Internship Report

VARADA DHABE
PES1UG22EC326

Project Goal

The objective of this project is to design a highly accurate and scalable Optical Character Recognition (OCR) system tailored for Indian languages, with a focus on scripts such as Kannada, Tamil, Telugu, and others. This system will specialise in recognizing text across diverse fonts, sizes, and layouts, while also enabling precise extraction and reconstruction of complex tables with multiple rows and columns. By harnessing advanced computing resources and cloud infrastructure, the project seeks to automate the digitization of printed books into accessible digital formats. The ultimate aim is to support visually impaired individuals and facilitate accessibility for libraries and institutions dedicated to inclusivity and knowledge dissemination.

Repository created: https://github.com/Varada01/Table_Extractor

In this project, the GPT API is integrated to significantly enhance the performance of the Optical Character Recognition (OCR) system. By leveraging the advanced natural language processing capabilities of the GPT API, the project aims to improve the accuracy and efficiency of text recognition across various fonts, sizes, and layouts, specifically focusing on Indic languages such as Kannada, Tamil, and Telugu. One of the key features of this system is its ability to comprehend the context of the extracted text, which helps in rectifying errors commonly found in traditional OCR systems, especially when dealing with complex scripts and diverse writing styles.

The GPT API is also used to facilitate the reconstruction of tables, ensuring that multi-row and multi-column structures are accurately captured and represented in the digital format. In addition, the API is employed to perform language-specific corrections, making it adaptable to the unique features and syntax of each Indic language, which improves the overall reliability of the digitization process. This integration ensures that the text output is not only more accurate but also semantically aligned with the nuances of the original language, providing a higher level of text fidelity.

By combining the GPT API's linguistic understanding with modern OCR technology, this project automates the digitization of printed books into accessible digital formats. This approach not only accelerates the process but also offers significant support to visually impaired individuals, making books and printed content more accessible. Furthermore, it provides a scalable solution for libraries and institutions working on digitization projects, enhancing their ability to preserve and share knowledge in an inclusive and accessible manner. The result is a robust, high-performance OCR system that can transform how printed materials, especially in Indian languages, are digitised and made available to a wider audience.

Improvement Scopes for the OCR System

Multi-Language Support Expansion:

While the system currently focuses on languages like Kannada, Tamil, and Telugu, expanding support to additional Indian languages such as Hindi, Malayalam, Bengali, and Gujarati could significantly increase its reach and impact. This could involve refining the GPT API's ability to handle various script complexities and dialectal variations in these languages.

Font and Handwriting Recognition:

Improving the system's ability to recognize not just printed fonts but also handwritten text would be a valuable enhancement. Training the OCR model on diverse handwriting styles and cursive scripts, which are commonly used in Indian languages, would increase its versatility and make it applicable to a wider range of documents.

Contextual Accuracy for Complex Documents:

While the GPT API improves contextual understanding, there is room to further enhance its ability to handle complex layouts, such as those found in academic papers, research documents, and newspapers. Enhancements could include better handling of footnotes, annotations, and references that often appear in such texts, ensuring all components are accurately recognized and digitised.

Improved Table and Data Extraction:

The current table reconstruction feature could be further refined to handle more complex table structures, such as those with merged cells, multi-level headers, and irregular row/column data. Developing advanced algorithms for better table recognition, including handling handwritten or scanned tables, would increase the system's utility in converting data-heavy documents.

Cross-Platform Integration:

Improving the system's accessibility by integrating it with various platforms, such as mobile apps, cloud storage, and library management systems, could enhance user experience and usability. A web-based or mobile version of the OCR tool would allow users to digitize and access books and documents on the go, making it even more practical for visually impaired users.

User-Feedback Loop for Continuous Improvement:

Implementing a feedback mechanism where users (especially visually impaired users) can report inaccuracies or improvements in the digitized text could help refine the system. This feedback could be used to train the model further, ensuring continuous improvements in accuracy and adaptability over time.

Support for Non-Standard Fonts and Poor Quality Scans:

Enhancing the system's ability to recognize non-standard fonts and poorly scanned documents (e.g., faded text, low-resolution images) would significantly improve its robustness. Integrating advanced image pre-processing techniques and training the OCR model on lower-quality scans could help overcome these challenges.

Integration with Assistive Technologies:

By integrating the OCR system with existing assistive technologies for visually impaired users, such as screen readers or braille displays, the accessibility of the system could be further enhanced. This would allow visually impaired users to directly interact with the digitized content in real-time, making it more accessible and inclusive.

Localization for Regional Variations:

Considering the wide variety of regional dialects and language variations in India, enhancing the system's ability to recognize regional differences in language and context could make it more adaptable. For example, words or phrases that are unique to certain regions could be better recognized and digitised with higher accuracy.

Speed Optimization:

As the project scales, ensuring that the OCR system operates efficiently even with large volumes of text and data will be critical. Optimising the processing speed, particularly for large book digitization projects, while maintaining high accuracy, will make the system more suitable for large-scale applications in libraries and institutions.

CONCLUSION

In conclusion, this OCR system, powered by the GPT API, holds great promise in revolutionising the digitization of printed Indian language materials, particularly for enhancing accessibility for visually impaired users. By focusing on Indic scripts, complex document layouts, and table reconstruction, the system aims to provide a high level of accuracy and reliability. While there are several avenues for improvement—such as expanding language support, enhancing handwriting recognition, and optimizing performance for diverse document types—the project is well-positioned to make a meaningful impact on libraries, educational institutions, and individuals. Through continuous advancements and the integration of assistive technologies, this system has the potential to transform how printed content in Indian languages is accessed and shared, contributing to a more inclusive and accessible digital world.