

VisionAssist: A Voice-Based Assistive System for the Visually Impaired

Anushka Jadhav¹ Nita Jadhav² Komal Jadhav³ Kunal Jadhav⁴ Prof.Yogesh.H.Palve⁵

^{1,2,3,4,5}Department of Information Technology

^{1,2,3,4,5}MVP's Rajashri Shahu Maharaj Polytechnic, Gangapur Road, Nashik, Maharashtra, India

Abstract — Visually impaired people find it difficult to know who or what is in front of them, and they depend on other people for navigation, reading and daily tasks. Existing aids such as the white cane detect only obstacles within contact range, smartphone applications are not hands-free and usually need an internet connection, and smart glasses are costly and heavy. This paper presents VisionAssist, a low-cost, wearable and voice-based assistive system. The user presses a single push button to capture an image with an ESP32-CAM, while an HC-SR04 ultrasonic sensor measures the distance to obstacles ahead. The ESP32 sends the image and the distance reading to a backend server over Wi-Fi. The backend identifies people, animals and objects using a deep-learning object detector, reads printed text using optical character recognition (OCR), checks the obstacle distance, and converts the result into speech using a text-to-speech (TTS) engine, which is played through a speaker or earphone. The system deliberately avoids heavy scene description so that the spoken output stays short, fast and useful.

Keywords: ESP32-CAM, Ultrasonic Sensor, Object Detection, OCR, Text-to-Speech, Assistive Technology

I. INTRODUCTION

Vision is the main sense through which people understand their surroundings. A person with visual impairment loses much of this awareness, which limits independent movement and makes everyday activities such as crossing a road, finding an object or reading a notice difficult. Traditional aids tell the user very little about what is actually ahead. A cane, for example, can sense a nearby obstacle but cannot say whether it is a person, an animal or a parked vehicle.

The aim of this work is to build a simple device that works like a “search engine for the surroundings” for a blind user: it looks at the scene in front of the user and tells the user, by voice, what is there, for example “person ahead” or “cat on the left”. In VisionAssist an ESP32-CAM captures the image, an ultrasonic sensor detects nearby obstacles, a backend server processes the data, and the result is spoken to the user. Input is by a single push button and output is by voice, so the device can be used without a smartphone screen and without any training in complex controls.

Recent advances in real-time object detection, open-source OCR and neural speech synthesis, together with cheap Wi-Fi enabled camera modules, make it possible to build such a device at low cost. This paper describes the problem, the existing systems and their limitations, the proposed system, its architecture and hardware, its modules, and its applications.

II. LITERATURE REVIEW

- 1) Redmon et al. (2016) introduced YOLO (You Only Look Once), which treats object detection as a single regression problem and predicts bounding boxes and class probabilities directly from the full image in one pass of a neural network. Because of this single-pass

design the detector runs in real time. VisionAssist uses this idea of fast, single-pass detection to identify people, animals and objects in the captured image [1].

- 2) Dakopoulos and Bourbakis (2010) presented a survey of wearable obstacle-avoidance electronic travel aids for the blind. The survey compares devices on the basis of sensors, size, user interface and the way information is delivered to the user, and shows that distance-sensing devices and audio feedback are widely used. This supports the use of an ultrasonic sensor and voice warnings in VisionAssist [2].
- 3) Poggi and Mattochia (2016) proposed a wearable mobility aid based on embedded 3D vision and deep learning that recognises objects in front of the user and communicates them through audio. Their work shows that a wearable and embedded design is practical, which motivates the small, body-worn form of VisionAssist [3].
- 4) Vinyals et al. (2015) developed Show and Tell, a neural image caption generator that combines a convolutional network with a recurrent language model to produce a sentence describing an image. Although such scene description is useful, it needs heavy computation and produces long sentences, so VisionAssist uses plain object detection instead [4].
- 5) Smith (2007) gave an overview of the Tesseract OCR engine, an open-source engine that converts images of printed text into machine-readable text. It is used in VisionAssist as the basis of the document-reading module [5].
- 6) Shen et al. (2018) presented Tacotron 2, which converts text to mel spectrograms and then uses a WaveNet vocoder to generate speech that is close to human quality. This work shows that natural-sounding speech output is feasible and supports the text-to-speech module of VisionAssist [6].

III. EXISTING SYSTEM

Several solutions are available today for visually impaired people, but each of them has limitations:

- The white cane detects only contact-range obstacles and gives no information about what the obstacle is.
- Smartphone applications are not hands-free, need the phone to be held steadily, and often depend on an internet connection or on a human volunteer.
- Smart glasses can provide recognition but are expensive and heavy.
- Scene-description systems based on image captioning are compute-heavy, slow and give long, wordy output.
- Most systems do not combine voice output, object recognition and obstacle alerts in one low-cost device.

IV. PROPOSED SYSTEM

The proposed work is the design and development of VisionAssist, a hands-free and voice-based assistive device. The user presses a push button; one press makes the ESP32-

CAM capture an image. At the same time the HC-SR04 ultrasonic sensor measures the distance to the nearest obstacle in front of the user. The ESP32 sends the image and the distance value to the backend server over Wi-Fi.

At the backend, the object-detection module identifies the people, animals and objects in the image, the document-reading module extracts any printed text using OCR, and the obstacle-distance module checks whether the user is too close to an obstacle. A decision and response generator combines these results into a short sentence, and the text-to-speech module converts it into audio that is played through a speaker or earphone. The system does not generate full scene descriptions, so the spoken message remains short and quick to understand.

V. SYSTEM ARCHITECTURE

The architecture of VisionAssist has three parts: the input side, the backend server, and the output side. The input side consists of the push button, the ESP32-CAM and the ultrasonic sensor. The backend server contains the object detection, document reading (OCR), obstacle distance check, decision and response generator, and text-to-speech modules. The output side is the speaker or earphone worn by the user. The flow of data is: button press → image capture → ESP32 → backend → voice output.

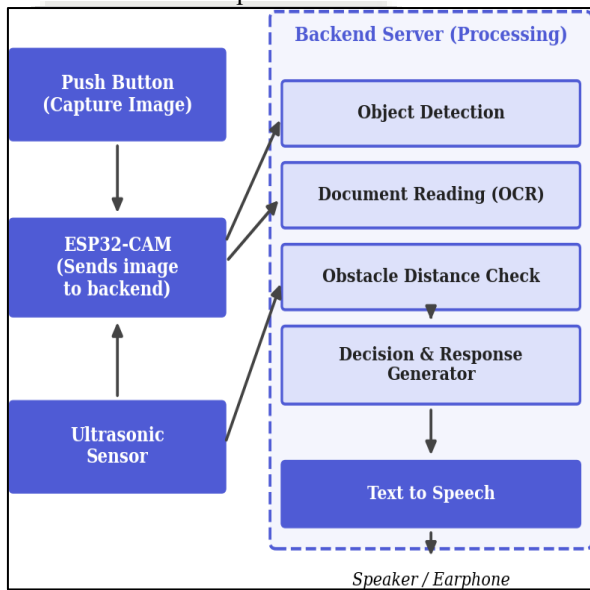


Fig. 1: System Architecture

VI. BLOCK DIAGRAM AND HARDWARE

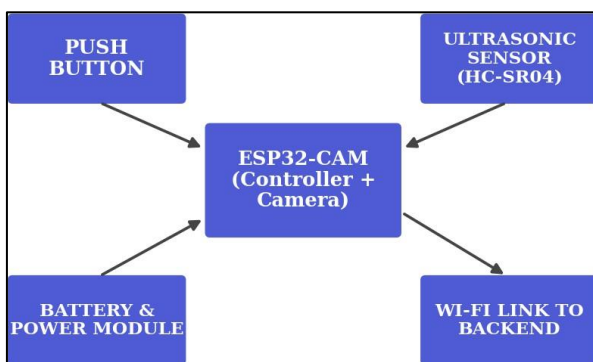


Fig. 2: Block diagram

- 1) ESP32-CAM: This is the main controller with an on-board camera. It captures the image when the button is pressed and sends the image and sensor data to the backend over Wi-Fi.
- 2) Ultrasonic sensor (HC-SR04): It sends out ultrasonic pulses and measures the time taken by the echo to return, which gives the distance to an obstacle in front of the user.
- 3) Push button: A single push button is the only input control. One press captures an image.
- 4) Battery and power module: It supplies power to the ESP32-CAM and the sensor so that the device is portable and wearable.
- 5) Speaker / earphone: It plays the spoken result so that the user hears it immediately.

VII. WORKING PRINCIPLE

When the user presses the push button, the ESP32-CAM captures a still image. The ESP32 also reads the distance from the ultrasonic sensor. Both are sent to the backend server over Wi-Fi. The backend runs the following steps:

- 1) the object detector finds people, animals and objects in the image and returns their names;
- 2) the OCR engine checks whether the image contains printed text and, if it does, extracts it;
- 3) the distance value is compared with a safety threshold to decide whether an obstacle warning is needed;
- 4) the decision and response generator selects the most useful information and forms a short sentence, for example "Two persons ahead, obstacle very close";
- 5) the text-to-speech engine converts the sentence into speech, which is played to the user.

VIII. MODULES

- 1) Object Detection: Identifies people, animals and common objects in the captured image and reports them by name. Scene description is intentionally not used.
- 2) Document Reading: Uses OCR to read printed text such as books, notices and labels, and passes the text to the speech module.
- 3) Obstacle Distance Check: Uses the ultrasonic sensor reading to warn the user when an obstacle is close.
- 4) Decision and Response Generator: Combines the outputs of all modules into one short, clear message.
- 5) Text to Speech: Converts the final message into natural-sounding voice output.

IX. REAL-WORLD APPLICATIONS

- 1) Outdoor navigation: spoken awareness of people, animals and obstacles on the road.
- 2) Education: reading books, notices and classroom boards aloud.
- 3) Home and daily living: finding objects, reading labels and supporting independence.
- 4) Public places: railway stations, hospitals, malls and offices.
- 5) Elderly care: assisting elderly people who have weak vision.
- 6) Smart city and accessibility: a part of inclusive and accessible infrastructure.

X. ADVANTAGES

- 1) Hands-free operation with a single push button.
- 2) Low cost, light weight and portable compared with smart glasses.
- 3) Tells the user what is ahead, not only that something is ahead.
- 4) Combines object detection, obstacle warning and document reading in one device.
- 5) Short voice messages instead of long scene descriptions.

XI. FUTURE SCOPE

The system can be extended with multi-language voice support, GPS-based navigation, face recognition of known people, on-device AI so that the device can work offline, and a compact smart-glasses form factor.

XII. CONCLUSION

VisionAssist gives visually impaired users a low-cost way to know what is around them. A push-button image capture with an ESP32-CAM, an ultrasonic sensor for obstacles, backend processing for object detection, document reading and text-to-speech, and voice output combine into a practical assistive device that improves safety and independence. By keeping the output short and avoiding heavy scene description, the system stays fast and easy to use. With the future extensions, VisionAssist can grow into a complete wearable assistant for people with visual impairment.

REFERENCES

- [1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE CVPR, 2016, pp. 779-788.
- [2] D. Dakopoulos and N. G. Bourbakis, "Wearable obstacle avoidance electronic travel aids for blind: A survey," IEEE Trans. Syst., Man, Cybern. C, vol. 40, no. 1, pp. 25-35, Jan. 2010.
- [3] P. Poggi and S. Mattoccia, "A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning," in Proc. IEEE ISCC, 2016, pp. 208-213.
- [4] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE CVPR, 2015, pp. 3156-3164.
- [5] R. Smith, "An overview of the Tesseract OCR engine," in Proc. IEEE ICDAR, 2007, vol. 2, pp. 629-633.
- [6] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. IEEE ICASSP, 2018, pp. 4779-4783.