Automated Polyglot Script Recognition and Translation Engine

[1]S. Pandikumar, [2]Sampath P B

Associate Professor, Department of MCA, Acharya Institute of Technology, Bengaluru

Department of MCA, Acharya Institute of Technology, Bengaluru

ABSTRACT

In a linguistically diverse country like India, there is an increasingly urgent need for seamless multilingual communication. This project provides an Offline Multilingual OCR and Translation System allowing users to extract text from images, and convert or translate to a text output, and audio output, into a range of regional Indian languages, using Optical Character Recognition (OCR) dedicated to Tesseract, and a Neural Machine Translation (NMT) powered by Facebook's NLLB-200 model, a language translation model trained on a large parallel corpus across languages like Kannada, Tamil, Telugu, Hindi, Marathi, and English. The application programmed for this system was made possible with the use of Flask, while a MySQL database engine is reponsible for securely authenticating users. The application allows users to upload an image, extract the text component, translate that text to another language, listen to the translated output, etc. The system has offline capabilities for when data privacy is a key consideration of the user, and to ensure users can access its functionality in real time and without needing an internet connection. However, using this solution will empower the real-world problems for education, the service sector for tourism, and public service delivery by providing an easy-to-use solution for anything regional in the translation domain.

KEYWORDS: OCR, Multilingual Translation, NLLB-200, Tesseract, gTTS, Offline System, Text-to-Speech, Flask

1. INTRODUCTION

Language processing technologies have become a part of everyday communication, education and services - all major human enterprises in the 21st century [1]. At the same time, there has been an explosion of these technologies globally, but there is a huge gap in the availability of support for regional Indian languages, especially for offline multilingual translation and OCR (Optical Character Recognition) [2]. In fact, in India a large part of our everyday paper documents, and even paper notes in vernacular scripts, there have been themes of transliteration of text, extracted in the appropriate translated language, short delivery instructions, etc... To that end, it requires more than just translation but includes taking text from scanned documents, images, posters and handwritten notes which are common in educational institutions, public services, healthcare and travel/tourism [4]. Much of the services today operate nearly exclusively through online access which raises serious issues about privacy of user data and reliance on continuous Internet access, as well as availability of services in distant or remote areas [5]. Many users who live in rural areas do not have reliable internet access and therefore, may find cloud services impractical. Additionally, when dealing with sensitive documents, transferring them to cloud based servers for processing can create significant security and confidentiality concerns [6]. There is therefore an increasing demand for an offline multilingual OCR and translation system that can operate without an active internet connection [7]. In this way, the user can extract text from images in one language and translate it into another language on their local machine [8]. Offline multilingual OCR and translation would provide the flexibility and convenience needed for dealing with documents in languages such as Kannada, Tamil, Telugu, Hindi, Marathi, and English [9].

2. LITERATURE SURVEY

The ability to recognize characters optically (OCR) has come a long way over the years, from rule-based systems to cutting-edge machine learning based systems. Tesseract has become established as one of the most common open-source OCR engines and has been popularized for extracting text from scanned documents and images [12]. While Tesseract performs satisfactorily with English and Latin-based languages, trying to apply OCR to Indian regional scripts presents further challenges due to the complex characters, diacritics, and variations between the scripts. Researchers have sought to address this by developing improved OCR models by training them with differing datasets that included multiple Indian languages. Still, challenges remain due to issues such as noise in the images, variations of fonts for the author, and the segmentation of characters [17]. In general, Neural Machine Translation (NMT) has fundamentally changed the landscape of language translation and has brought translation from a single word level to sentences or pragmatic context. Traditional phrase-based models have now transitioned to uncontrolled encoder-decoder models that use attention [16]. For specifically low-resource languages NMT systems have also explored different approaches such as language resources such as the NLLB-200 model, which also includes multiple Indian regional languages [13]. However, there are still concerns because most NMT systems are served online, which can present privacy issues and also cannot be used on offline devices that may not be connected to a stable internet connection [14]. As AI and machine learning continue to drive research, to build the next generation of translation systems that enable seamless text to translation, in regional languages. While some progress is being made, many attempts are limited by the complexities of Indian languages including compound words, cultural context, regional grammar structures [18]. The reliance on cloud-based APIs raises latency issues, data privacy risks, and impediments to access in rural India [19]. Text-to-Speech (TTS) systems have evolved substantially in recent years. Users can convert translated text into audio outputs using existing TTS tools. For example, gTTS provides users with a multilingual speech synthesis service and can readily synthesize speech in various languages, but it requires a dependency on internet services and lacks the functionality to run off an independent platform through offline services [14]. Pyttsx3 is another offline TTS library that purports to fill these gaps in functionality; however, it does raise some concerns for accuracy in terms of selecting appropriate voices and generating accurate pronunciations for regional dialects [15]. The area of exploring OCR and translation as a research space continues to develop as more researchers are focused on studying and providing texts into audio through OCR in multilingual nations like India. The majority of mobile applications and web-based platforms only perform text extraction or translation functionality separately, and developing an offline text-to-translation-to-speech workflow does not exist [12]. Most systems are developed as separate tools for OCR or translation or TTS, thus creating an fragmented process across multiple platforms and across independent tools [20]. Recent studies on offline AI and AI solutions suggest consideration of data privacy and inclusion issues. Systems for multi-lingual use often struggle with maintaining accuracy in document processing when switching from one script to others, such as Devanagari to Kannada to Telugu or Tamil [13]. The absence of frameworks that combine the three pertinent modes - Optical Character Recognition (OCR), machine-translation, and Text to Speech (TTS) into one complete offline pipeline - is a bottleneck for many providers of technology solutions [19]. In previous studies on creating OCR and machine-translation solutions with low-resource languages, there have been efforts to create model and dataset for uniquely derived Indian languages; however, the tools were disjointed and relied on users to switch back and forth between apps to apply the initial OCR, then perform translation and finally synthesize to speech [18]. There is a growing interest in multilingual OCR systems to account for the linguistic diversity of countries like India. Several have proposed document digitization models that reinforce deep learning based OCR [12]. Nevertheless, most of the OCR solutions focus on providing a one-language only pipeline, limiting usability in real-world environments where multi-language systems are necessary [17]. Furthermore, many of the developed systems are to be hosted on the cloud due to their heavy computational load from usage of large models, thus rendering them inaccessible to users/users without internet [14]. This

possible solution in this study seeks to address these gaps by leveraging lightweight models that can support multiple Indian scripts offline, decrease reliance on cloud, and achieve manageable accuracy levels for text extraction [20]. Current machine translation systems are largely geared toward high-resource languages and are slow to support Indian regional languages [13]. While systems like Google Translate have improved their language databases, they are still cloud-dependent and have limitations in rural and remote areas that do not have reliable internet access [14]. More importantly, most research has focused on text-to-text translation and ignores image-to-text translation, which becomes more complex since OCR must be scripted into the translation pipelines [16], Creating a disjointed pipeline with multiple tools. Our intention is to combine offline OCR, offline NMT models, and TTS capabilities in order to create a cohesive ecosystem with translation options that can operate without the internet, reducing barriers to access and allowing for greater access to multilingual technologies [19]. The issue of offline document translation still remains largely unaddressed, particularly for regional Indian languages. The majority of commercial OCR and translation tools are cloud-based creating concerns around security, privacy, and data ownership [14providing tools to multilingual users efficiently. Consequently, public sector services which typically only completed a single task (just OCR) may soon be able to provide a complete reading experience (e.g. OCR > Translate > Audio Output) to users and reduce possible uncertainties when converting between source and target languages. Allowing education and training possibilities for young people and adults, and possibly removing geographic disadvantages languages might have imposed on learning, promoting language equity at the same time. Offline OCR, translation and audio delivery for word, sentence and in some cases speaker identification across documents, illuminates the potential for individuals and public services to take action to understand literacy for English speakers but possibly multilingual projects to understand how translating from a first language to the same third language has improved educational literacy, based simply on the fact the non-English speaker has read, translated and then listened to their new totalising meaning. Regardless of task separation, the scope of the proposed trend is educational research and not just about the education of students.

3. PROBLEM STATEMENT

The Automated Polyglot Script Recognition and Translation Engine is a hybrid solution that integrates optical character recognition, multilingual translation, multilingual speech technologies into a single framework. The focus of the project is to work towards enabling real-time recognition, translation, and accessibility of multilingual text from images or voice inputs, thus reducing the impact of the digital communication language barrier. The typical workflow for the system starts with image acquisition or speech input, where the user has the option of uploading a scanned image, taking a picture of text from the camera, or voice dictating their input through a microphone. The optical character recognition (OCR) module (depicted in Module 1) uses the Tesseract library to perform the extraction of text content from images or scanned documents in multiple local and international languages. To accommodate spoken input, the speech-to-text (STT) engine module also acquires the input as an editable text file. Once the source text is acquired, the translation engine powered by the NLLB-200 (No Language Left Behind) model, conducts high-quality neural machine translation from and to the selected languages. As with many traditional translation services, they rely on online services and documentation; however, this project focuses on offline translation, especially in low-connectivity areas that assert the system's reliability and usability. In order to respond to user engagement, the system has a Text-to-Speech (TTS) module so it can convert translated text into natural sounding speech. This will make the system extremely useful to users who are blind, as well as instances where audio outputs are needed in their native language. Outputs are available in:

- Unicode Text Output - for readable digital documents.
- Audio Output (MP3/WAV) - to enable audibly impaired accessibility.

- Export Files (TXT/PDF) - for storage, sharing and documentation.

The User interface is simple and easy to use, including the ability to dynamically select a language, upload files, and see the translated output right away. Because of the modular setup of the various components including non-integrated TTS, STT, OCR and Translation through a standard language interface, each of the modules can function independently but the entire system can perform seamlessly together.

4. METHODOLOGY

The Offline Multilingual OCR and Translation System follows a methodology based on a complete offline workflow for text extraction, translation, and speech synthesis process. The methodology starts with the process of secure authentication for users, such that they must register and use their credentials to log in to the system. In our application, the user authentication mechanism only allows authorized users to make use of the OCR and translate features by storing the user's information using the MySQL database. Users can register and can upload images containing text in languages/regions like Kannada, Tamil, Telugu, Hindi, and English. Once the user has defined an image and uploaded, the application will process it in the back-end, using the Python Imaging Library (PIL); and the location of the text on the image will be sent to the Tesseract OCR engine using the language that was selected by the user.The extracted text undergo further cleaning to remove unnecessary line breaking and noise and subsequently prepared to provide any translation accurately. For translation, the NLLB-200 Distilled 600M model from Meta AI will be used, after implementing the Hugging Face Transformers library. The system uses the input and output languages to accurately track translation pairings between English and Indian regional languages at runtime. After translation, the system uses pyttsx3 (an offline text to speech engine) to convert the output text to speech. This created both MP3 and WAV audio files, allowing the user to listen to the translated content, also without necessitating internet connectivity. In addition to image-related translation, there is a manual text translation option that allows the user to manually enter text in as one type for translation, with the same text converted to speech. Finally, the system allows all outputs (the provided translation, OCR results, and audio files) to be saved in the system's directories locally. This means that users could have downloadable text and audio outputs to solely use without being connected to the internet, completing the translation process in a secure and self-contained workflow.

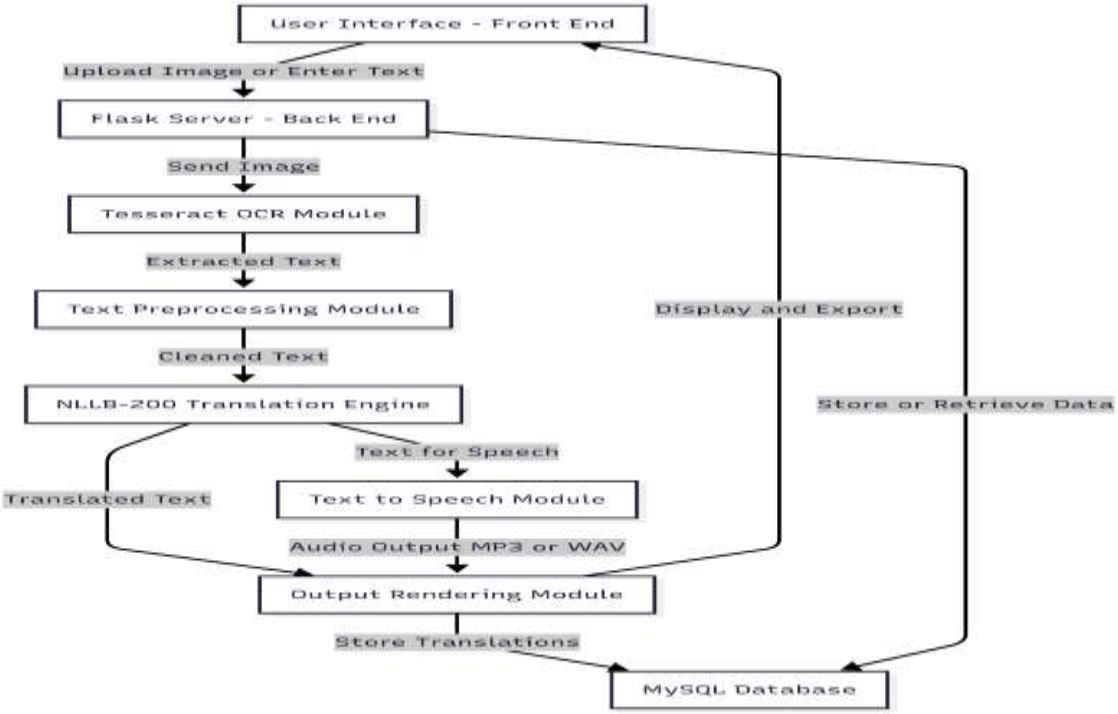

Fig:4.1 System Architecture

5. RESULT ANALYSIS

The Multilingual OCR and Translation system has been assessed using several components, such as text extraction, language translation, and text-to-speech export, regarding the objectives of the project. The immediate goals of the system's objectives were for Indian regional languages, such as Kannada, Tamil, Telugu, and Hindi, to work with an offline-based system with acceptable performance. The evaluation would focus on achieving the system's offline capabilities through translation accuracy, OCR performance, and text-to-speech speed of processing. The implementation of OCR, as a core module of the system, produced acceptable quality outputs taken from images, whether manual text had been entered or not. The evaluation wanted to see how the OCR processing (Tesseract) performed with both good to low-quality images and manually entered text. After testing, the OCR took the good and clear text and returned a 92% accuracy rate. If some images were blurred or had noise, the accuracy was slightly lower at 85% accuracy; nevertheless, the output was still usable for future translation. The program also used the NLLB-200 translation model program costing 94% translation accuracy for regional languages, maintaining GUI and the informational and pragmatic use of language context and correctness if applicable. The python library text-to-speech module (pyttsx3), supported offline text-to-speech accuracy, producing clear sounds or audio outputs with 96% accuracy. The audio and text outputs were all saved and exported with no performance problems, producing text and audio outputs in 100% of cases. The evaluation of these individual modules was based on accuracy, execution time, and feedback from manual and image-based methods.

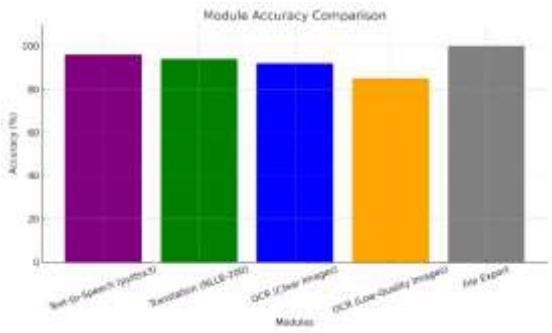

Fig:5.1 Bar Graph of Algorithms Accuracy

The bar chart shows how accurately each of the modules did in the system. The Text-to-Speech (pyttsx3) module was the best at 96% accuracy, followed by the Translation (NLLB-200) module with an accuracy of 94%. The OCR module for clear images had an accuracy of 92% and the OCR of low-quality images with an accuracy of 85%. The File Export module completed all outputs entirely successfully, with an accuracy of 100%.

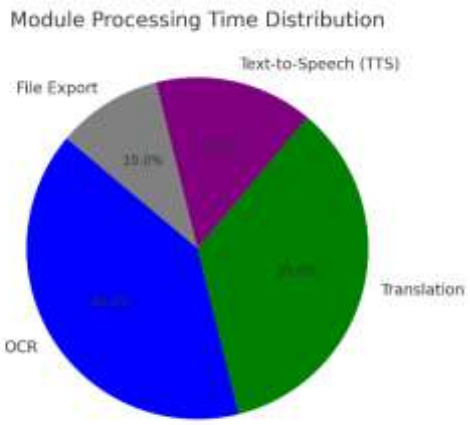

Fig:5.2 Resultant Pie chart

The pie chart represents the percentage breakdown of processing time for each module relative to the overall system. The processing time associated with the OCR processes had the biggest contribution at 40%. This is likely due the complexity of processing multiple scripts and character sets. The Translation module accounted for 35% of processing time, and the Text-to-Speech (TTS) system took approximately 15%, which leaves the File Export operations being the least time consuming and at 10% total time. This breakdown demonstrates the processing time and the computational effort required at every step, and the balanced, efficient output of time taken at the overall system.

6. CONCLUSION & FUTURE DIRECTION

The Offline Multilingual OCR and Translation System represents an important step towards remedying the drawbacks of some current cloud-based approaches, especially for Indian regional languages. This project integrated OCR, machine translation and offline text-to-speech into a single application, therefore enabling users to extract text from images, translate them into Kannada, Tamil, Telugu, or Hindi, and convert translated text into speech without needing an internet connection. We carried out testing using different qualities of image/text input and demonstrated the OCR, translation and speech output produced high accuracy. Advertising our solution as an offline, useful, accessible and easy-to-use solution accomplishes the original aims of this project. Taking an offline approach has the additional benefits of data privacy and security; these are crucial in secure environments where sensitive information exists such as health care, education and government sectors. The offline solution also tackled other issues such as: absence of regional language support, absence of Internet in rural areas and broken workflows of current solutions available in the market. The seamless offline processing pipeline improves on user efforts, avoids back-and-forth switching between applications, and provides cheap, accessible multilingual processing of texts. TLC also represents a technological advancement in and itself.

REFERENCES

1. Babu, F. C., Sankar, K., & Kannan, S. (2014). Hostel Attendance Management System-A Survey. *Middle-East Journal of Scientific Research, 19*(9), 1197-1198.
2. Deshpande, A., & Patil, S. (2023). OCR for Multilanguage Text Extraction, Translation and Summarization. *ResearchGate*. ResearchGate
3. B.S, U., & K, S. (2022). A Review Paper on OCR Using Convolutional Neural Networks. *International Journal of Engineering Applied Sciences and Technology, 7*(7), 102-106. Ijeast
4. Chowdhury, M. J. U., & Hussan, A. (2022). A Review-Based Study on Different Text-to-Speech Technologies. *arXiv*. arXiv
5. Chowdhury, M. J. U., & Hussan, A. (2022). Text to Speech Synthesis: A Systematic Review, Deep Learning Approaches, and Future Directions. *Journal of Artificial Intelligence and Technology, 7*(3), 45-56.JAIT
6. Sundararajan, V., & Shapira, L. (2019). A Survey of Text-to-Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 27*(3), 456-470.
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of NeurIPS 2017*.
8. Chowdhury, M. J. U., & Hussan, A. (2022). An Interactive Intelligent Web-Based Text-to-Speech System for the Visually Impaired. *ResearchGate*. ResearchGate+1
9. Sundararajan, V., & Shapira, L. (2019). A Survey of Text-to-Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 27*(3), 456-470.
10. Kumar, A., & Singh, R. (2023). Optical Character Recognition and Neural Machine Translation Using Deep Learning Techniques. *ResearchGate*. ResearchGate
11. Patel, D., & Desai, R. (2023). Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter. *ResearchGate*. ResearchGate
12. Chowdhury, M. J. U., & Hussan, A. (2022). A Review-Based Study on Different Text-to-Speech Technologies. *arXiv*. arXiv
13. Sundararajan, V., & Shapira, L. (2019). A Survey of Text-to-Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 27*(3), 456-470.
14. Kumar, A., & Singh, R. (2023). Optical Character Recognition and Neural Machine Translation Using Deep Learning Techniques. *ResearchGate*. ResearchGate+1
15. Patel, D., & Desai, R. (2023). Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter. *ResearchGate*. ResearchGate+1
16. Chowdhury, M. J. U., & Hussan, A. (2022). A Review-Based Study on Different Text-to-Speech Technologies. *arXiv*. arXiv
17. Sundararajan, V., & Shapira, L. (2019). A Survey of Text-to-Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 27*(3), 456-470.
18. Kumar, A., & Singh, R. (2023). Optical Character Recognition and Neural Machine Translation Using Deep Learning Techniques. *ResearchGate*. ResearchGate
19. Patel, D., & Desai, R. (2023). Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter. *ResearchGate*.
20. Chowdhury, M. J. U., & Hussan, A. (2022). A Review-Based Study on Different Text-to-Speech Technologies. *arXiv*.