e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

SCRIPT SCRIBE: EMPOWERING LESS KNOWN LANGUAGES THROUGH HANDWRITING RECOGNITION

Mr. Srinidhi Kulkarni*1, Amrutha S*2, H Ashwin Kumar*3, H V Pavani*4, Jeevan N M*5

*1Assistant Professor, Dept of CSE Jyothy Institute Of Technology Bangalore, India.

*2,3,4,5Computer Science Engineering Jyothy Institute Of Technology Bangalore, India.

ABSTRACT

This project endeavors to safeguard the linguistic and cultural heritage of the Tulu-speaking community by employing advanced technologies to digitize ancient palm leaf manuscripts. Combining Optical Character Recognition (OCR) for Tulu letters and digits, as well as linguistic analysis, we aim to bridge traditional scripts with contemporary communication methods. The project not only focuses on converting Tulu scripts into the Kannada script but also delves into unraveling the intricate linguistic evolution within the South Indian Dravidian context. By creating specialized datasets, our approach optimizes OCR processes for Tulu language nuances. This interdisciplinary effort not only contributes to the digitization and preservation of Tulu heritage but also holds the potential to advance the broader field of Dravidian language studies, offering insights into the evolution of both linguistic and numerical notations in the South Indian context.

Keyword - South Dravidian Language, Optical Character Recognition.

I. INTRODUCTION

The rich cultural heritage of the Tulu-speaking community is embedded in its ancient palm leaf manuscripts, providing invaluable insights into the linguistic and historical evolution of the region. With a commitment to preserving this unique linguistic legacy, our project focuses on the meticulous task of extracting Tulu letters from these aged palm leaves and employing cutting-edge Optical Character Recognition (OCR) technology to recognize and digitize the text.

Our primary objective is to seamlessly bridge the transition from traditional Tulu scripts to contemporary communication by converting recognized Tulu letters and digits into the Kannada script. This transformative process not only facilitates broader accessibility to Tulu literature but also contributes significantly to its preservation, ensuring its endurance for future generations.

Furthermore, our initiative extends beyond mere transcription, aiming to optimize the recognition of the South Indian Dravidian evolution embedded within the Tulu language. By leveraging advanced linguistic analysis and historical linguistics, we seek to unravel the intricate linguistic threads that connect Tulu to the broader Dravidian language family, shedding light on its evolution over centuries.

Handwriting recognition is a process that involves converting handwritten text into machine-readable text. The recognition can be categorized into online and offline handwriting recognition, each with its specific characteristics. Below is an elaborate explanation of the steps involved in the handwriting recognition process, including image preprocessing, resizing, feature extraction, training, and testing.

A. IMAGE PREPROCESSING:

a. Noise Reduction:

Objective: Remove unwanted elements from the image that may interfere with character recognition.

Methods: Techniques like Gaussian blurring, median filtering, or morphological operations are applied to reduce noise.

b. Contrast Enhancement:

Objective: Improve the visibility of characters by adjusting the contrast.

depends on the quality of the preprocessing steps, the effectiveness of the feature extraction techniques, and the choice of a suitable machine learning algorithm.

Methods: Histogram equalization, contrast stretching, or adaptive histogram equalization can be used to enhance the contrast of the image.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

B. RESIZING THE IMAGE:

Objective: Standardize the size of the images in the dataset to ensure consistency.

Method: Images are resized to a predefined dimension, often converting them to a common resolution suitable for processing.

C. FEATURE EXTRACTION:

Objective: Transform the raw image data into a format suitable for machine learning algorithms.

Methods:

Pixel-based Features: Extracting features based on pixel values, such as intensity gradients, edges, and textures.

Shape-based Features: Analyzing the shape and structure of characters.

Statistical Features: Calculating statistical properties of the image, like mean, variance, etc.

Zoning and Grid-based Features: Dividing the image into zones or grids and extracting features from each section.

D. TRAINING THE MODEL:

Objective: Train a machine learning model to learn patterns and relationships in the preprocessed data.

Algorithm: Various algorithms can be used, such as neural networks, support vector machines (SVM), k-nearest neighbors (KNN), or deep learning models.

E. TESTING FOR CHARACTER IDENTIFICATION:

Objective: Evaluate the trained model's performance on new, unseen data.

Method: The model is tested on a separate dataset to assess its Beyond the realms of transcription, our initiative delves into the fascinating world of South Indian Dravidian evolution embedded within the Tulu language. Utilizing advanced linguistic analysis and historical linguistics, we strive to untangle the intricate linguistic threads connecting Tulu to the broader Dravidian language family, offering insights into its evolution over centuries.

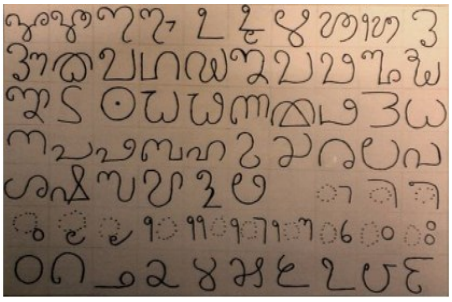

Fig 1: Tulu Script

Fig 2: Tulu Digits

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

Recognizing the need for precision, we extend our focus to Tulu digit recognition, enhancing the digitization process for a more comprehensive understanding of numerical notations within Tulu manuscripts. Simultaneously, we acknowledge the significance of crafting specialized datasets tailored to the unique characteristics of the Tulu language, contributing to the accuracy and efficiency of OCR processes.

This interdisciplinary approach not only addresses the digitization and preservation of Tulu heritage but also holds the potential to advance the broader field of Dravidian language studies. Through this project, we aspire to build a bridge between the ancient wisdom encoded in Tulu palm leaf manuscripts and the digital age, fostering a deeper understanding of linguistic and numerical evolution in the South Indian context.

II. LITERATURE SURVEY

SL	TITLE	AUTHORS	CONCLUSION
1	Recognition Of Handwritten Word Images	Umesh D. Dixit Rahul Hiraskar Raghavendra Purohit Sagar Shivanagutti	The paper proposes a Handwritten Character Recognition (HCR) system for word image recognition using Histogram of Oriented Gradients (HOG) features. Comparing K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) classifiers, the SVM achieves a 75% word recognition rate, demonstrating the effectiveness of the proposed approach in digitizing manual scripts for document image processing.
2	Handwritten Arabic Optical Character Recognition Approach Based on Hybrid Whale Optimization Algorithm With Neighborhood Rough Set	Ahmed Talat Sahlol Mohamed ABD Elaziz Mohammed A.A.AL-Qaness Sunghwan Kim	The paper introduces a novel approach for handwritten Arabic character recognition, emphasizing the need for digitization in handling the intricacies of Arabic script. It proposes a four-stage method, combining Binary Whale Optimization Algorithm (BWOA) with Neighborhood Rough Set (NRS) for feature selection, showcasing potential improvements in Arabic character recognition efficiency on the CENPARMI dataset.
3	RNN Based Online Handwritten Word Recognition in Devanagari Script	Rajib Ghosh Pooja Keshri Prabhat Kumar	The paper presents a novel approach for online handwritten word recognition in Devanagari script using Recurrent Neural Network (RNN), specifically LSTM and BLSTM, addressing unique challenges in Indic scripts. By employing local zone wise stroke analysis and feature extraction, the proposed method achieves superior recognition rates compared to existing methods, including those utilizing Hidden Markov Model (HMM), and organizes its content to discuss related works, theoretical background, data collection, preprocessing, RNN- based recognition, and results.
4	An Efficient Translation of Tulu to Kannada South Indian Scripts using Optical Character Recognition	Dr.I.Manimozhi Dr.Manoj challa	A proposed framework aims to preserve Tulu language palm leaves using MATLAB and OCR techniques for character recognition. The Tulu script closely resembles Malayalam, and manual identification is employed initially, followed by mapping characters through OCR. The digitized Tulu script is crucial for cultural preservation. Efforts by

			the Tulu Academy and the Karnataka government include teaching Tulu at various levels, conducting research programs, and promoting it as a third optional language in schools. The overall goal is to empower Tulu through education, research, and cultural initiatives.
5	Bangla Handwritten Word Recognition System Using Convolutional Neural Network	Md. Tanvir Hossain Md. Wahid Hasan Amit Kumar Das	Bangla, an Eastern South Asian language, is spoken by nearly 300 million people and is the sixth most spoken language globally. It serves as the national language of Bangladesh and is also official in West Bengal and Tripura. Bangla's alphabet includes 11 vowels, 39 consonants, and compound characters. Handwritten word recognition in Bangla is a current research focus, leveraging Convolutional Neural Networks (CNN) for efficient and accurate results. CNN, widely used in image processing and Natural Language Processing, plays a crucial role in overcoming challenges in Bangla handwritten character recognition.
8	Towards Spotting and Recognition of Handwritten Words in Indic Scripts	Kartik Dutta Praveen Krishnan Minesh Mathew C.V. Jawahar	The document proposes a comprehensive scheme for collecting and annotating large-scale handwritten data for low-resource languages like Indic scripts. It releases a significant annotated dataset for Telugu, benchmarking major Indic scripts for text spotting and handwritten recognition with state-of-the-art deep neural architectures, emphasizing the effectiveness of synthetic and real handwriting data in improving word recognition results. The paper focuses on challenging handwritten word images, utilizing an End2End deep network for word spotting and recognition, while discussing the efficacy of different architectural choices and training pipelines.
9	RNN based online handwriting recognition in Devanagari Script	Authors-Rajib Ghosh Pooja Keshri Prabhat Kumar	The paper presents a groundbreaking approach to online handwritten word recognition in Devanagari, utilizing Recurrent Neural Networks (LSTM and BLSTM). The proposed method focuses on local zone-wise analysis of basic strokes, departing from traditional holistic approaches. With a 10K-word lexicon, the system achieves a remarkable accuracy of 99.50%, outperforming existing Hidden Markov Model (HMM)-based systems. The study envisions practical deployment in hand-held devices and outlines future extensions to sentence and paragraph recognition.
10	Read and Recognition of	Imran Khan	The abstract outlines a novel approach, the Mean, Standard Deviation, and Sum of Absolute Difference

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

	old Kannada Stone Inscriptions Characters using MSDD Algorithm	Ancy Elizabeth Megha S G Prakruti K S Gowthami G R	Algorithm (MSDDA), for the recognition of ancient Kannada characters from stone inscriptions using computer vision. This method offers a cost-
6	Deep Learning Network Architecture based Kannada Handwritten Character Recognition	N. Shobha Rani Subramani A C Akshay Kumar P B R Pushpa	The document proposes a character recognition system for Kannada handwritten characters using transfer learning and VGG NET 19. It leverages the knowledge of Devanagari characters for classification, achieving satisfactory accuracy. The methodology involves transfer learning with a convolutional neural network (CNN) and global average pooling, while the results are represented through graphs depicting accuracy and loss, suggesting potential enhancements with increased resolution and instances.
7	Off-line Telugu Handwritten Characters Recognition using optical character recognition	N Prameela P Anjusha R Karthik	The document outlines a Telugu character recognition system employing normalization and skeletonization techniques. It introduces the use of Principal Component Analysis (PCA) for dimensionality reduction of curvature feature vectors extracted from Telugu script. Various classifiers are employed, with Support Vector Machines (SVM) achieving a recognition rate of 80.6%, and Quadratic Discriminant Analysis (QDA) achieving 87.6%. The proposed OCR system for Telugu is shape and font dependent, involving pre-processing, feature extraction, and is designed to fulfil the practical need for effective Telugu character recognition.
			effective solution to the digitization of old stone inscriptions, bypassing the time and resource-intensive conventional methods. The algorithm, focusing on Hoysala and Ganga periods, employs mean, standard deviation, and sum of absolute difference values for character recognition, demonstrating high accuracy and efficiency, especially in handling faded and unprotected inscriptions. The proposed technique involves capturing, pre-processing, and utilizing a database for recognition, achieving a recognition rate of 98.75%. This approach not only contributes to the preservation of historical artifacts but also facilitates the accessibility of ancient Kannada literature through digitalization.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

III. METHODOLOGY

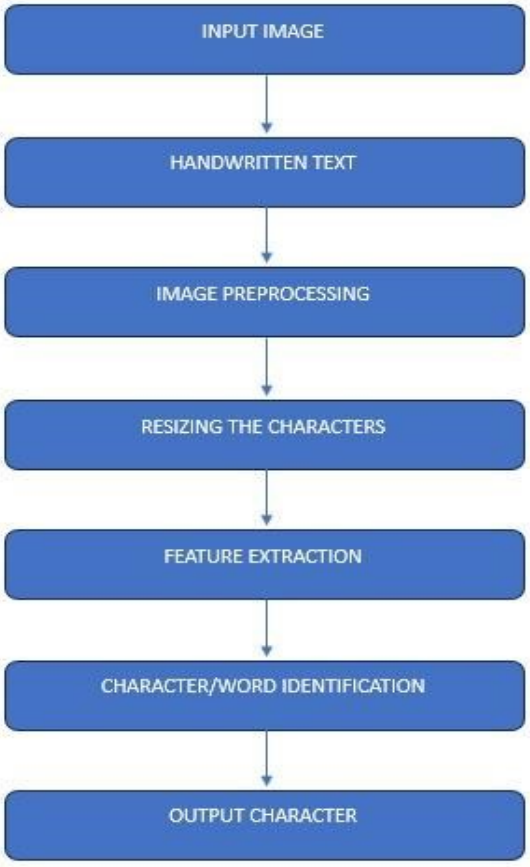

Fig3: Flow chart for methodology accuracy and effectiveness in identifying characters.

F. OUTPUT:

Objective: Provide the most probable character(s) based on the model's predictions.

Method: The model outputs the characters it predicts with a certain level of confidence, and the character with the highest confidence is considered the final output.

Handwriting recognition systems often require fine-tuning and optimization based on the specific dataset and application requirements. The success of the system

IV. CONCLUSION

In summary, our project employs cutting-edge Optical Character Recognition (OCR) technology to digitize ancient Tulu palm leaf manuscripts, encompassing both Tulu letters and digits. The conversion of Tulu scripts into Kannada facilitates digital accessibility to Tulu literature and bridges traditional and contemporary linguistic expressions. Our exploration of the South Indian Dravidian evolution within Tulu enriches our understanding of regional linguistic history. The creation of specialized datasets, tailored to the unique nuances of the Tulu language, enhances OCR accuracy, emphasizing the importance of customization. Our project not only preserves Tulu heritage but also contributes to broader research on linguistic and numerical evolution in the South Indian cultural context.

V. REFERENCES

[1] Sahlol, A. T., Abd Elaziz, M., Al-Qaness, M. A. A., & Kim, S. (2020). Handwritten Arabic Optical Character Recognition Approach Based on Hybrid Whale Optimization Algorithm With Neighborhood Rough Set. IEEE Access, 8, 23011–23021.

[2] Dixit, U. D., Hiraskar, R., Purohit, R., & Shivanagutti, S. (2020). Recognition Of Handwritten Word Images. 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC).

[3] Keshri, P., Kumar, P., & Ghosh, R. (2018). RNN Based Online Handwritten Word Recognition in Devanagari Script. 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR).
[4] Manimozhi, I., & challa, M. (2021). An Efficient Translation of Tulu to Kannada South Indian Scripts using Optical Character Recognition. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC).
[5] Hossain, M. T., Hasan, M. W., & Das, A. K. (2021). Bangla Handwritten Word Recognition System Using Convolutional Neural Network. 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM).
[6] C K Savitha, P J Antony, Machine Learning Approaches for recognition of offline Tulu Handwritten Scripts
[7] November 2018Journal of Physics Conference Series 1142(1):012005
[8] Rani, N. S., Subramani, A. C., Kumar P., A., & Pushpa, B. R. (2020). Deep Learning Network Architecture based Kannada Handwritten Character Recognition. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA).
[9] Prameela, N., Anjusha, P., & Karthik, R. (2017). Off-line Telugu handwritten characters recognition using optical character recognition. 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA).
[10] Prameela, N., Anjusha, P., & Karthik, R. (2017). Off-line Telugu handwritten characters recognition using optical character recognition. 2017 International Conference of Electronics,
[11] Zhiqi, Y., & Kai, F. (2018). Design and implementation of handwritten digit recognition system based on template method. 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC).