OPTICAL CHARACTER RECOGNITION USING DEEP LEARNING

Raj Singh
Dept of Computer Engineering
PCE New Panvel
raj2002a@gmail.com

Vishnu Kulathunkal
Dept of Computer Engineering
PCE New Panvel
vishnukulathunkal24@gmail.com

Saras Shirgaonkar
Dept of Computer Engineering
PCE New Panvel
sshirgaonkar20comp@student.mes.ac.in

Nishad Mokal
Dept of Computer Engineering
PCE New Panvel
nmokal20comp@student.mes.ac.in

Pooja Bhise
Computer Engineering
PCE New Panvel
poojabhise@mes.ac.in

***Abstract*— One of the main areas of data science is optical character recognition, or OCR. OCR is being used to turn picture data into text data in many different domains. Nowadays, a lot of photos are saved as pictures or in pdf formats, which takes up a lot of memory and storage space and makes them challenging to understand if they are subsequently referenced since the data is derived from several sources. Any sort of text or text-containing document, including handwritten, printed, or scanned text images, may be modified or converted using OCR into a digital format that can be altered and utilised for more extensive processing. The project's objective is to transform visual information into text, which is useful in a variety of fields. It will convert old documents images into texts which can be used where required such as in government offices.**

Keywords: Optical character recognition, OCR, Text to Image, Tesseract.

1.Introduction

In today's digital era, the demand for efficient text extraction and translation from images is paramount across various industries. Optical Character Recognition (OCR) stands at the forefront of this technological advancement, enabling the swift conversion of printed or handwritten text into digital formats.

Leveraging OCR alongside machine translation techniques opens up a realm of possibilities for businesses and organizations seeking to process multilingual content seamlessly. This project embarks on a journey to fuse OCR and machine translation, presenting a robust image-to-text system. Beyond mere translation, it integrates text summarization to distil essential information and a question extraction module to foster deeper engagement. By amalgamating these components, the project aims to offer users a comprehensive solution for extracting, translating, summarizing, and interacting with text content from images efficiently.

2. Literature Survey

M. G. Marne et al. (2018): "Identification of Optimal Optical Character Recognition (OCR) Engine for Proposed System":
This paper focuses on identifying the optimal OCR engine for a proposed system, highlighting Tesseract as the preferred choice. It discusses the stages of OCR and conducts a comparative study of different OCR engines based on parameters like OS support, language support, and more.

P. Sona et al. (2018): "OCR (Optical Character Recognition) Based Reading Aid": The paper introduces an OCR-based reading aid application for the visually impaired, utilizing cloud-based OCR and the MARY TTS engine for text extraction and speech synthesis. It emphasizes multi-language support, user-friendly interfaces, and cost-effectiveness.

S. Dome and A. P. Sathe (2021): "Optical Character Recognition using Tesseract and Classification":
This research presents an OCR web application integrating Tesseract for printed document OCR and a deep learning model for handwritten text recognition. It highlights the use of preprocessing techniques and real-time OCR capabilities.

M. Brisinello et al. (2018): "Optical Character Recognition on images with colorful background": The paper proposes a preprocessing method to enhance Tesseract OCR accuracy on images with colorful backgrounds, demonstrating a significant improvement in OCR performance through image segmentation and classification.

Table 1 Summary of techniques for image enhancement in the literature survey

Paper Name	Advantages	Disadvantages
M. G. Marne et al.	User-friendly, Preprocessing	Memory issues, Manual border removal
P. Sona et al.	Accurate, Multi-language, Cost-effective	Not specified
S. Dome, A. P. Sathe	Accurate, Real-time OCR, User-friendly	Not specified
J.Memom et al,	Works Best for handwritten	Doesn't work for different font styles.

Table 2 Technical Summary of Literature Survey

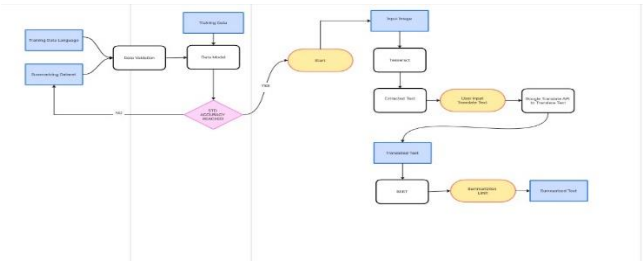
Paper Authors	Models Used	Accuracy
M. G. Marne et al. (2018)	Tesseract OCR	Not given
P. Sona et al. (2018)	HMM-based TTS in MARY TTS	Not given
S. Dome, A. P. Sathe (2021)	Tesseract, Deep learning	96%
M. Brisinello et al. (2018)	Tesseract with preprocessing	62.47%
J.Memom et al, (2020)	HMM, ANN, CNN SVM	Not Available

3. Methodology

The proposed system integrates cutting-edge technologies such as Optical Character Recognition (OCR), Summarization, and Question & Answer (Q&A) systems to provide a comprehensive solution for text extraction, content summarization, and interactive knowledge retrieval. Using OCR technology, the system accurately converts images of printed or handwritten text into machine-readable formats, enabling seamless digitization and analysis. The Summarization module condenses textual information into concise representations, enhancing accessibility and understanding of content. The Q&A system, powered by Natural Language Processing (NLP), automates the extraction of meaningful questions and their corresponding answers from textual data, facilitating knowledge acquisition and comprehension. Together, these modules offer a versatile tool for various applications, including data digitization, information retrieval, and educational support, catering to a wide range of users' needs with efficiency and accuracy.

3.1 System Architecture

The system architecture is given in Figure 2. Important blocks are described in this Section.



Flow Diagram

OCR

The OCR methodology involves a systematic approach to converting images of printed or handwritten text into machine-readable format. It begins with image preprocessing, including noise reduction and binarization, to enhance the clarity of the text in the image. Subsequently, text detection techniques are applied to identify and locate text regions within the pre-processed image. Character segmentation is then performed to break down the text into individual characters, followed by feature extraction to analyze unique characteristics like curves and lines. Character recognition using machine learning models or neural networks matches these features with known characters, resulting in accurate identification. Post-processing steps, including error correction and spatial relationship analysis, ensure the assembled characters form meaningful words, sentences, and paragraphs. The output is a digital text representation that can be further processed or saved in various formats, facilitating text extraction from images for diverse applications such as data digitization and information retrieval.

Language Translation Methodology

The language translation methodology utilizes Natural Language Processing (NLP) techniques and translation models to provide accurate and context-aware translations. It begins with text analysis using NLP tools to preprocess and understand the input text, ensuring compatibility with the translation model. The system integrates a language translation model such as Google Translate or a custom-trained model capable of translating text into multiple languages. Users are provided with the option to select their desired target language for translation, enhancing user flexibility and customization. The translated text is then generated and displayed in the graphical user interface (GUI) for user interaction and comprehension, contributing to a seamless user experience across different languages and content types.

Summarization Methodology

The summarization methodology involves condensing textual information into concise representations using extractive or abstractive summarization techniques. It starts with text preprocessing using NLP tools to tokenize, remove stop words, and prepare the text for summarization. Depending on the system's requirements and complexity, either extractive methods, which select and combine existing text elements based on importance criteria, or abstractive methods, which generate human-like summaries using NLP techniques, are applied. The system may also implement hybrid approaches that combine both extractive and abstractive methods for more informative and coherent summaries. Length control mechanisms are integrated to manage the length of generated summaries based on input text complexity and user preferences. Evaluation metrics such as ROUGE scores or human evaluation are used to assess the quality and coherence of generated summaries, ensuring the effectiveness of the summarization module.

Question Answering (QA) using BERT

The QA methodology leverages pre-trained BERT (Bidirectional Encoder Representations from Transformers) models for accurate and context-aware question answering. It involves tokenizing and processing user questions, ensuring compatibility with BERT's input format and semantic understanding capabilities. The extracted text serves as context for BERT to extract answers accurately, utilizing BERT's ability to comprehend context and provide relevant answers. The system then presents the extracted answers along with the relevant context in the graphical user interface (GUI) for user interaction and comprehension. This methodology ensures that the QA system delivers accurate and context-aware answers, enhancing the overall usability and functionality of the system for knowledge acquisition and comprehension tasks.

3.2. Requirement Analysis

The various hardware and software requirements, datasets used and evaluation metrics are mentioned in detail in this section.

A. Technical Implementation Requirements

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 3.2 and Table 3.3 respectively.

Table 3.1 Hardware details

Processor	Intel(R)Core (TM) i5-9300H CPU @ 2.40GHz
HDD	180 GB
RAM	2 GB

Table 3.2 Software details

Operating System	Windows 11
Programming Language	Python

B. Dataset

We used a custom dataset for Language Detection it can detect 7 languages. Including languages English, Hindi, Tamil, Malayalam, Punjabi, Gujrati, Marathi. Each had approximately 150 tuples. The dataset had two rows Text Data and Target Language.

Table 3.3 Sample Dataset Used for Experiment

Index	Language	No. Of Entries	Total Tuples
1	English	150	250000
2	Hindi	150	180000
3	Marathi	150	120000
4	Malayalam	150	230000
5	Gujarati	150	90000
6	Tamil	150	120000
7	Kanada	150	110000

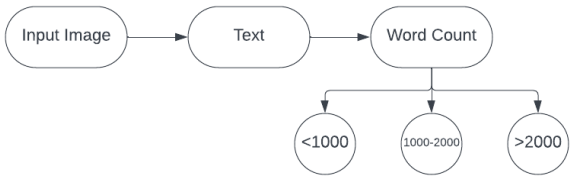
C. Evaluation Metrics

Due to the limited availability of datasets catering to Optical Character Recognition (OCR) in Indian languages, we undertook the initiative to curate a custom dataset. This was accomplished by meticulously sourcing a diverse array of Indian document images from reputable online sources, a testament to our commitment to ensuring the availability of high-quality resources for OCR applications in Indian languages. This tailored dataset not only addresses the prevailing scarcity of such resources but also empowers our system with the capacity to effectively process and analyse textual content in these languages, further enhancing the reach and utility of our project.

Accuracy for Language Detection Model: 98.84 %
Accuracy for Text Summarization: 94.60 %

Evaluation rules for text summarization:

Flow Diagram



Index	Input Text Length	No. of Sentences	Summary Percentage	Expected Output	Actual Output	Resulting Output Sentence
1	904	9	50%	452	502	4
2	1262	17	40%	504	438	7
3	3884	40	30%	1165	1432	13
4	600	6	50%	300	297	4
5	2000	21	40%	800	695	8
6	1500	14	40%	600	573	54
7	1050	9	40%	410	380	4

Table 4.1 Accuracy of Text Summarization

4. Result and Analysis

The OCR Translator and Summarizer application is a comprehensive tool offering a range of functionalities. It utilizes the Tesseract OCR engine for image text extraction, allowing users to upload images with text and extract that text for further processing. The extracted text can then be translated into various languages using the Google Translate API, expanding its usability for multilingual audiences.

Additionally, the application employs TF-IDF vectorization and spaCy for NLP processing to provide text summarization capabilities. This summarization feature condenses the extracted text into a concise summary, aiding users in quickly grasping the main points of the content.

Moreover, the inclusion of a text-to-speech feature through pygame mixer enhances accessibility by enabling users to listen to the translated text in an audible format. This functionality is particularly beneficial for visually impaired individuals or those who prefer auditory learning methods.

Furthermore, the application integrates a BERT-based question-answering model from the transformers library. This allows users to input questions related to the extracted text, and the model generates accurate answers based on the content.

With its diverse set of features ranging from OCR and translation to summarization and question answering, the OCR Translator and Summarizer application caters to various use cases such as language learning, content consumption, accessibility, and research analysis.

The accuracy of the OCR Translator and Summarizer application's models is crucial for delivering reliable results to users. The Tesseract OCR engine, utilized for text extraction from images, has been widely recognized for its robustness and accuracy in recognizing text across various languages and fonts. This ensures that the extracted text is highly accurate, even when dealing with complex or stylized fonts.

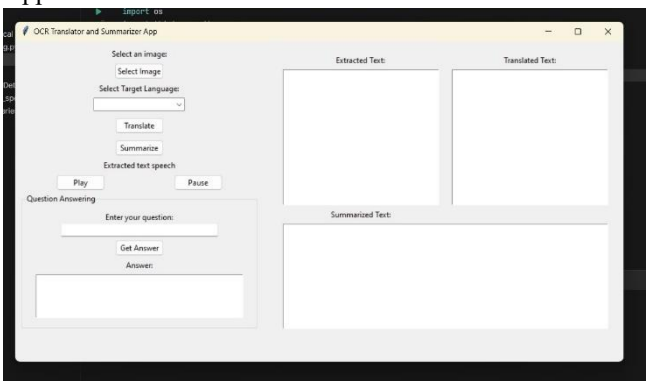
In terms of translation accuracy, the Google Translate API offers state-of-the-art machine translation capabilities, leveraging neural networks and large-scale training data to provide accurate translations between different languages. Users can trust the translated text to maintain the original meaning and context effectively.

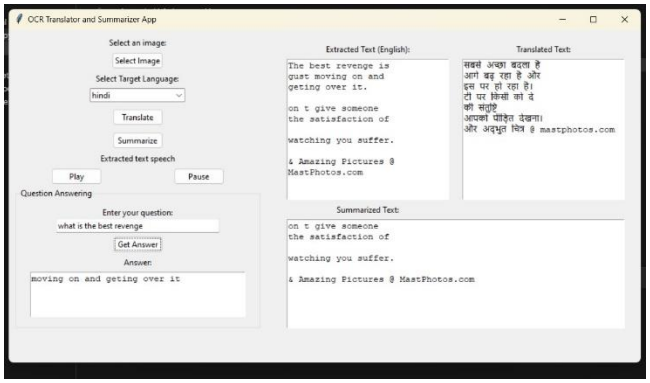
For text summarization, the application's use of TF-IDF vectorization and spaCy for NLP processing contributes to generating concise and relevant summaries. While summarization accuracy can vary based on the complexity and length of the input text, the models employed strive to capture the essential information accurately.

Moreover, the BERT-based question-answering model enhances accuracy in generating answers to user queries based on the extracted text. BERT's contextual understanding and fine-tuned training contribute to providing precise and contextually relevant answers, improving overall user satisfaction and confidence in the application's capabilities.

the OCR Translator and Summarizer application's models prioritize accuracy across all functionalities, ensuring that users receive reliable and meaningful results for their text processing-needs.

Application:





5. Conclusion

The project presents a comprehensive image-to-text system integrating Optical Character Recognition (OCR), text summarization, and a question answering system.

The OCR component leverages advanced techniques to accurately extract text from images, enhancing accessibility and enabling efficient digitization of printed or handwritten content. Additionally, language detection functionality ensures adaptability to diverse linguistic contexts, facilitating seamless processing of multilingual text.

The text summarization module condenses extracted text into concise and coherent summaries, preserving essential information while facilitating efficient content comprehension. This feature enhances accessibility and usability, particularly for large volumes of textual data. Furthermore, the question answering system employs Natural Language Processing (NLP) to automatically generate relevant questions and extract corresponding answers from the extracted text. This functionality fosters deeper engagement and comprehension, empowering users to interact meaningfully with the content.

Moreover, the integration of machine translation techniques enables seamless translation of extracted text into user-specified languages, further enhancing accessibility and usability on a global scale.

Overall, the project's multifaceted approach addresses key challenges in text extraction, comprehension, and interaction, offering a robust solution for efficiently accessing, understanding, and engaging with multilingual text content from images.

6. Future Work

In the expansive domain of Optical Character Recognition (OCR), a promising frontier awaits further exploration and refinement. A key avenue for future research involves optimizing OCR systems to seamlessly handle larger documents, transcending the confines of individual images to encompass comprehensive texts and documents. This optimization necessitates the refinement of algorithms to adeptly process and extract text from diverse document formats, thereby meeting the demands of users grappling with extensive textual data. Concurrently, the ubiquitous integration of mobile devices into contemporary lifestyles presents an opportune path for advancement. Developing OCR-based mobile applications stands as a compelling endeavour, facilitating users in conveniently scanning and extracting text directly from their smartphones or tablets. Such applications offer a portable and user-friendly alternative to traditional desktop solutions, aligning with the evolving needs and preferences of modern users.

Moreover, future endeavours in OCR research must prioritize the enhancement of accuracy through innovative machine learning techniques and meticulous dataset curation. By delving into advanced recognition algorithms and meticulous system parameter fine-tuning, significant strides can be taken towards attaining heightened levels of precision and reliability in text extraction. These advancements not only fortify the utility and effectiveness of OCR technology but also unlock its potential to revolutionize various domains, ranging from document management to language processing and beyond.

References

[1] J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," in IEEE Access, vol. 8, pp. 142642-142668, 2020

[2] M. Brisinello, R. Grbić, D. Stefanovič and R. Pečkai-Kovač, "Optical Character Recognition on images with colourful background," 2018 IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin), Berlin, Germany, 2018

[3] P. Sona, G. V. Mini and K. S. A. Viji, "OCR (Optical Character Recognition) Based Reading Aid," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2018

[4] M. G. Marne, P. R. Futane, S. B. Kolekar, A. D. Lakhadive and S. K. Marathe, "Identification of Optimal Optical Character Recognition (OCR) Engine for Proposed System," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018

[5] S. Dome and A. P. Sathe, "Optical Character Recognition using Tesseract and Classification," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021