IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Halegannada To Hosa Kannada Translator

Prof. Vijay Kumar M S [*1], Bhavana M[*2], Neha L[*2], Prerana B[*2], Thanushree G P[*2]

*1 Assistant Professor, Department of Information Science and Engineering, Maharaja Institute of Technology Mysore, Mysore, Karnataka, India.

*2Undergraduate Student, Department of Information Science and Engineering, Maharaja Institute of Technology Mysore, Mysore, Karnataka, India.

Abstract: Rooted in a rich historical tapestry, Kannada, a language that has evolved over time, presents shifts in pronunciation, sentence structure, words, and meanings. This project aimed at rendering old Kannada more accessible by offering written and spoken interpretations. The project introduces a "smart translator" converting old Kannada into the modern variant, facilitating comprehension through word and sentence mapping. Additionally, it employs Optical Character Recognition (OCR) to recognize handwritten characters within scanned images. The system also integrates Text-to-Speech (TTS) technology, enhancing accessibility for users. By digitizing handwritten text and accurately mapping translations, the project bridges the gap between old and modern Kannada. This comprehensive approach seeks to understand and preserve the linguistic and cultural heritage of old Kannada, ensuring its continued enrichment of modern Kannada and the broader cultural landscape. However, challenges remain in effectively translating and comprehending old Kannada text, hindering access to valuable historical and literary resources. Through ongoing development and refinement, the proposed system aims to overcome these obstacles and contribute to the preservation and accessibility of Kannada heritage.

Keywords: Analysis, investigation, research, Kannada, OCR, TTS.

INTRODUCTION

Kannada, a language steeped in history and evolution, stands as a testament to the cultural richness of its origins. In recent years, there has been a growing interest in exploring the depths of old Kannada literature and understanding its relevance in contemporary times. This introduction seeks to highlight the significance of bridging the linguistic and temporal gap between old and modern Kannada, focusing on the technological advancements aimed at preserving and making accessible the treasures of Kannada heritage. The importance of this topic lies in its contribution to the preservation of linguistic and cultural heritage. Old Kannada texts offer insights into the historical development of the language, reflecting the social, political, and cultural contexts of their time. However, accessing and comprehending these texts pose significant challenges due to linguistic differences and the lack of standardized resources.

Current research in this area emphasizes the utilization of technology to overcome these challenges. Projects integrating Optical Character Recognition (OCR) and Text-to-Speech (TTS) technologies aim to digitize and translate old Kannada texts, making them accessible to a wider audience. Furthermore, efforts to develop smart translators and linguistic databases facilitate the understanding and interpretation of ancient Kannada literature. By providing an overview of the current research landscape, this introduction sets the stage for further exploration into the innovative solutions and methodologies employed in preserving and revitalizing old Kannada literature. Through technological advancements and interdisciplinary collaborations, the rich cultural heritage embedded within old Kannada texts can be safeguarded for future generations, ensuring its enduring legacy in the modern world.

LITERATURE REVIEW

1. Thakare, S., Kamble, A., Thengne, V. and Kamble, U.R., 2018, December. Document segmentation and language translation using tesseract-OCR. In 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) (pp. 148-151). IEEE.

The paper focuses on document segmentation and translation using Python-tesseract and Google Translator. It emphasizes language codes and showcases a web application for image text extraction and translation. Despite manual input limitations, efforts are ongoing to automate language detection. Integration of OCR and translation simplifies document understanding, offering real-time conversion for diverse documents. Future directions include enhancing user experience and automatic language determination, aiming for broader accessibility. The paper contributes to pattern recognition and natural language processing, demonstrating practical application through a web interface. It acknowledges achievements, outlines limitations, and proposes advancements for improved technology.

2. Acharya, Minal, Priti Chouhan, and Asmita Deshmukh. "Scan. It-Text Recognition, Translation and Conversion." 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2019.

Electronic documents are essential in workplaces, government offices, and educational institutions. However, many texts, such as letters and documents, require conversion into readable regional languages. Conventional scanners are bulky and inconvenient. This paper discusses the need for document scanning and introduces Scan.it, a portable application addressing language barriers by converting scanned documents. Scan.it utilizes Natural Language Processing for text recognition and translation, resolving ambiguity between similar words and classifying data contextually. Implemented with Node.js, it employs Tesseract OCR for text recognition and Translator for translation. Unlike other OCR techniques, Scan.it accurately recognizes Marathi text and preserves soft copies of documents.

3. Ramesh, G., Sandeep Kumar, and H. N. Champa. "Recognition of Kannada handwritten words using SVM classifier with convolutional neural network." 2020 IEEE Region 10 Symposium (TENSYMP). IEEE, 2020.

The disorganized layout and poor print quality present difficulties for many recognition algorithms for Indian document images. A character segmentation approach is suggested for Kannada handwriting recognition in order to solve these problems. It uses graph distance theory to separate overlapped characters and major segmentation paths based on character structural features. The Support Vector Machine (SVM) classifier is used for validation. Comparing simulations over a range of databases, traditional methods are outperformed in terms of accuracy and sensitivity.

4. Dome, Saurabh, and Asha P. Sathe. "Optical charater recognition using tesseract and classification." 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2021.

To transform text from digital images into editable data, OCR is essential. Our WebApp offers a high-accuracy, ad-free user experience that satisfies the need for speedy, dependable tools. OCR was developed initially for printed text, but it can now read handwritten text as well. Our approach seeks to reduce expensive data entering by automatically extracting information from paper documents. By doing away with human interaction, document processing drastically cuts expenses and time.

5. Narang, Sonika Rani, Munish Kumar, and Manish Kumar Jindal. "DeepNetDevanagari: a deep learning model for Devanagari ancient character recognition." *Multimedia Tools and Applications* 80 (2021): 20671-20686.

Many historical manuscripts are preserved in the Devanagari script, which is extensively used in Asia and India. Optical Character Recognition is a key component of the digitization efforts for these documents (OCR). Convolutional Neural Network (CNN) has great performance in character and pattern recognition, however it has not yet been fully utilized for Devanagari manuscript recognition. Our goal is to use CNN's capabilities to glean information from these texts. We test various CNN design alternatives and suggest using it as a classifier and feature extractor for 33 fundamental Devanagari letters. Our studies show greater accuracy (93.73%) over state-of-the-art methods using a dataset of 5484 characters. With this method, the wealth of Devanagari manuscripts will become accessible to a wider audience for study and accessibility.

METHODOLOGY

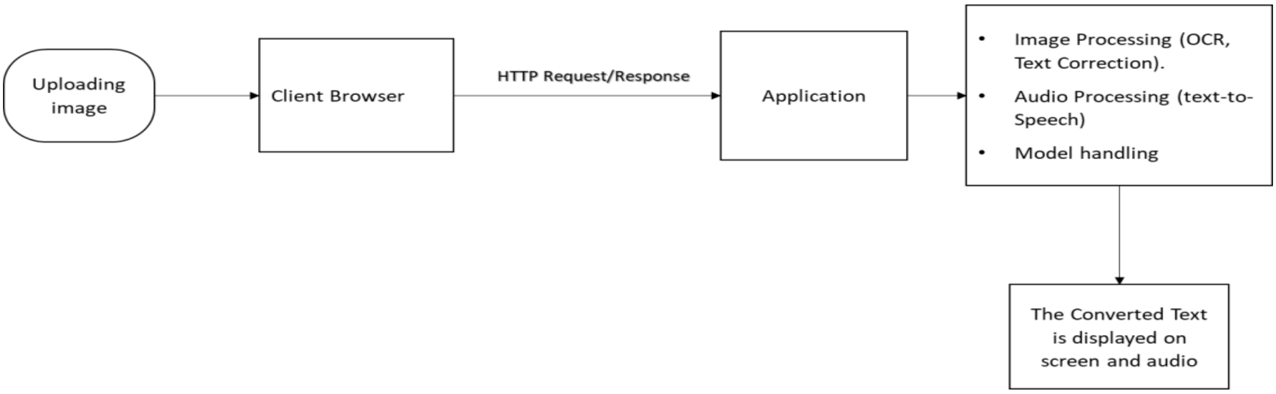

Uploading image
Client Browser
HTTP Request/Response
Application
Image Processing (OCR, Text Correction).
Audio Processing (text-to-Speech)
Model handling
The Converted Text is displayed on screen and audio

Figure 1: Architecture for proposed System

The proposed system integrates Optical Character Recognition (OCR) and Text-to-Speech (TTS) technologies to bridge the linguistic gap between old and modern Kannada. Through OCR, handwritten text from scanned documents is digitized, while the smart translator ensures accurate mapping, and TTS enhances accessibility. This system offers a comprehensive solution for translating outdated Kannada script, originally presented as image formats, into legible text enriched by dictionary mapping and a voice-over function.

The process begins with the upload of outdated Kannada script images, initiating a sequence that may require iteration to ensure full transformation into an accessible format. The system interprets the visual script, converting it into an internal code for systematic processing. This internal code serves as the foundation for constructing individual words from the picture text, enabling a more thorough comprehension and manipulation of the content.

Integral to the system is a comprehensive dictionary containing pairs of old Kannada words and their modern Kannada counterparts. Leveraging the internal code, each constructed word is mapped with entries in the dictionary, ensuring a faithful and accurate transformation of old Kannada into the modern language. As a result, old Kannada words seamlessly find their place alongside their modern counterparts, rendering ancient scripts accessible and comprehensible to contemporary readers.

Furthermore, the system incorporates a voice-over feature, allowing users to listen to the text in addition to reading it. This enhancement significantly increases accessibility, particularly benefiting individuals with visual impairments and those who prefer auditory learning methods. By combining OCR, smart translation, and TTS technologies, the proposed system offers a holistic solution for bridging linguistic gaps and making ancient Kannada literature accessible to a broader audience.

Optical Character Recognition (OCR) using Tesseract

Optical Character Recognition (OCR) revolutionizes text digitization by converting printed or handwritten content into machine- readable text. Tesseract, an open-source OCR engine developed by Google, stands out for its accuracy and versatility. Exploring Tesseract's anatomy reveals its robust features tailored for diverse OCR tasks, supporting multiple languages and font styles across various platforms. Its workflow involves preprocessing, layout analysis, text recognition, and postprocessing, ensuring precise results. Key features include extensive language support, adaptability to diverse fonts, and open-source collaboration for continuous enhancement. Tesseract's practical applications range from document digitization to data extraction and automated text recognition in images. Seamlessly integrating with languages like Python and frameworks like OpenCV, Tesseract empowers developers to leverage its capabilities for diverse applications and workflows.

SVM Model Creation: Unveiling the Architectural Choices:

The sentiment analysis model hinges on SVM, known for multiclass classification proficiency. Selecting the SVM kernel is pivotal, with a linear kernel chosen for simplicity and interpretability. Linear kernels draw straight decision boundaries, aiding comprehension for practitioners. The regularization parameter (C) is calibrated at 1.0, striking a balance between model simplicity and adaptability to training data intricacies. This prevents overfitting, ensuring the model generalizes well to unseen data. The aim is not only accurate sentiment predictions but also a transparent model, accessible to stakeholders seeking insights. This strategic combination equips the SVM model to navigate the complexities of sentiment analysis effectively. It merges the sophistication required for multiclass sentiment classification with a commitment to transparency, laying the groundwork for comprehensive model evaluation and interpretation.

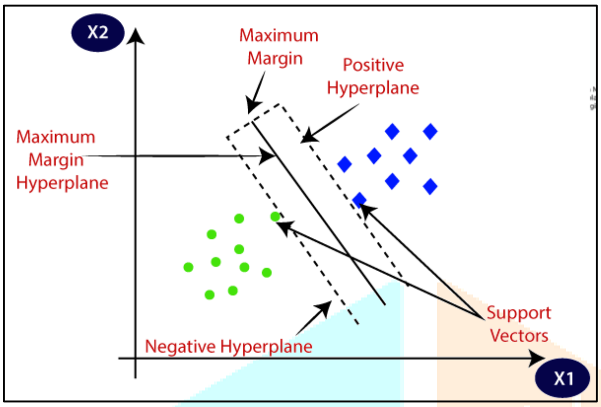


Figure 2: Support Vector Machine Classifier

EXPERIMENT AND ANALYSIS

The experiment setup entails developing an application for Optical Character Recognition (OCR) and text-tospeech (TTS) capabilities. This application incorporates a pre-trained Support Vector Machine (SVM) model for text classification. Upon receiving a POST request with an uploaded image, the application initiates image processing using Tesseract OCR to extract textual content. The extracted text undergoes correction utilizing a symbol mapping dictionary retrieved from a JSON file. Subsequently, the corrected text is synthesized into speech format through the gTTS library and saved as an MP3 file. The application's interface renders the generated audio file, providing users with accessible speech output.

The application architecture revolves around Flask, a Python web framework, facilitating the deployment of a user-friendly interface. The application's main routes include '/ocr' and '/audio/<filename>', responsible for handling OCR and audio file serving functionalities, respectively. The OCR route processes uploaded images, performs text extraction, correction, and converts the corrected text into speech. The audio file is then made available for download or playback within the application.

The core components of the application consist of Tesseract OCR, responsible for optical character recognition from images, and gTTS for converting text into speech. Additionally, a symbol mapping dictionary aids in correcting recognized text, enhancing accuracy and readability.

For robustness and flexibility, the application is configured to run locally with debug mode enabled, facilitating real-time testing and development adjustments. This setup streamlines the process of image-based text extraction, correction, and synthesis into speech, catering to users seeking seamless OCR and TTS functionalities.

The experiment setup encompasses the development and deployment of an application with OCR and TTS capabilities. Leveraging Tesseract OCR, SVM classification, and gTTS synthesis, the application provides a comprehensive solution for extracting, correcting, and vocalizing text from images. With a user-friendly interface and robust functionality, the application aims to enhance accessibility and usability in text processing tasks.

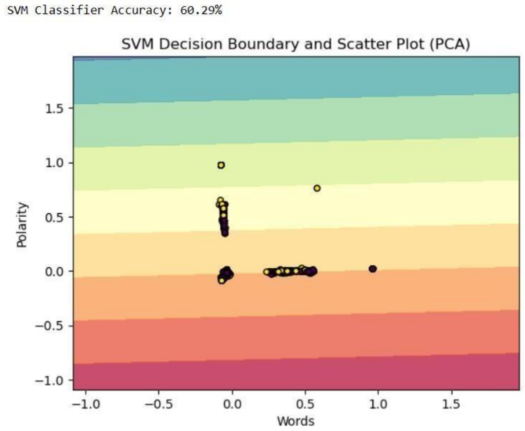
Analysis Result

The integration of both code segments results in a tailored application for Kannada text sentiment analysis. Leveraging a Support Vector Machine (SVM) classifier trained on a comprehensive Kannada lexicon dataset, the application accurately discerns sentiment polarity. Utilizing TF-IDF features extracted from textual data and

Principal Component Analysis (PCA) for dimensionality reduction, the SVM model achieves a 60.29% accuracy on the test dataset, validating its efficacy.

The application offers a visual depiction of the SVM classifier's decision boundary through a scatter plot, effectively illustrating sentiment polarity segregation. This visualization aids users in understanding sentiment analysis outcomes within Kannada text. Overall, the application provides a robust solution, merging advanced machine learning techniques with user-friendly visualization capabilities.

Empowering users to analyze sentiment polarity in Kannada text seamlessly, the application facilitates various natural language processing tasks. From social media sentiment analysis to tailored content creation, its utility spans diverse domains. This amalgamation of cutting-edge methodologies and intuitive visualization signifies a significant stride in Kannada text analysis tools. Positioned to advance Kannada language processing, the application promises deeper insights into sentiment dynamics within Kannada textual content.



```
SVM Classifier Accuracy: 60.29%
```

Figure 3: Graph of SVM model

CONCLUSION

The preservation of Halegannada script, vital to Karnataka's cultural heritage, is a cornerstone of our proposed system. Integrating Optical Character Recognition (OCR) and Text-to-Speech (TTS), it transforms ancient scripts into spoken formats, fostering accessibility and comprehension of Kannada's evolution. This initiative expands historical content accessibility, aiding scholars and readers in exploring Kannada's rich heritage. It embodies a cultural mission to safeguard linguistic and cultural treasures, ensuring their vibrancy for future generations. By marrying technology with cultural preservation, the system empowers individuals to delve into Karnataka's history, culture, and literary traditions, preserving their significance for posterity.

REFERENCES

[1] Thakare, S., Kamble, A., Thengne, V. and Kamble, U.R., 2018, December. Document segmentation and language translation using tesseract-OCR. In 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) (pp. 148- 151). IEEE.

[2] Acharya, Minal, Priti Chouhan, and Asmita Deshmukh. "Scan. It-Text Recognition, Translation and Conversion." 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2019.

[3] Ramesh, G., Sandeep Kumar, and H. N. Champa. "Recognition of Kannada handwritten words using SVM classifier with convolutional neural network." 2020 IEEE Region 10 Symposium (TENSYMP). IEEE, 2020.

[4] Dome, Saurabh, and Asha P. Sathe. "Optical charater recognition using tesseract and classification." 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2021.

[5] Dome, S. and Sathe, A.P., 2021, March. Optical charater recognition using tesseract and classification. In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 153-158). IEEE.

[6] Tang, M., Xie, S. and Liu, X., 2023. Ancient character recognition: a novel image dataset of Shui manuscript characters and classification model. Chinese Journal of Electronics, 32(1), pp.64-75.

[7] Jain, P.H., Kumar, V., Samuel, J., Singh, S., Mannepalli, A. and Anderson, R., 2023. Artificially intelligent readers: an adaptive framework for original handwritten numerical digits recognition with OCR Methods. Information, 14(6), p.305..

[8] Krishnan, P. and Jawahar, C.V., 2019. HWNet v2: an efficient word image representation for handwritten documents. International Journal on Document Analysis and Recognition (IJDAR), 22(4), pp.387-405.