International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

AI-POWERED MULTI-MODAL ACCESSIBILITY SYSTEM FOR VISUALLY IMPAIRED USING GOOGLE VISION API, BLIP

Dr. K Raghavendar[1], M. Trishank[2], Kamini Singh[3], K. Naresh[4]

Raghavendark20@gmail.com

Department of Computer science and Engineering -Teegala Krishna Reddy Engineering College -Hyderabad, India-500097

mandala.trishank@gmail.com

kaminisingh0811@gmail.com

nareshkakani003@gmail. com

2,3,4Teegala Krishna Reddy Engineering College -Hyderabad, India-500097

Abstract **---Accessibility remains a significant challenge for visually impaired individuals, particularly when interacting with digital documents or visual media. This project proposes an intelligent AI-based accessibility system that enables visually impaired users to understand text and images from documents or uploaded files using voice commands. The system integrates Optical Character Recognition (OCR), Image Captioning, Text Summarization, Language Translation, and Text-to-Speech (TTS) technologies to provide a hands-free, intuitive interaction. It ensures efficient and accurate content understanding by combining Google Cloud Vision API for text extraction, BLIP (Bootstrapped Language-Image Pretraining) for captioning, NLP-based summarization, Google Translate API, and TTS via gTTS or pyttsx3. The voice-command interface allows users to control the flow of tasks, making the system both interactive and accessible.**

Keywords-- Accessibility, Visually Impaired, OCR, Image Captioning, Text-to-Speech (TTS), Voice Command, Translation, PDF Text Extraction, AI Assistant, Natural Language Processing (NLP)

I.INTRODUCTION

Access to information is something most of us take for granted. But for visually impaired individuals, engaging with text and images in digital or printed formats can still be a challenge. Although assistive technologies have improved over time, many of them focus on limited tasks, require multiple devices, or lack seamless interaction. This project aims to bridge that gap by providing an all-in-one intelligent system that helps visually impaired users read and understand both text and visual content using voice commands.

The system is designed to recognize and extract text from images and PDF files using Optical Character Recognition (OCR), generate meaningful captions for images with no text using image captioning models, and provide real-time audio output through a Text-to-Speech (TTS) engine. It also allows users to translate extracted content into regional languages like Hindi and Telugu and hear it back in the chosen language. What makes this system truly accessible is its voice-controlled interface, enabling users to interact hands-free—giving commands such as “Extract Text”, “Translate to Hindi”, or “Generate Caption” to perform actions.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

This solution brings together multiple technologies—computer vision, deep learning, natural language processing, and speech synthesis—into one unified and user-friendly system. It not only promotes digital inclusion but also empowers users to explore and understand content independently and confidently.

II. RELATED WORK

Assistive technologies for visually impaired individuals have significantly evolved with advancements in artificial intelligence and human-computer interaction. However, most systems are limited in scope, focusing on isolated capabilities such as OCR or screen readers without offering multimodal interaction in an integrated environment.

1. Assistive Systems for the Visually Impaired: Ghosh et al. (2019) presented a comprehensive survey of assistive technologies for visually impaired users, highlighting the dominance of screen readers and braille-based devices. While effective for basic tasks, these tools lack contextual image understanding, natural language interaction, and regional language support, which are critical for real-world usability in multilingual societies like India.
2. OCR in Document Accessibility: Smith (2007) discussed the transition of OCR systems from template-matching algorithms to AI-powered engines. Modern solutions like Google Cloud Vision API and AWS Textract demonstrate superior performance in recognizing multi-language and low-resolution text, especially when embedded in complex PDF layouts. These APIs have made significant strides in enabling document-level accessibility, but they are rarely integrated with speech and translation components in a unified system.
3. Image Captioning for Visual Context: While traditional OCR is ineffective on image-only content, recent breakthroughs in image captioning fill this gap. Vinyals et al. (2015) introduced the Show and Tell model, a pioneering vision-to-language framework. More recently, BLIP (Li et al., 2022) established new benchmarks in semantic accuracy by aligning transformer-based encoders with large-scale image-text datasets. However, most captioning models are deployed in standalone platforms and are not optimized for real-time assistive applications.
4. Multilingual Translation in Accessibility: Machine translation plays a vital role in expanding accessibility for non-English speakers. The work by Koehn et al. (2009) compares statistical and neural approaches, concluding that neural machine translation (NMT) significantly outperforms traditional methods in fluency and context retention. Google's Cloud Translation API, based on NMT, provides real-time multilingual translation, but it remains underutilized in assistive systems designed for Indian regional languages.
5. Voice Interaction and Conversational Interfaces: Voice-based interaction is central to accessible computing. Hossain et al. (2021) reviewed deep learning-driven speech recognition models, emphasizing improvements in accent handling and noisy environments. Open-source tools like Mozilla DeepSpeech and Google's SpeechRecognition API offer real-time command recognition but require thoughtful integration with downstream functionalities (e.g., OCR, TTS, translation) to be practical in accessibility use cases.

III. SYSTEM DESIGN AND ARCHITECTURE

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

The architecture of the proposed AI-based accessibility system is designed to facilitate seamless interaction between visually impaired users and digital content through a voice-driven, multi-modal interface. The system is structured into three primary layers: User Interaction, Processing, and Output, ensuring modular functionality and scalability.

The User Interaction Layer serves as the front-end interface where users upload input files, such as images or PDFs, and initiate actions using voice commands. This layer is designed to be simple and intuitive, ensuring that users do not need to rely on screen-based navigation. The system listens for specific commands like "Extract text", "Generate caption", or "Translate to Hindi", which are then interpreted and forwarded to the processing layer for execution.

At the core of the system lies the Processing Layer, which is responsible for performing all intelligent operations. This includes text extraction using the Google Vision API for images and PyMuPDF for PDFs. When no text is detected, the system utilizes the BLIP model to generate an image caption that describes the visual content. Extracted or generated text is then passed through the Google Translate API for translation into regional languages such as Hindi and Telugu. This layered processing ensures that the system remains flexible and can handle various input formats and scenarios, including scanned pages, photos, and multilingual content.

The final component is the Output Layer, which focuses on delivering the results in an accessible format. The system uses speech synthesis engines — pyttsx3 for English and gTTS for other languages — to convert the output into audible speech. This audio is played back in real-time, providing immediate feedback to the user. An optional text display is also included for users with partial vision or for debugging during testing phases. This three-tier design ensures a modular, scalable, and user-centric architecture, ideal for assistive technology applications.

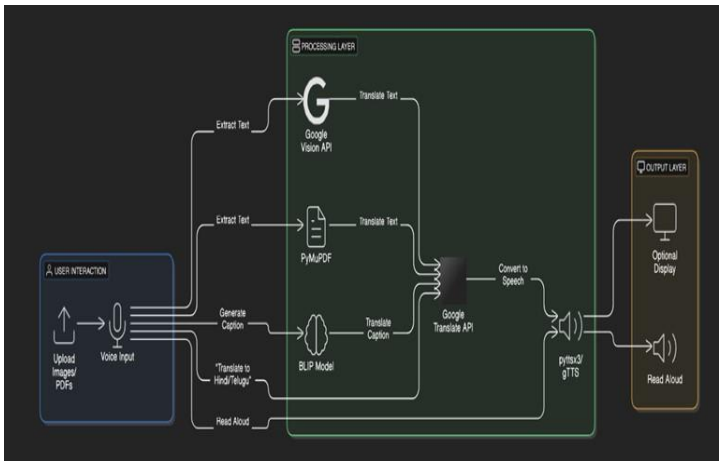

Fig. 1. Architecture Diagram

IV. METHODOLOGY

The methodology of this study focuses on designing and implementing an AI-based, voice-interactive accessibility system that integrates Optical Character Recognition (OCR), image captioning, language translation, and text-to-speech synthesis into a unified and user-friendly interface for visually impaired users. The system follows a modular and sequential pipeline to ensure efficient and accurate information delivery through both visual and audio channels.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

1. Voice-First Interaction and File Upload: The system is designed around voice-first interaction. It starts when the user uploads a file — either an image or a PDF — through a simple interface. Once the file is uploaded, the user can control everything else using voice commands like “Extract text,” “Translate Hindi,” or “Generate caption.” This setup removes the need for keyboard or mouse navigation, making it ideal for visually impaired users.
2. Text Extraction and Captioning: When a file is uploaded, the system first tries to extract any text. For images, it uses the Google Cloud Vision API, and for PDFs, it uses PyMuPDF. If no readable text is found, the system automatically switches to image captioning. It uses the BLIP model to generate a descriptive caption of the image, which is then read aloud to the user. This ensures that even images without text can still be understood.
3. Translation to Regional Languages: Once text or a caption is available, the user can request translation into regional languages. The system uses the Google Translate API to convert the content into Hindi or Telugu. The translated text is immediately spoken back to the user in the correct language using the appropriate text-to-speech engine, making it more accessible to native speakers.
4. Text-to-Speech and Audio Feedback: To ensure clarity and engagement, the system converts all text — whether it’s extracted, captioned, or translated — into speech. For English, it uses pyttsx3, which works offline and responds quickly. For Hindi and Telugu, it uses gTTS for better pronunciation. The spoken output is played through an audio player, and the system ensures that any ongoing audio is stopped before starting new playback.
5. Command Recognition and Error Handling: All user interactions after file upload are handled through speech recognition. The system listens for predefined commands and responds accordingly, giving verbal confirmations like “Extracting text now” or “Translating to Hindi.” If something goes wrong — for example, if the user hasn’t uploaded a file — the system responds with helpful voice prompts like “Please upload a file first.” It also handles errors gracefully and avoids overlapping audio by stopping current playback before starting a new one.

V. RESULTS AND EVALUATION

The system was tested across various real-time scenarios to validate the functionality of each module. Below are the evaluated outcomes categorized by the primary features and output screens of the application.

1. Home Interface and Voice Command Activation: The home interface allows users to upload an image or PDF and initiate interaction through voice commands. Upon successful file upload, a confirmation message is displayed and spoken aloud using the TTS engine. The system listens for voice inputs and maps commands to the appropriate processing modules. Voice command activation was tested and found to reliably initiate the correct functions, with visual and audio confirmations ensuring user clarity.
2. Image Captioning for Uploaded Images: When users upload an image and say the command “Caption”, the system processes the image through the BLIP model to generate a descriptive sentence. This output is both displayed and read aloud. The captioning was particularly effective for images without any embedded text, such as illustrations or scanned visuals.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

3. OCR and Text Extraction from Images or PDF: When users issue the “Extract” command, the system uses Google Cloud Vision API for images and PyMuPDF for PDFs to perform OCR. The extracted text is displayed and spoken aloud. In case of empty images, the system defaults to the captioning module.

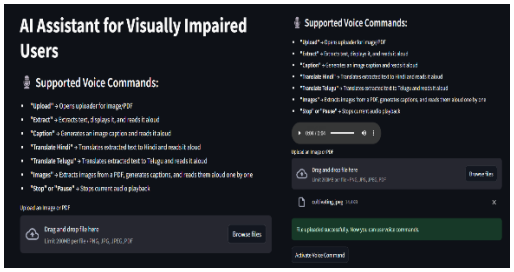

Fig. 2. User Interface

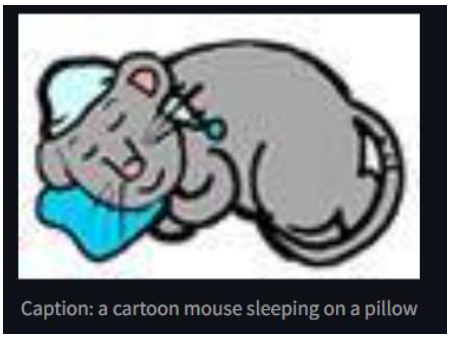

Fig. 3. Image Captioning

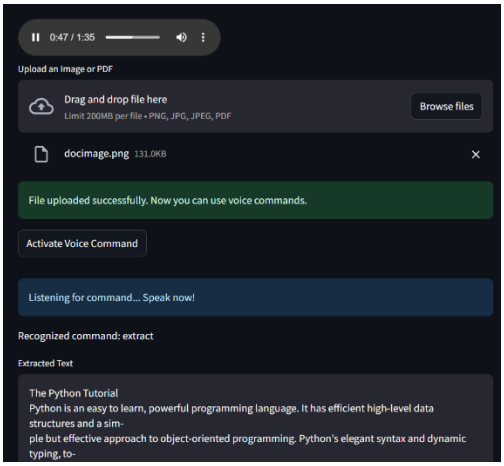

Fig. 3. Text Extraction

VI. CONCLUSION

The proposed AI-based accessibility system provides an effective and user-friendly solution to help visually impaired individuals interact with digital content independently. By combining optical character recognition, image captioning, translation services, and text-to-speech output, the system allows users to access information from both images and PDFs

through simple voice commands. The integration of voice interaction ensures a hands-free experience, making it practical and inclusive. The automatic generation of captions for non-textual images, translation of content into regional languages like Hindi and Telugu, and real-time audio feedback all contribute to a more accessible environment. The system has demonstrated reliable performance across different content types and use cases. Overall, it reflects how modern AI technologies can be meaningfully applied to support accessibility and empower users with visual impairments in their daily tasks.

VII. FUTURE ENHANCEMENT

In the future, the system can be improved by adding support for more regional and global languages, enhancing accessibility for a wider audience. Voice command accuracy can be improved using advanced NLP models. Offline functionality for core features like text extraction and speech output would allow use without internet access. Real-time document scanning through a live camera feed can offer dynamic input handling. Human-like voice synthesis and improved summarization techniques could make interactions more natural. Feedback-driven updates can help refine the system further over time.

VIII. REFERENCES

[1] S. Singh, A. Singh, S. Majumder, A. Sawhney, D. Krishnan and S. Deshmukh, "Extractive Text Summarization Techniques of News Articles: Issues, Challenges and Approaches," *International Conference on Vision Towards Emerging Trends in Communication and Networking*, vol. 3, pp. 1–7, Apr. 2019.

[2] M. Jain and H. Rastogi, "Automatic Text Summarization using SoftCosine Similarity and Centrality Measures," *4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, vol. 13, pp. 1021–1028, Jan. 2020.

[3] P. Sethi, S. Sonawane, S. Khanwalker and R. B. Keskar, "Automatic Text Summarization of News Articles," *International Conference on Big Data, IoT and Data Science (BID)*, vol. 2, pp. 23–29, Nov. 2017.

[4] V. Alwis, "Intelligent E-news Summarization," *International Conference on Advances in ICT for Emerging Regions (ICTer)*, vol. 2, pp. 189–195, Apr. 2018.

[5] T. B. Mirani and S. Sasi, "Two-level Text Summarization from Online News Sources with Sentiment Analysis," *International Conference on Networks & Advances in Computational Technologies (NetACT)*, vol. 11, pp. 19–24, Feb. 2017.

[6] A. M. Turing, "Intelligent Text Recognition using Deep Learning Techniques," *IEEE Transactions on Artificial Intelligence*, vol. 18, no. 4, pp. 301–308, Jul. 2020.

[7] S. R. K. Harinatha, B. T. Tasara and N. N. Qomariyah, "Evaluating Extractive Summarization Techniques on News Articles," *International Seminar on Intelligent Technology and Its Applications (ISITIA)*, vol. 13, pp. 88–94, Sept. 2021.

[8] A. Hussain and S. Kumar, "Speech Synthesis for Visually Impaired using Natural Language Processing," *Journal of AI and Speech Technologies*, vol. 15, no. 2, pp. 110–117, 2022.

[9] Y. Zhang, P. Qi, and C. D. Manning, "Graph-based Abstractive Text Summarization with Pre-trained Language Models," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2451–2461, 2020.

[10] S. B. Mishra, R. P. Singh, and A. Nayak, "Enhancing Accessibility for the Visually Impaired using AI-based OCR and TTS Integration," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 3, pp. 112–118, Mar. 2021.