International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

DocuVoice: Survey of Intelligent Document Analysis Solutions for the Visually Impaired

Abhishek Chavan
Department of Computer Engineering,
Pune Institute of Computer Technology,
Pune, India.
abhishekchavan9394@gmail.com

Roshan Patil
Department of Computer Engineering,
Pune Institute of Computer Technology,
Pune, India.
roshanvpatil2004@gmail.com

Sahil Patil
Department of Computer Engineering,
Pune Institute of Computer Technology,
Pune, India.
ssp221004@gmail.com

Pankaj Ahirrao
Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India.
pdahirrao25@gmail.com

Prof. P. T. Kohok
Assistant Professor, Department of Computer Engineering, Pune Institute of Computer Technology,
Pune, India.
ptkohok@pict.edu

***Abstract*—The availability of written information is a challenge for visually impaired people. The traditional screen reader and text-to-speech technology enable linear reading but do not support understanding and interaction. This paper presents a survey of the current state of the art and research in document analysis and accessibility, including Optical Character Recognition, layout analysis, Natural Language Processing, summarization, and speech interfaces. The survey highlights the key gaps in the current state of the art, including dealing with complex layouts, explaining figures and tables, multilingual support, and data privacy. This paper also introduces a unified framework for an intelligent document analyzer using AI technology called Docu- Voice. Index Terms—Accessibility, visually impaired, document analysis, OCR, NLP, text-to-speech, assistive technology, artificial intelligence. *DocuVoice*.**

***Index Terms*—Accessibility, visually impaired, document analysis, OCR, NLP, text-to-speech, assistive technology, artificial intelligence.**

I. INTRODUCTION

Access to printed and digital documents is fundamental to education, employment, and independent living. Despite advances in assistive technology, visually impaired users still face significant challenges when accessing arbitrary docu- ments, especially scanned pages, images, or poorly tagged PDFs [1], [2]. Screen readers such as JAWS, NVDA, and VoiceOver enable linear reading but provide limited compre- hension, navigation, and interaction capabilities. Users often need high-level overviews, summaries, or the ability to ask targeted questions about a document rather than listening to entire files linearly [5].

The purpose of this survey is to discuss the technologies that can be integrated to build a smarter and more useful system for the visually impaired. We term this new concept *DocuVoice* — an AI-powered, voice-centric document analyzer. DocuVoice will attempt to (1) identify and organize information from various document types, (2) offer contextual summaries and question-answering capabilities, (3) transform non-text data into useful descriptions, and (4) offer a straightforward voice interface for interaction and control.

The structure of this paper is as follows: Section II intro- duces previous work and systems; Section III introduces core methods and an integrated framework; Section IV identifies key challenges and trade-offs; Section V suggests future research avenues and DocuVoice functionality; VI concludes.

II. LITERATURE REVIEW AND EXISTING WORK

A. Screen Readers and Simple TTS

Screen readers are well-established software tools intended for the playback of structured electronic text as audio. They work well for well-tagged documents (e.g., HTML or accessi- ble PDF), but they are often ineffective for scanned or poorly formatted documents [1]. TTS systems translate text to speech but rarely support structure-aware summary, navigation, or Q&A systems [6].

B. OCR and Text Extraction

Optical Character Recognition (OCR) is the process of recognizing printed or handwritten text and converting it into machine-readable form. Open-source OCR engines such as Tesseract remain widely used [7], although commercial solutions (e.g., Google Cloud Vision, Microsoft OCR) are more robust. Contemporary systems employ deep CNN and sequence models to deal with different fonts and noisy inputs [12], [13]. Nonetheless, raw text extraction is only the first step, and document layout and semantics are vital for accessibility purposes [8].

C. Document Layout Analysis

Layout analysis involves the detection of regions such as title, headings, paragraphs, lists, tables, and images. Recent work includes transformer models (LayoutLM series) and visually-grounded models such as Donut and DocFormer [9], [10]. PubLayNet and DocBank datasets fuel advancements [8]. Proper layout analysis facilitates logical navigation (e.g., jump to next heading) and proper reading order in multi-column documents.

D. Natural Language Processing (NLP)

NLP models support summarization, entity recognition, and question answering (Q&A). Extractive approaches pick sentences, while abstractive approaches (BART, T5) synthesize summaries [16], [19]. Retrieval-Augmented Generation (RAG) models integrate retrieval for grounding and generation for fluent response writing. This is an attractive paradigm for document-based Q&A [17]

E. Image Captioning and Multimodal Models

Non-text components (graphs, images, diagrams) demand multimodal models. CLIP and BLIP models facilitate captioning and vision-language alignment [11]. For graphs and tables, dedicated parsers can translate structure to text summaries [14]. Math equation OCR software (Mathpix) assists in translating equations into speech-supportive representations [15].

F. User Studies and Accessibility Research

User studies highlight the importance of context, fast navigation, and interactivity for visually impaired individuals, rather than mere reading [1], [5]. User surveys show the demand for summarization, voice interaction, and accurate figure description. Nevertheless, user-centric assessments with specific target groups are relatively rare in technical literature [2].

III. Methodologies Proposed Architecture

A. High-Level Architecture

Figure 1 gives an architectural view of DocuVoice. The pipeline comprises:

1) **Input Layer:** Accepts PDFs, images, DOCX, or camera captures.
2) **Preprocessing:** Image denoising, deskewing, contrast enhancement.
3) **OCR Layout Detection:** Extracts text with bounding boxes; detects regions (headings, tables, figures).
4) **Semantic Analysis:** Performs entity recognition, topical segmentation, and keyphrase extraction.
5) **Summarization QA:** Produces section summaries and answers user queries using RAG or fine-tuned trans- former models.
6) **Accessibility Interface:** TTS rendering, voice command parser, navigation primitives.
7) **Storage/Privacy Layer:** Local encrypted caching; op-

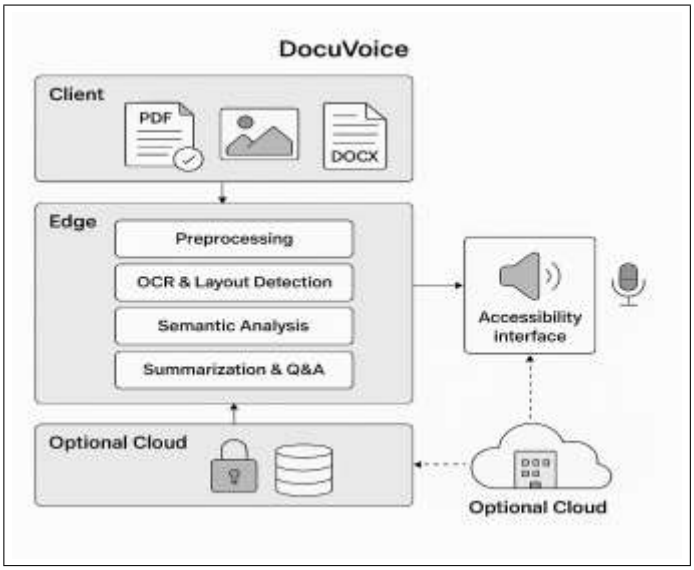

Fig. 1: DocuVoice system architecture

B. Text Extraction and Layout Analysis

The OCR system should provide both the recognized text and the positional information (coordinates) [7], [12]. The coordinates are used by the layout model to infer the reading order and semantic labels [8], [9]. The text can be extracted from PDFs with layout coordinates using PyMuPDF, and together with PubLayNet-trained layout detectors, the system is layout-independent [8].

C. Semantic Parsing and Topic Segmentation

After text extraction, sentence segmentation and tokeniza- tion follow. Named-entity recognition (NER) identifies people, dates, and other entities; topic segmentation identifies topics by grouping sentences into meaningful chunks. For large documents, these chunks facilitate focused summary extraction [16].

D. Summarization Strategies

DocuVoice adopts a hybrid summarization strategy:

- **Extractive:** Embeddings (SBERT) and ranking (TextRank) are used to select informative sentences [18].
- **Abstractive:** BART/T5 models are fine-tuned to generate condensed paraphrases suitable for audio presentation [19], [20].

Extractive summarization is useful for content selection, while abstractive summarization enhances the summary's brevity and naturalness for audio presentation.

E. Retrieval-Based Q&A (RAG)

For Q&A, candidate passages are ranked based on a user query, and the top-scoring passages are used by a generator

to construct an answer. This helps to ensure that the answer is document-grounded and prevents hallucinations [17].

F. Non-Text Modalities: Tables and Figures

Tables: The system includes table detection followed by structural analysis (rows, columns, and headers). Table-to-text pipelines are used to summarize the table by describing header information and important rows (e.g., maximum/minimum values or key totals) [14]. **Images/Figures:** Image captioning models (e.g., BLIP) provide short descriptions; chart under- standing may use specialized pipelines to extract axes and data points [11].

G. Voice Interface Design

A small set of commands can decrease the cognitive burden:

- *Summarize document"*, *Summarize section"*, *Explain figure X"*, *Jump to next heading"*, *"Search for keyword"*.

Speech recognition should be able to tolerate background noise; offline speech recognition models (such as Whisper or VOSK) could be used to support privacy-friendly voice interaction [23].

IV. Challenges and Gap Analysis

A. Layout Complexity and Reading Order

Complex layouts, such as multi-column texts, sidebars, floating figures, and diverse typographic practices,

B. Handling OCR Errors

OCR errors (incorrect recognition, merged/split words) affect the performance of subsequent tasks. Confidence measures, language model-based post-processing, and human cor- rection (if needed) can improve robustness [7], [12].

C. Describing Non-Textual Data

Representing images, charts, and mathematical expressions accurately and concisely is difficult. State-of-the-art image captioning models can generate generic captions, and understanding charts often needs domain knowledge [11], [15].

D. Latency, On-device vs Cloud

Real-time interaction is desirable for on-device inference, but on-device complex models can consume all available resources. Edge-cloud hybrids must strike a balance [13].

E. Multilingual and Low-resource Languages

There are many places where OCR/TTS/NLP for non-Latin scripts (e.g., Devanagari for Hindi/Marathi) is required. Developing good models for such languages is a challenge, and IndicBERT and multilingual mT5 are good starting points [20], [22].

F. Evaluation and User Studies

Objective measures like ROUGE and BLEU are not very representative of the quality of accessibility. Human-centric evaluation, including comprehension tests, success rates, and user satisfaction, involving visually impaired users is required [1], [5].

V. Future Scope DocuVoice Contributions

A. Adaptive and Personalized Interfaces

DocuVoice should be able to adapt to user preferences regarding verbosity, voice style, and summary level. A simple user profile could allow for personalized approaches to summarization (e.g., more technical information for students, summary-level control for casual reading).

B. Multilingual Pipeline

We suggest adding Indic language OCR/TTS and multilingual NLP (IndicBERT, mT5) to the pipeline to provide local language content [22].

C. Privacy-first Processing

Local processing and encrypted caching should be the norm. Cloud computing should be opt-in for resource-intensive processing, and anonymized data should be used wherever possible.

D. Assistive Integrations

DocuVoice can integrate with:

- Mobile camera capture apps for instant scanning.
- Braille displays via BRF conversion.
- Learning management systems (LMS) to provide accessible course materials.

E. Community and Feedback Loop

Allow users to correct transcriptions and captions; use these corrections to incrementally fine-tune models (federated or centralized) while preserving privacy.

VI. Conclusion

This study investigated the components and difficulties involved in developing an intelligent document assistant for the visually impaired. Although each technology (OCR, layout analysis, NLP, TTS) is well-established, their integration for accessibility purposes is still in its infancy. DocuVoice envisions an integrated, voice-centric, and privacy-friendly system that supports active understanding via summarization and QA. Future research should concentrate on the following: multilinguality, robust multimodal understanding, efficiency on-device, and, most importantly, human-centric studies involving the visually impaired.

REFERENCES

[1]M. Dorigo, B. Harriehausen-M"uhlbauer, I. Stengel, and P. S. Dowland, "Survey: Improving Document Accessibility from the Blind and Visu- ally Impaired User's Point of View," in *Universal Access in Human- Computer Interaction. Applications and Services*, LNCS, vol. 6768, 2011.
[2]L. Kaczmirek and K. G. Wolff, "Survey Design for Visually Impaired and Blind People," *Universal Access in HCI*, 2007.
[3]M. Bamdad, D. Scaramuzza, and A. Darvishy, "SLAM for Visually Impaired People: a Survey," arXiv:2212.04745, Dec. 2022.
[4]P. Kathiria, "Assistive systems for visually impaired people: A survey,"
ScienceDirect, 2024.
[5]I. Xie *et al.*, "How to conduct blind and visually impaired (BVI) user studies in mobile environment," University of Wisconsin–Milwaukee, 2023.
[6]ACM Digital Library, "A Survey on Assistive Technologies for Visually Impaired Users," ACM, 2025.
[7]R. Smith, "An Overview of the Tesseract OCR Engine," in *Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR)*, 2007.
[8]H. Xu *et al.*, "PubLayNet: Dataset for Document Layout Analysis,"
Proc. CVPR, 2019.
[9]Z. Hu *et al.*, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in *Proc. KDD*, 2020.
[10] Y. Xu *et al.*, "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," *ArXiv*, 2022.
[11] J. Li *et al.*, "BLIP: Bootstrapped Language-Image Pretraining for Unified Vision-Language Understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.
[12] J. Jaided, "EasyOCR: Ready-to-Use OCR with Deep Learning," GitHub Repository, 2021.
[13] PaddleOCR authors, "PaddleOCR: Open-source OCR system based on deep learning," 2020.
[14] A. Deng *et al.*, "Table structure recognition and parsing: a survey," *arXiv preprint*, 2020.
[15] Mathpix, "Mathpix API and Math OCR," 2017. [Online]. Available: https://mathpix.com
[16] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge- Intensive NLP Tasks," NeurIPS, 2020.
[17] P. Lewis *et al.*, "RAG: Retrieval-Augmented Generation," NeurIPS Workshop (extended), 2020.
[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP*, 2019.
[19] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation," ACL 2020.
[20] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," J. Mach. Learn. Res., 2020. (T5)
[21] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," arXiv (CLIP), 2021.
[22] A. Khanuja *et al.*, "IndicBERT: A Multilingual ALBERT Model for Indic Languages," 2020.
[23] OpenAI, "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," (2022).