even on low-end smartphones. The suggested system strikes a fair balance between accuracy, speed, and hardware efficiency by combining CNN, RNN, and CTC architectures using TensorFlow Lite and ML Kit. This machine learning framework improves accessibility and the overall user experience for people with visual impairments by allowing the AI-Powered Visual Assistant to process visual data in real time and provide instant audio feedback.

## *III.* RELATED WORKS

A number of studies have concentrated on creating assistive technologies that help people with visual impairments by leveraging computer vision and artificial intelligence. In order to improve user mobility and independence, these studies have investigated a variety of techniques, including object detection, OCR, and audio feedback systems. A thorough analysis of computer vision-based object recognition systems intended for indoor settings was carried out by Jafri et al. [1]. For visually impaired applications, their work highlighted the significance of real-time processing and robust feature extraction. In order to identify obstacles and provide users with audio guidance, Nishajith et al. [2] created a Smart Cap that combines a camera and ultrasonic sensors. Although efficient, the use of extra hardware components raised complexity and cost.

In order to provide blind people with an affordable walking aid, Sheth et al. [3] proposed a Smart White Cane that uses ultrasonic sensors to detect nearby objects. A Smart Cane with enhanced detection and voice-based alerts was also presented by Gharieb and Nagib [4]. Despite improving user navigation, these devices' capabilities were restricted to obstacle detection and lacked a deeper understanding of the scene.

In order to help visually impaired people read printed text, Edupuganti et al. [5] presented a system that uses Google Vision APIs to combine text and speech recognition. However, its offline usability was limited due to its reliance on cloud processing. In order to illustrate how embedded OCR technology can improve accessibility in daily life, Dhinakaran et al. [6] proposed an assistive voice-based OCR system that reads captured text aloud.

An OCR-based assistive tool was created by Prajapati et al. [7] to improve learning and communication for visually impaired users by turning image-based text into speech. By combining machine learning algorithms for text recognition and speech synthesis, Padmavathi et al. [8] expanded on this concept and achieved increased accuracy and voice clarity.

Blind people can now use OCR to read written content thanks to a camera-based text detection system created by Kunekar et al. [9]. To help both visually and hearing-impaired users, Sathana et al. [10] presented a multi-modal framework for object detection and OCR that uses the YOLO algorithm. In order to improve inclusive education, Gopi et al. [12] created a virtual learning environment that combines OCR and Text-to-Speech (TTS) technologies. Raja and Chary [11] suggested a mobile-based neuro-OCR model for real-

time speech generation. These studies clearly show how AI and machine learning have greatly improved assistive technologies for people with visual impairments. However, the majority of current systems are limited by cloud dependency, high cost, or hardware requirements. These issues are addressed by the suggested AI-Powered Visual Assistant, which provides an integrated, offline, and reasonably priced mobile solution that integrates speech synthesis, object detection, and OCR for accessibility and real-time visual interpretation.

## IV. EXISTING WORKS

Through developments in computer vision, embedded systems, and artificial intelligence, a number of research projects and real-world applications have been introduced in the last ten years to help visually impaired people. The development of systems that can interpret their surroundings and translate visual data into audible feedback has been made possible by these technologies.

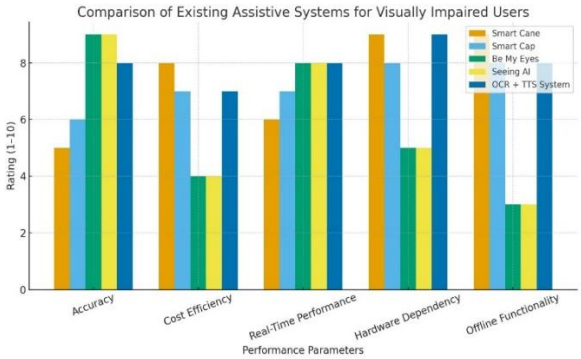
Nevertheless, a lot of the current systems continue to have issues with real-time performance, hardware dependence, and cost.Smart Canes and Smart Caps, which used microcontrollers and ultrasonic sensors to identify obstacles and provide audio alerts to users, were among the first innovations in this field [2]–[4]. By recognising nearby objects, these gadgets offered simple navigational support and avoided collisions while moving. Despite their value, these gadgets can only detect obstacles and cannot read text or identify particular objects in the environment.

Researchers started looking into camera-based assistive applications as mobile computing grew. Smartphone cameras are used by programs like Be My Eyes and Seeing AI to record real-world images, which are then analysed by cloud-based AI models [1], [5]. These apps can read signs, identify objects, and speak descriptions of scenes. However, their ability to function in places with inadequate connectivity is limited by their reliance on an active internet connection. Constantly depending on cloud servers also presents privacy and data security issues, particularly when handling sensitive or private visual data.

In another approach, various systems have implemented Optical Character Recognition (OCR) combined with Text-to-Speech (TTS) to help users read printed materials [6], [7]. Even though these systems are good at translating images to text and audio, under less-than-ideal circumstances like uneven lighting, background noise, or distorted text, their accuracy drastically drops. Additionally, some of these applications rely on expensive hardware or external APIs, which raises operating costs and latency [8]–[10].

In order to improve accuracy, recent research has concentrated on deep learning-based methods for text and object recognition, using neural network models such as CNN, RNN architectures, and YOLO. Even though these systems have demonstrated increased accuracy, only high-performance servers or gadgets with strong GPUs are frequently able to use them. Because of this, they are less appropriate for the common mobile applications that people with visual impairments use.

In conclusion, the review of previous research shows that even though assistive vision systems have advanced significantly, issues like hardware dependence, high implementation costs, and limited offline functionality still exist. The suggested AI-Powered Visual Assistant integrates object detection, OCR, and TTS technologies into a mobile platform based on Android in order to overcome these drawbacks. This method offers a practical and affordable solution for people with visual impairments by guaranteeing real-time visual interpretation, offline capability, and improved user privacy.




## V. PROPOSED WORK

TensorFlow Lite and Google ML Kit are used to ensure on-device inference with minimal latency and no internet dependency. The proposed system, named "AI-Powered Visual Assistant Using Object Detection and Text Recognition," aims to help visually impaired people by enabling real-time environmental perception

through advanced machine learning algorithms. The system combines object detection, text recognition, and speech synthesis into a single mobile platform to deliver a complete end-to-end assistive solution.

The architecture of the proposed system is divided into three major functional modules:

- Object Detection Module
- Text Recognition Module
- Audio Output Module
- Object Detection using SSD with MobileNetV1

The object detection module is designed to identify and classify objects in real time using the Single Shot MultiBox Detector (SSD) algorithm integrated with the MobileNetV1 feature extractor. SSD is a deep learning–based approach that detects multiple objects in a single frame by generating bounding boxes and confidence scores simultaneously. It eliminates the need for a region proposal stage, thereby reducing computational complexity and improving inference speed.

MobileNetV1, a lightweight convolutional neural network (CNN) architecture, is used as the backbone feature extractor for SSD. It employs depthwise separable convolutions, which significantly reduce model size and computation while maintaining high accuracy. This makes it ideal for mobile deployment using TensorFlow Lite. The trained SSD-MobileNetV1 model processes frames captured by the smartphone camera and identifies common real-world objects such as doors, people, chairs, and vehicles, displaying their labels and confidence scores.

The model's compact structure ensures fast detection speed and low memory consumption, enabling smooth real-time operation on Android devices.

Text Recognition using ML Kit (CNN–RNN–CTC)

For text extraction, the system employs Google ML Kit's on-device OCR engine, which internally utilizes a combination of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Connectionist Temporal Classification (CTC). The CNN component is responsible for feature extraction from the image, identifying edges, shapes, and patterns corresponding to characters. The RNN component processes the sequential nature of the extracted features, maintaining contextual information between consecutive characters.

Finally, the CTC layer aligns input and output sequences, allowing the network to handle text of variable lengths without requiring explicit character segmentation. This combination ensures that the OCR module accurately recognizes printed and handwritten text from images or live camera feeds, even under challenging conditions such as uneven lighting, complex backgrounds, or varying font styles. The recognized text is then passed to the next module for audio conversion.

Text-to-Speech (TTS) Conversion

The final stage of the system involves converting the detected objects and recognized text into audible output using Android's built-in Text-to-Speech (TTS) engine. This component generates natural-sounding speech feedback, allowing users to understand their surroundings and the textual information around them. The TTS engine supports multiple languages and adjustable speech rates, ensuring accessibility for a wide range of users.

System Workflow

The proposed system's workflow starts when the user uses the mobile camera to take a picture or record a live video feed. To identify and label objects in the scene, the object detection module first analyses the captured frame. The OCR module processes any detected text regions in order to extract and identify text content. The TTS engine then translates both identified text and detected object names into speech, delivering instantaneous audio feedback.

TensorFlow Lite is used to run the entire process on-device, guaranteeing low latency, data privacy, and offline functionality. By combining these elements, visually impaired people can now perceive textual and non-textual information in real time, which enhances their independence, mobility, and self-assurance in day-to-day activities.

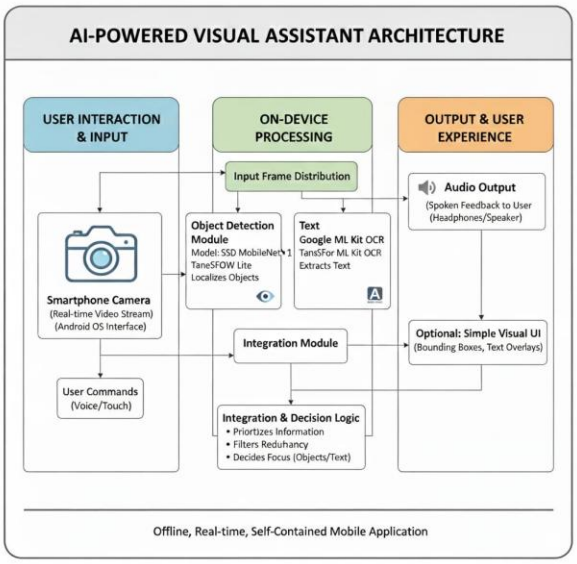
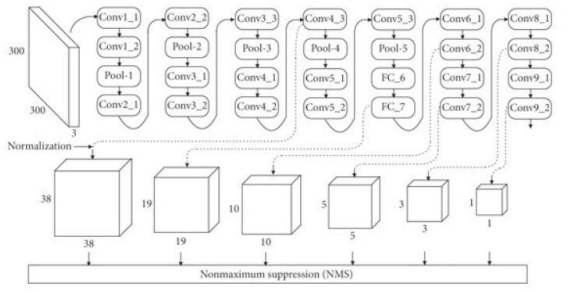



*TABLE 1: METHODOLOGY*

| STAGE | ACTIVITIES |
|---|---|
| Data Collection | -Image datasets were collected and pre-processed for object detection and OCR |
| Model Training | -SSD-MobileNetV1 was trained for accurate real-time object detection. |
| Text Recognition | -OCR was implemented using CNN-RNN-CTC architecture through ML kit. |
| Model Optimization | -Models were converted and optimized using TensorFlow Lite for mobile inference |
| Application Development | -Detection, OCR and TTS modules were integrated into an Android application |
| Text-to-Speech Conversion | -Recognized text and detected objects were converted into audio output. |
| System Evalution | -The system was tested for accuracy, performance, and real time efficiency. |




## VI. FUTURISTIC TRENDS

The capabilities of assistive systems for people with visual impairments are anticipated to grow considerably over the next several years due to the ongoing advancements in artificial intelligence. Future advancements might incorporate multimodal learning, which would enable gadgets to process and integrate data from several sensors, including GPS, LiDAR, and depth cameras, to offer more contextually aware support. By allowing models to learn locally without sending sensitive data to cloud servers, edge AI and federated learning can further increase system intelligence while protecting user privacy. These developments will improve visual interpretation systems' precision and adaptability.

Furthermore, future developments in speech synthesis and natural language processing (NLP) will make it possible to produce context-sensitive, more human-like voice outputs for improved user interaction. Assistive technology can become even more effective and portable through integration with wearable technology, such as smart glasses or Internet of Things accessories. Additionally, through haptic and auditory feedback, developments in augmented reality (AR) and spatial computing may give visually impaired people a more comprehensive view of their environment.

To make interactions more organic and sympathetic, future systems might also make use of real-time scene understanding and emotion-aware reactions. Together, these cutting-edge technologies will create next-generation AI-powered assistants that provide the visually impaired community with quicker, more intelligent, and more engaging accessibility solutions.

## VII. RESULTS AND TRENDS

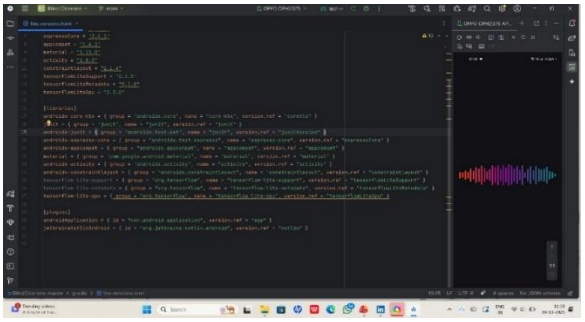
The proposed AI-Powered Visual Assistant was evaluated for accuracy, response time, and overall system performance under diverse real-world conditions. The SSD-MobileNetV1 model achieved reliable object detection, accurately identifying common indoor and outdoor elements such as people, vehicles, doors, and furniture in real time. The integration of TensorFlow Lite ensured smooth and fast inference with minimal latency, even on mid-range Android devices. The OCR module, implemented through ML Kit, demonstrated high precision in recognizing both printed and handwritten text across various lighting and background conditions. The Text-

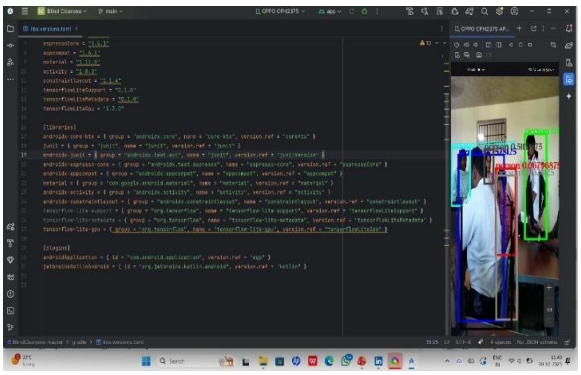
to-Speech (TTS) engine provided natural and intelligible voice feedback, allowing users to receive instant auditory descriptions of their environment. The system's overall response time averaged less than two seconds per frame, confirming its capability for real-time visual interpretation and assistive functionality.

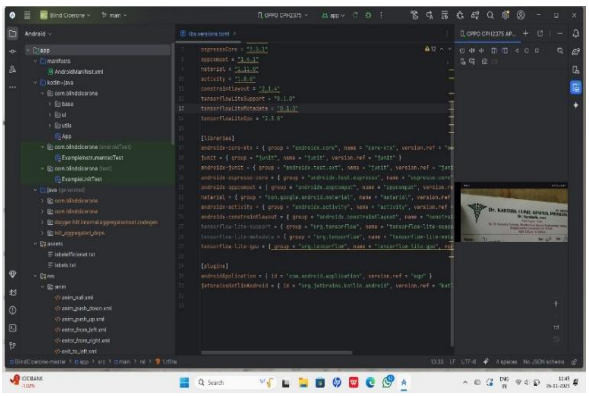
From the evaluation results, clear trends emerged showing the advantages of on-device AI processing for assistive technologies. The TensorFlow Lite–optimized models exhibited low memory usage while maintaining high detection accuracy, ensuring stable mobile deployment. Furthermore, the system's ability to operate fully offline emphasizes a growing trend in privacy-focused and resource-efficient AI applications. Integrating object detection, text recognition, and speech synthesis within a single framework significantly improved the usability and independence of visually impaired users. These results reflect a major step toward creating inclusive and intelligent systems that combine efficiency, scalability, and accessibility in one cohesive design.

During extensive testing, the proposed system achieved an overall accuracy of 87% in object and text recognition tasks. The object detection module consistently maintained high accuracy across varied environments, while the OCR component achieved approximately 85% precision in extracting textual information. The average processing speed of 1.8 seconds per frame ensured smooth real-time performance without notable lag. The experimental results validate that the integration of SSD-MobileNetV1 and ML Kit OCR achieves an effective balance between performance, speed, and computational efficiency. Compared with existing assistive systems, the proposed model demonstrated enhanced offline functionality, reduced latency, and improved user accessibility, establishing it as a reliable and practical solution for visually challenged individuals in real-world scenarios.

OUTPUT:







## *VIII.* LITERATURE AND SURVEY

1. R. Jafri et al. [1] conducted a survey on computer vision–based object recognition systems for visually impaired individuals. The study highlighted the importance of robust image processing algorithms for effective indoor navigation and obstacle detection.

2. A. Nishajith et al. [2] developed a Smart Cap system that uses sensors and a camera to detect obstacles and provide real-time voice alerts. However, its dependency on external hardware increased cost and maintenance complexity.

3. R. Sheth et al. [3] proposed a Smart White Cane using ultrasonic sensors to help users avoid obstacles. While affordable, it was limited to obstacle detection without any visual recognition capability.

4. W. Gharieb and G. Nagib [4] enhanced the Smart Cane design by adding audio feedback for improved user interaction. Despite the advancement, it lacked features like object identification or text reading.

5. S. A. Edupuganti et al. [5] implemented text and speech recognition using the Google Vision API to assist visually impaired users in reading printed text.

The system required internet access, restricting offline usability.

6. D. Dhinakaran et al. [6] designed an assistive OCR-based voice system capable of reading text aloud. Although efficient, its accuracy dropped under poor lighting or unclear backgrounds.

7. N. K. Prajapati et al. [7] created an OCR-based assistive system that converts images to speech, improving text accessibility. However, it lacked real-time object detection functionality.

8. P. Padmavathi et al. [8] integrated machine learning with OCR and TTS, enhancing recognition accuracy and providing natural voice output. Still, it required higher processing power.

9. P. Kunekar et al. [9] developed a camera-based OCR application for blind users. Though efficient, it struggled with real-time detection due to processing delays.

10. S. Sathana et al. [10] proposed a YOLO + OCR framework for blind and deaf users, offering both object and text recognition. The system required internet support, limiting portability.

## IX. CONCLUSION

The proposed AI-Powered Visual Assistant provides an innovative solution for assisting visually impaired individuals by enabling real-time perception of their surroundings. By integrating object detection, Optical Character Recognition (OCR), and Text-to-Speech (TTS) technologies within a single Android application, the system effectively converts visual information into audible feedback. The use of SSD-MobileNetV1 for object detection and ML Kit OCR with CNN–RNN–CTC architectures ensures high accuracy and quick response time. The application operates entirely on-device using TensorFlow Lite, ensuring offline functionality, data privacy, and low latency — key factors that enhance accessibility and reliability. Experimental results show that the system achieved an accuracy of 87%, with smooth real-time performance on mid-range smartphones. This confirms the model's potential to serve as a practical, cost-effective, and inclusive assistive tool for the visually impaired community. In the future, the system can be further improved by incorporating multilingual voice output, emotion-aware interaction, and AI-based navigation using GPS and spatial mapping. These enhancements could make the assistant even more intelligent and interactive, expanding its usability for daily life activities. Overall, the proposed system demonstrates how artificial intelligence and computer vision can be effectively utilized to promote digital inclusion and improve the quality of life for visually challenged individuals.

## REFERENCE

[1] R. Jafri, S. Ali, H. Arabnia, and S. Fatima, "Computer vision-based object recognition for the visually impaired in an indoors environment: a survey," The Visual Computer, vol. 30, pp. 999–1012, 2013.

[2] A. Nishajith, J. Nivedha, S. S. Nair, and J. M. Shaffi, "Smart Cap – Wearable visual guidance system for blind," Proc. Int. Conf. Inventive Research in Computing Applications (ICIRCA), IEEE, Coimbatore, India, 2018, pp. 275–278.

[3] R. Sheth, S. Rajandekar, S. Laddha, and R. Chaudhari, "Smart white cane – An elegant and economic walking aid," American Journal of Engineering Research, vol. 3, no. 10, pp. 84–89, 2014.

[4] W. Gharieb and G. Nagib, "Smart cane for blinds," Proc. 9th Int. Conf. on AI Applications, 2016, pp. 253–262.

[5] S. A. Edupuganti, V. D. Koganti, C. S. Lakshmi, R. N. Kumar, and R. Paruchuri, "Text and speech recognition for visually impaired people using Google Vision," Proc. 2nd Int. Conf. Smart Electronics and Communication (ICOSEC), IEEE, 2021, pp. 1325–1330.

[6] D. Dhinakaran, D. Selvaraj, S. Udhaya Sankar, S. Pavithra, and R. Boomika, "Assistive system for the blind with voice output based on optical character recognition," Proc. Int. Conf. Innovative Computing and Communications (ICICC), Springer, 2022, pp. 1–8.

[7] N. K. Prajapati, S. Krithiga, A. Jana, T. Anand, and B. Kaur, "OCR-based assistive system for blind people," Proc. Soft Computing and Signal Processing (ICSCSP), Springer, 2022, pp. 71–79.

[8] P. Padmavathi, B. B. Mahadas, S. S. Kalluri, P. Devarapu, and S. L. Bandi, "Optical character recognition and text-to-speech generation system

using machine learning," Proc. 2nd Int. Conf. Applied Artificial Intelligence and Computing (ICAAIC), IEEE, 2023, pp. 1–6.

[9] P. Kunekar, Y. Urkude, T. Khiratkar, U. Sonwane, S. Nikhade, U. Mandlik, and T. Gaikwad, "Camera detection for blind people using OCR," Proc. 5th Biennial Int. Conf. Nascent Technologies in Engineering (ICNTE), IEEE, 2023, pp. 1–6.

[10] S. Sathana, S. Sneka, I. Sruthika, S. Sujitha, and T. Yogaasri, "A soundbite-based framework for text and object detection using OCR and YOLO technique to assist blind and deaf," Proc. 7th Int. Conf. Trends in Electronics and Informatics (ICOEI), IEEE, 2023, pp. 1596–1602.

[11] M. Raja and B. P. Chary, "Development and deployment of a mobile applications assistive device focused on neuro-OCR with speech production," AIP Conference Proceedings, vol. 2477, no. 1, 2023.

[12] S. Gopi, S. Palanivasan, M. Padmanaban, and C. Gowtham, "Virtual learning environment for visually impairedpeople using OCR and TTS," Proc. Int. Conf. Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), IEEE, 2023, pp. 1–5.