An Intelligent Voice-Based Application for Assisting Visually Impaired Users in Navigating Real-World Environments

Mariam El-Saqa[1], Aliaa Rehan Youssef[1]
[1]*Systems and Biomedical Engineering Department, Cairo University, Giza*
mariam.aziz99@eng-st.cu.edu.eg, aliaa.rehan@eng.cu.edu.eg

Abstract - Visually impaired individuals often face significant challenges in perceiving and understanding their surrounding environment, especially in unfamiliar or dynamic outdoor settings. While several assistive technologies have emerged in recent years, many still lack intuitive interaction, detailed scene understanding, and real-time feedback. To address these limitations, we present VisionPal, a voice-controlled desktop application designed to assist blind and visually impaired users in navigating the world more independently. VisionPal integrates automatic speech recognition (ASR), text-to-speech (TTS), and a multimodal vision-language model to generate natural language descriptions of real-world scenes captured via camera or image selection. The user interacts entirely through voice commands, choosing between Arabic or English, selecting image input modes, and receiving audio feedback that describes scene contents. In addition to initial scene description, users can ask follow-up questions about objects, positions, or safety-related details in the image. To evaluate the model's performance, we collected images under diverse lighting and environmental conditions and analyzed the clarity of the generated descriptions. Furthermore, user feedback was gathered via a structured form to assess real-world applicability. Results indicate that VisionPal has strong potential as an accessible, multilingual assistant for individuals with vision loss. This work represents an initial prototype designed for desktop use; future development will focus on deploying VisionPal as a mobile application to enhance portability and real-time accessibility for users in everyday environments.

Keywords: *VisionPal, visually impaired, blind, image captioning, speech recognition, text-to-speech, assistive technology, voice-controlled application, scene understanding, multimodal model, LLMs.*

Introduction

According to the World Health Organization (WHO), over 2.2 billion people globally suffer from some form of vision impairment or blindness [1]. Vision loss significantly impacts individuals' ability to navigate safely and independently, especially in unfamiliar or dynamic environments. Traditional navigation aids such as white canes, guide dogs, and human assistants provide a degree of support but come with limitations in terms of cost, convenience, and user autonomy [2], [3].

Recent advances in artificial intelligence, computer vision, and speech technologies have opened new possibilities for creating intelligent assistive tools. Research has shown that visually impaired users benefit from auditory feedback systems that offer scene understanding and obstacle awareness [4], [5]. However, many existing systems either rely on complex wearable hardware or require significant training and adjustment periods [6]. Despite these advancements, challenges remain. Many current solutions are cost-prohibitive, lack multilingual support, or require extensive user training. Moreover, few applications offer seamless voice-controlled interaction combined with detailed scene understanding, limiting their practicality for daily use.

To overcome challenges such as discomfort in public spaces, limited portability, high costs and time requirements for training, and continuous reliance on another person, numerous navigation assistant systems have been proposed in the literature [7-9]. These systems vary in their intended environments, with some tailored for indoor use, others for outdoor settings, and a few designed to function in both [10]. Nevertheless, many of these devices are often found to be inconvenient and may contribute to a sense of social stigma among users [11].

To address these limitations, researchers have explored various technological solutions, particularly in artificial intelligence and machine learning, for developing advanced navigation aids for individuals with visual impairments. In recent years, deep learning models have gained significant attention for their potential in obstacle detection within such systems. However, selecting a model that balances low inference time and small memory footprint—both critical for real-time use—requires careful study and design considerations.

Given that portability is a crucial aspect of effective navigation systems, there has been growing interest in leveraging the capabilities of smartphones [12]. These devices provide a promising platform due to their built-in sensors, high-quality cameras, increasing processing power, and widespread availability [12]. Moreover, many solutions still require external specialized hardware, which can be expensive and cumbersome.

In contrast, VisionPal operates primarily through cloud-based services such as Together AI for vision-language understanding, eliminating the need for heavy local hardware. This approach leverages internet connectivity to provide advanced scene captioning and question answering in real-time without sacrificing portability or user convenience. Additionally, VisionPal integrates Google Speech Recognition and Google Text to Speech for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) models respectively, ensuring reliable voice control. While currently dependent on online services, future development plans include transitioning to locally run speech and vision models to enable offline functionality and enhance privacy.

We introduce VisionPal, a desktop-based assistive application designed to provide a natural, voice-driven interface for users with visual impairments. VisionPal enables users to interact via voice commands, capture or upload images, and receive comprehensive audio feedback describing the scene. The application integrates ASR, TTS, and a vision-language model (LLaVA) to perform image captioning and question answering.

VisionPal aims to balance accessibility, affordability, and performance by focusing on a fully voice-controlled experience that works in real-time and can later be adapted to mobile platforms. Unlike many prior solutions, VisionPal reduces reliance on external hardware while increasing interactivity and environmental understanding.

Unlike systems that provide simple object detection or label-based results, VisionPal leverages advanced multimodal models to interpret scene details and answer follow-up questions. This allows users to better assess safety, locate specific elements, and gain spatial awareness through detailed audio descriptions.

In addition to developing the application, we evaluate its performance under various lighting conditions and environments, and collect human feedback through structured forms to assess the accuracy and usefulness of its responses. The long-term goal is to transition VisionPal into a portable mobile application, expanding its real-world applicability and convenience.

The main contributions of this paper are: (1) A voice-controlled desktop assistive tool for visually impaired users integrating ASR, TTS, and a vision-language model; (2) A methodology for evaluating image captioning under varying environmental conditions through a user study assessing the accessibility, reliability, and practical value of the tool; (3) The possibility of using the application in Arabic language.

Materials and methods

VisionPal is a voice-controlled desktop application designed to assist visually impaired users by providing audio descriptions of images. The system integrates computer vision, automatic speech recognition (ASR), text-to-speech (TTS), and natural language understanding into a unified interface. VisionPal operates in English and Arabic, offering a seamless multimodal experience without the need for complex or expensive external hardware.

The application runs on a standard desktop or laptop with internet access, leveraging cloud-based services to minimize local computational requirements.

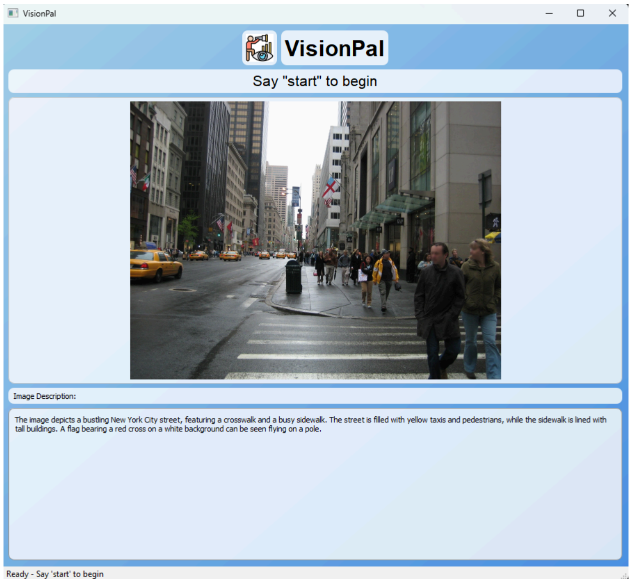
I. *Software Components*

The system is implemented using the Python programming language, with the following core components:

- GUI Framework: [PyQt5] is used to build the graphical interface, although interaction is entirely voice-based in the primary use case.
- Vision-Language Model: LLaVA (Large Language and Vision Assistant) is used for image captioning and visual question answering. VisionPal uses the (meta-llama/Llama-3.2-11B-Vision-Instruct-Turbo) model hosted on Together AI, a cloud platform offering hosted inference endpoints for advanced multimodal models. This is a state-of-the-art multimodal vision-language model developed by Meta, capable of generating detailed image captions and answering complex questions about visual content. The model

supports both general scene understanding and fine-grained reasoning, making it suitable for assistive technologies for visually impaired users. The model is accessed via Together AI's cloud API, which eliminates the need for local GPU-based hardware and ensures lightweight deployment. Only an internet connection is required, making the system highly portable and cost-effective.

- Speech Recognition: Google Speech Recognition API is used for transcribing spoken input into text. It supports high accuracy in both English and Arabic.
- Text-to-Speech: The gTTS (Google Text-to-Speech) library is used to convert textual responses into spoken audio in the user's selected language.
- PyGame for seamless audio playback.
- Voice Commands Handling: The application listens for key voice prompts such as "start," "camera," or "gallery" to guide the user through the process.
- Threading ensures the UI remains responsive during: Speech recognition, AI model inference and Audio playback.



Say "start" to begin

Image Description:

The image depicts a bustling New York City street, featuring a crosswalk and a busy sidewalk. The street is filled with yellow taxis and pedestrians, while the sidewalk is lined with tall buildings. A flag bearing a red cross on a white background can be seen flying on a pole.

Ready - Say 'start' to begin

FIGURE 1. VISIONPAL DESKTOP INTERFACE SHOWING IMAGE CAPTURE AND SCENE DESCRIPTION OUTPUT.

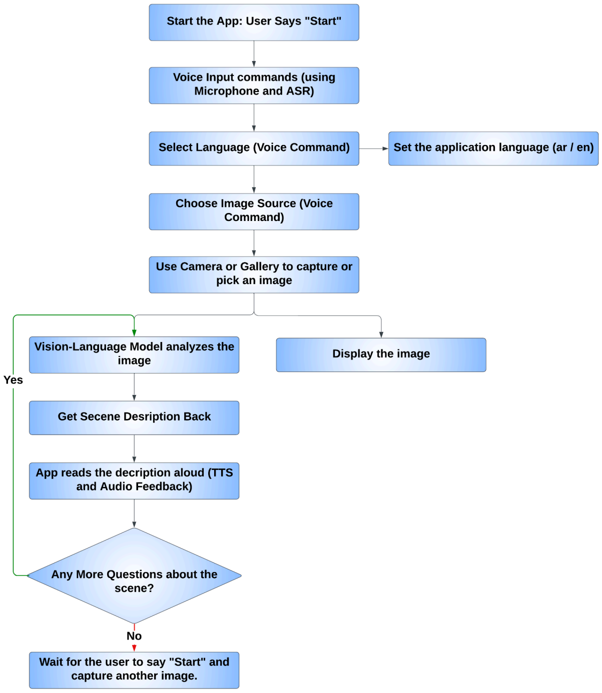

Start the App: User Says "Start"
Voice Input commands (using Microphone and ASR)
Select Language (Voice Command)
Set the application language (ar / en)
Choose Image Source (Voice Command)
Use Camera or Gallery to capture or pick an image
Vision-Language Model analyzes the image
Display the image
Yes
Get Secene Desription Back
App reads the decription aloud (TTS and Audio Feedback)
Any More Questions about the scene?
No
Wait for the user to say "Start" and capture another image.

FIGURE 2. VISIONPAL SYSTEM ARCHITECTURE SHOWING THE FLOW FROM USER VOICE INPUT TO AUDIO FEEDBACK USING INTEGRATED ASR, VISION-LANGUAGE MODEL, AND TTS MODULES.

II. Hardware and Deployment Environment

VisionPal is designed to operate on low-to-mid-range desktop systems without GPU acceleration. The system has been tested on:

- Operating System: Windows 10
- CPU: Intel i5 or higher
- RAM: 8GB
- Camera: Built-in or external webcam

An internet connection is required to access cloud-based services such as Together AI and Google APIs

III. Evaluation Methodology

To assess the usability and effectiveness of VisionPal, we conducted a multi-stage evaluation focusing on:

1. Environmental Conditions: The system was tested under varying lighting conditions (natural light, artificial light, and low light).
2. Scene Complexity: Performance was evaluated on a range of images with different object densities and backgrounds.
3. User Study: Feedback was collected from sighted users. Participants completed structured questionnaires assessing:

- Accuracy of scene descriptions
- Overall user experience and perceived usefulness

To evaluate VisionPal's usability and real-world performance, **a structured feedback form was created using Google Forms.** Participants were asked to rate the application's performance across multiple dimensions.

Responses were collected from a diverse group of participants, The feedback provided valuable insights for refining the system and identifying areas for future improvement.

Here is the link to the online form: https://forms.gle/Wt2QakNjSF3Xn1ia7

RESULTS AND DISCUSSION

I. System Performance

The VisionPal application was tested in various environments to evaluate its usability, accuracy, and responsiveness. The core components—Automatic Speech Recognition (ASR), vision-language understanding, and Text-to-Speech (TTS)—worked together to create an interactive voice-controlled experience. The model used for image captioning and question answering was meta-llama/Llama-3.2-11B-Vision-Instruct-Turbo, accessed via the Together AI API. This model consistently generated relevant and coherent descriptions for a wide range of indoor and outdoor scenes.

Experiments were conducted under different lighting conditions (bright daylight, dim lighting, and artificial indoor lighting). In well-lit scenes, the system produced highly detailed and accurate descriptions. In lower-light scenarios, performance declined slightly due to the reduced image quality, but key elements of the scene were still captured. The captioning output was generally robust, though clarity was reduced when images were blurry or overexposed. These findings suggest that while the model is tolerant of minor visual noise, future enhancements like image preprocessing (e.g., auto-enhancement or blur detection) could improve overall reliability.

One of the key strengths of VisionPal is its ability to provide descriptive, natural-language feedback to users in response to both images and voice commands. Unlike traditional object detection systems that return isolated labels or bounding boxes, VisionPal delivers full-sentence captions and can answer follow-up questions, making it more aligned with how visually impaired users process and understand their environment.

Moreover, the application supports both English and Arabic, broadening its accessibility to users in different linguistic regions. The voice-controlled interface further reduces the need for manual input, aligning the user experience with the needs of those who cannot rely on visual UIs. This also makes VisionPal potentially more usable in real-world conditions where quick hands-free operation is critical.

The use of free and publicly accessible APIs also allows VisionPal to remain affordable and deployable at scale, especially in low-resource settings. Nevertheless, long-term stability and data privacy of cloud services must be taken into account in future iterations. A potential improvement for the next version of VisionPal is integrating on-device speech recognition and vision-language models to allow offline functionality, reduce latency, and enhance privacy.

II. User Feedback and Evaluation

A structured evaluation was carried out using a Google Forms questionnaire distributed to users. The feedback was gathered from 6 participants. The questions covered the app's ease of use, usefulness of the image descriptions, and potential areas for improvement.

Results Summary from User Feedback:

I. *Accuracy and Usefulness:*

- Most users (66.68%) rated the image descriptions as very accurate or somewhat accurate.
- The majority (55.55%) agreed that the descriptions are helpful or sufficient to help a blind person navigate safely .
- A few users felt the descriptions could be improved for better accuracy or more detail.

II. *Strongest Aspects:*

- Users highlighted object detection as the strongest feature.
- Other praised aspects included scene layout, spatial relationships, and clear indication of the direction of the path.
- Some noted the inclusion of obstacles, signs, vehicles, and presence of people as valuable information.

III. *Areas for Improvement:*

- Users suggested adding more details on distance to objects and spatial relationships.
- Some feedback mentioned missing specific environmental details (e.g., weather, lighting conditions).
- A few comments pointed out the need for warnings about hazards like holes, stairs under construction, or slippery surfaces.
- Better description of signage content and its precise location was requested.

IV. *Key Information to Include:*

- Most users agreed on the importance of describing direction of the path, obstacles, presence of people, vehicles, signs or symbols, and stairs or ramps.
- Other important points included open path availability, nearby objects, weather conditions, and traffic signals.

V. *User Comments:*

- Feedback was generally positive, with many users finding VisionPal useful and promising.
- Suggestions included focusing on neglecting irrelevant details (e.g., night or irrelevant info) and adding more real-time reaction capabilities.
- Some users emphasized the importance of interactive feedback to clarify descriptions or provide warnings.

VI. *Overall Usefulness:*

- Most respondents believe that the VisionPal application is truly useful for visually impaired or blind users (83.3%).
- A majority expressed willingness to use the app if they were part of the target user group (66.7%).

This feedback highlights the strength of using cloud-based models like Together AI for high-quality responses, while also pointing to the importance of reducing reliance on stable internet connections. Users expressed interest in offline support, which can be achieved in the future by integrating smaller local models for both ASR and vision-language processing.

III. *Multilingual Capability*

The application supports both English and Arabic as many vision-based tools tend to focus on English only. While Google's speech recognition API handled Arabic input well, there were occasional misinterpretations with regional dialects. Enhancing this feature through dialect-specific tuning or future local ASR models could improve performance further.

IV. *Accessibility and Hardware Considerations*

Unlike traditional assistive tools requiring external sensors or wearable devices, VisionPal runs entirely on a standard desktop environment using only a webcam and microphone. By relying on free online models and APIs, VisionPal eliminates the cost and complexity of specialized hardware. This increases accessibility for users in low-resource regions or those who cannot afford proprietary assistive technologies.

The system's architecture is designed to be easily transferable to mobile platforms, which will enhance portability and usability. Mobile integration is planned for future iterations, along with the use of lightweight, quantized models to enable offline functionality.

Conclusion

This paper presented VisionPal, a voice-controlled assistive application designed to enhance the independence and navigation capabilities of visually impaired users. By integrating advanced vision-language models with automatic speech recognition and text-to-speech technologies, VisionPal provides rich, real-time audio descriptions and interactive question-answering about the user's environment. The system's reliance on cloud-based AI models eliminates the need for bulky, expensive hardware, making it more accessible and scalable.

User feedback collected via Google Forms indicates promising usability and practical benefits, although further refinement and testing across diverse scenarios are needed. Future work will focus on optimizing the application for mobile platforms, improving offline capabilities by incorporating local speech recognition models, and expanding language support. VisionPal's design prioritizes user convenience and interactivity, aiming to fill existing gaps in assistive technologies for people with visual impairments.

Ultimately, VisionPal demonstrates the potential of combining state-of-the-art AI with accessible interface design to empower visually impaired individuals in their daily lives.

References

[1] World Health Organization. (2021). World report on vision. [Online]. Available: https://www.who.int/publications/i/item/9789241516570

[2] Hersh, M. A., & Johnson, M. A. (2010). Assistive technology for visually impaired and blind people. Springer Science & Business Media.

[3] Aspinall, P. (2012). Guide Dogs and Mobility Training: Costs and Limitations. International Journal of Visual Impairment.

[4] Dos Santos, D. G., Ferrari, G. L., Medola, F. O., & Sandnes, F. E. (2022). "A Review of Wearable and Portable Technologies for the Visually Impaired," Sensors, vol. 22, no. 1, pp. 1–21.

[5] Mukhiddinov, M., & Cho, Y. (2021). "A Scene Recognition System Using Deep Neural Networks for the Visually Impaired," Applied Sciences, vol. 11, no. 2.

[6] Kuriakose, S., Shrestha, R., & Sandnes, F. E. (2020). "Design of Smart Navigation Tools for the Visually Impaired: A User-Centered Perspective," Disability and Rehabilitation: Assistive Technology, vol. 15, no. 6, pp. 635–642.

[7] Bhowmick, Alexy, and Shyamanta M. Hazarika. "An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends." Journal on Multimodal User Interfaces 11 (2017): 149-172.

[8] Chanana, Piyush, et al. "Assistive technology solutions for aiding travel of pedestrians with visual impairment." Journal of rehabilitation and assistive technologies engineering 4 (2017): 2055668317725993.

[9] Real, Santiago, and Alvaro Araujo. "Navigation systems for the blind and visually impaired: Past work, challenges, and open problems." Sensors 19.15 (2019): 3404.

[10] Real, Santiago, and Alvaro Araujo. "Navigation systems for the blind and visually impaired: Past work, challenges, and open problems." Sensors 19.15 (2019): 3404.

[11] Dos Santos, Aline Darc Piculo, et al. "Aesthetics and the perceived stigma of assistive technology for visual impairment." Disability and Rehabilitation: Assistive Technology 17.2 (2022): 152-158.

[12] Kuriakose, Bineeth, Raju Shrestha, and Frode Eika Sandnes. "Smartphone navigation support for blind and visually impaired people-a comprehensive analysis of potentials and opportunities." International conference on human-computer interaction. Cham: Springer International Publishing, 2020.