OPEN ACCESS

Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 04 | Issue 02 | March-April | 2024
Journal homepage: https://gjrpublication.com/gjrecs/

Review Article

A Comprehensive Review on BARAVU-Tulu Lipi Identification

[1]Midhun Varghese, [2]Suraksha*, [3]Abhinav Vinod, [4]Suhail Abdul Nazir, [5]Dr.Anoop B K

[1,2,3,4,5]Department of AIML Srinivas Institute of Technology Mangaluru, India

DOI: 10.5281/zenodo.10969139 Submission Date: 10 March 2024 | Published Date: 13 April 2024

***Corresponding author:** Suraksha
Department of AIML Srinivas Institute of Technology Mangaluru, India

Abstract

In this project, we present BARAVU-Tulu, a system dedicated to Tulu lipi identification leveraging Convolutional Neural Networks (CNNs). The intersection of image processing and deep learning has spurred significant advancements, particularly in handwritten text recognition. Our endeavor focuses on crafting an advanced machine learning model proficient in analyzing document images featuring Tulu script, with a specific emphasis on the linguistic nuances prevalent in the Northern part of Kerala and the Southern part of Karnataka in India. Initially, we meticulously collect a diverse dataset of printed Tulu script, encompassing various styles, fonts, sizes, and orientations. Employing CNN architectures, such as VGG or ResNet, we train a model capable of accurately identifying printed Tulu script from input images. Subsequently, we extend our efforts to gather a dataset of handwritten Tulu script, catering to the distinctive characteristics of individual handwriting styles. By fine-tuning the CNN model with the handwritten dataset, we augment its capabilities to encompass both printed and handwritten Tulu script identification seamlessly. A user-friendly interface facilitates interaction with the BARAVU-Tulu system, enabling users to input images containing Tulu script and obtain precise identification results. Beyond academic contributions, our system holds practical implications, particularly in facilitating communication for travelers in the targeted regions, overcoming barriers posed by diverse communication mediums and languages. Through meticulous methodology and rigorous evaluation, our project exemplifies the synergy between image processing, deep learning, and linguistics, culminating in an effective solution for Tulu lipi identification.

Keywords: Handwritten Text Recognition, Language Recognition, Convolution Neural Networks (CNN).

I. INTRODUCTION

Tulu is the most spoken language in the coastal regions of southern Karnataka and northern regions of Kerala. The script used for this language is a script with great heritage and is practiced by only a handful of people. To enhance exposure for this language around the world this project focuses on building a “Tulu Lipi Identification” system that can understand the scripts in Tulu and convert them into readable text in English. The unique script is intricately woven into the deep heritage of the Tulu-speaking communities. At its core, this project is propelled by advanced technologies, notably Convolutional Neural Networks (CNN) and deep learning algorithms, to automate the task of identifying and comprehending handwritten Tulu Lipi. The significance of this project extends far beyond the mere recognition of script; it embodies a commitment to the preservation of cultural identity and a chance for exposure of the language to the external world. The project’s initial stride involves the meticulous creation of a comprehensive dataset, capturing the diverse nuances and stylistic variations inherent in handwritten Tulu Lipi. This collection of data, like a small sample showcasing the detailed aspects of the Tulu script, becomes the essential building block for creating deep learning algorithms. Convolution Neural Networks (CNN’s), known for their ability to understand complex patterns in images, play a key role in the design of the algorithm. These networks, like the pathways in the human brain, can recognize and understand the unique characteristics and subtleties of Tulu Lipi from the carefully curated dataset. The model, thus trained, evolves into a sophisticated tool capable of not only recognizing but also understanding the intricate subtleties of

Tulu script across a spectrum of handwriting styles. This model aims to make a common person such as a traveler easy to comprehend and understand the language thereby exposing this language to the larger community. This project initially focuses on building a trainable model that can detect and automatically predict digitally typed text into its appropriate English transcript. Post completion of developing the model the model will be updated to analyse a deeper array of handwritten texts and eventually analyse handwritten texts anddetect correct predictions.

II. LITERATURE REVIEW

A. Review on Methods of Script Identification for Printed and Handwritten Documents [1]

The paper titled "2019 Innovations in Power and Advanced Computing Technologies (i-PACT): A Review on Methods of Script Identification for Printed and Handwritten Documents" is authored by Aditi Gaygole from the Department of Electronics Engineering at Government College of Engineering, Amravati, India. It addresses the significance of Optical Character Recognition (OCR) in the digitization of documents. The challenge of OCR being script-specific prompts the need for an algorithm capable of recognizing characters irrespective of the script. The paper categorizes methods based on document format—printed or handwritten. The methodology involves initially identifying the script of a document and selecting the appropriate OCR module accordingly. It emphasizes the complexity of script identification for handwritten documents, given the individualistic writing styles. The paper reviews existing work in the field, classifying it by document type and further categorizing papers based on the methods used for script identification. This classification aids in comparing different approaches and selecting suitable methods for future work in script identification. Additionally, the paper briefly mentions hardware requirements and introduces the importance of digitizing documents for preserving ancient texts and facilitating efficient document analysis. The specified hardware requirement, "Template Matching," is briefly described as a method introduced in 1997 by Lila Kerns, Timothy Thomas, Patrick Kelly, and Judith Hochberg for the automatic script identification of image documents using templates based on clusters.

B. Script and Language Identification for Document Images and Scene Texts [2]

The research explores the application of Recurrent Neural Networks (RNNs) for script and language identification at the word and line levels, focusing on multilingual settings. The proposed RNN model is designed to learn the distribution of feature vectors and demonstrate effectiveness in identifying scripts and languages directly from document images. The study utilizes a dataset of 15 scripts and languages, comprising nearly 15.03 million words from 55,000 document images. The RNN-based approach achieves significant accuracy in script and language identification, outperforming some previous methods. The investigation distinguishes itself by addressing higher-level tasks without explicit recognition, such as identifying topic models or categorizing documents. The method's simplicity, efficiency, and accuracy are high- lighted, emphasizing its potential for integrated solutions in multilingual recognition settings. Comparative analyses with existing methods, including Gabor features with SVM, show competitive or superior results. The study also introduces the concept of script separation in a Multilingual Optical Character Recognition (M-OCR) pipeline, demonstrating the importance of identifying the script before sending the word to the corresponding OCR engine. The research contributes to advancing language and script identification in the field of computer vision, showcasing the potential of RNNs for such tasks.

C. Script Identification: A Review [3]

The abstract provides an overview of script identification in a multilingual context, emphasizing its importance in diverse applications such as text filtering, automatic translation, OCR, and text location identification. The paper explores various script identification methods, including Scene Text Script Identification with Convolutional Recurrent Neural Networks, Sequence-to-label script identification for multilingual OCR, attention-based Convolutional LSTM network, convolutional triplets for script identification in scene text, Discriminative Convolutional Neural Network, script identification method based on hand-crafted texture features and an artificial neural network, fully differentiable method for multilanguage scene text localization and recognition, and integrating local CNN and global CNN. The introduction discusses the complexity of human languages' origin and the challenges in studying them. The paper categorizes writing systems into alphabets, syllabaries, and logography. Script identification is introduced as a crucial task, particularly for multilingual programming. The detailed discussion of various script identification systems includes their architectures, methodologies, and applications. The datasets section categorizes datasets into synthetic and realistic, emphasizing the significance of data in training deep learning models. The applications of script identification in various fields are outlined. In conclusion, the paper provides a comprehensive review of script identification methods, datasets, and applications, offering insights into the challengesand advancements in this domain.

D. A Classification of Script Identification Systems [4]

The paper titled "A Classification of Script Identification Systems" presents a comprehensive detail on script identification, a vital aspect of document image analysis. Script identification involves determining the script or scripts used in a document image, playing a crucial role in selecting appropriate optical character recognition (OCR) systems, language identification, and document classification. The granularity of script identification varies across different levels, ranging from document to character level, depending on specific applications and data availability. Script identification poses significant challenges attributed to the diversity and complexity of scripts, variability, and noise in document images, and a lack of standard datasets and evaluation metrics. Despite these challenges, script identification finds applications across various domains, including digital libraries, multilingual document processing, postal services, security, education, and cultural heritage preservation. Methods for script identification can be classified based on various criteria, such as the method of acquisition (offline or online), method of writing (handwritten or typeset), type of script (domestic or foreign), feature used (local or global), number of scripts (unilingual, bilingual, or multilingual), and technique used (spatial-domain or frequency-domain). Each criterion carries its own set of advantages and disadvantages, and the choice of the best method depends on the specific problem and data characteristics. Recent advancements in script identification have witnessed the application of deep learning techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms. These techniques have demonstrated superior performance and robustness compared to traditional methods. Future directions in script identification research involve the development of more comprehensive and representative datasets, exploration of cross-modal and multi-task learning, incorporation of linguistic and semantic information, and evaluating the impact of script identification on downstream tasks.

E. Script and Language Identification for Document Images and Scene Texts [5]

The research explores the application of Recurrent Neural Networks (RNNs) for script and language identification at the word and line levels, focusing on multilingual settings. The proposed RNN model is designed to learn the distribution of feature vectors and demonstrate effectiveness in identifying scripts and languages directly from document images. The study utilizes a dataset of 15 scripts and languages, comprising nearly 15.03 million words from 55,000 document images. The RNN-based approach achieves significant accuracy in script and language identification, outperforming some previous methods. The investigation distinguishes itself by addressing higher-level tasks without explicit recognition, such as identifying topic models or categorizing documents. The method's simplicity, efficiency, and accuracy are high- lighted, emphasizing its potential for integrated solutions in multilingual recognition settings. Comparative analyses with existing methods, including Gabor features with SVM, show competitive or superior results. The study also introduces the concept of script separation in a Multilingual Optical Character Recognition (M-OCR) pipeline, demonstrating the importance of identifying the script before sending the word to the corresponding OCR engine. The research contributes to advancing language and script identification in the field of computer vision, showcasing the potential of RNNs for such tasks.

F. Text Detection and Script Identification in Natural Scene Images Using Deep Learning [6]

The paper "Text Detection and Script Identification in Natu- ral Scene Images using Deep Learning" by Khalil et al. introduces fresh approaches for tackling the complex challenges of recognizing text in natural scene images. The authors present two methods, Multi-Channel Mask (MCM) and Multi-Channel Segmentation (MCS), both leveraging advanced convolutional networks. The goal is to not only detect text but also identify the script it belongs to, crucial in handling diverse fonts, colors, sizes, orientations, and languages in real-world images filled with noise and variations. The authors claim superiority, particularly in recall, over existing methods through rigorous evaluations on datasets like ICDAR MLT 2017, MLe2e, and Arabic–Latin, where they employ techniques like under sampling and oversampling to address imbalanced data issues. The document delves into the broader field of optical character recognition (OCR), stressing its applications in autonomous vehicles, robots, and drones. It provides a comprehensive re- view of existing methodologies, categorizing them into region- based and segmentation-based approaches for text detection and distinguishing between patch-based and end-to-end methods for script identification. The proposed MCM and MCS methods share a common feature map but differ in their prediction phases, offering a nuanced understanding of the joint challenges of text detection and script identification. The authors employ well-defined evaluation metrics like F- score, precision, and recall, comparing their outcomes with other state-of-the-art methods to support their claim of superior performance. The paper not only contributes novel techniques to the OCR research arena but also critically reflects on limitations, such as overlapping boxes and annotation errors, providing insights into potential areas for improvement. Its clear structure, comprehensive literature review, and insightful discussions make it a valuable resource for both researchers and practitioners engaged in optical character recognition and its diverse applications in the real world.

G. Identification of Telugu Script in a Bilingual [7]

In recent years, the field of script identification has seen increased research efforts due to the growing demand for accurate Optical Character Recognition (OCR) systems, particularly in multilingual contexts like India. This study, focusing on identifying Telugu and English scripts within bilingual documents, explores various methodologies and findings. It highlights the significance of script identification in India's diverse linguistic landscape, where many languages and scripts coexist. Telugu, an official language in South Indian states, is the specific focus of investigation. It addresses the challenge of designing a universal recognizer capable of distinguishing between different scripts and languages. The importance of identifying the language region before applying OCR is emphasized, forming the foundation of the proposed methodology. It reveals previous work in script identification for various languages, including

Chinese and Indian languages like Hindi, Bangla, Kannada, Tamil, and Telugu. Methodologies in these studies range from multi-channel filters to statistical techniques. In the context of Indian languages, the document discusses specific studies on separating text lines, using Gabor filters, and employing probabilistic neural networks. The proposed Telugu script identification methodology is presented, involving the extraction of visual features like top and bottom profiles, tick components, holes, and vertical lines. Two identification methods, a heuristic approach and a K-Nearest Neighbor (K-NN) method, become the focal points of innovation. As a result, document dissects the outcomes of applying the proposed methodology to bilingual document images containing Telugu and English text. The accuracy of the methods is evaluated, shedding light on the effectiveness of the heuristic approach and the K-NN method. Amid technical intricacies, the document also highlights the limited work on Telugu language document image analysis, positioning the current study as a pioneering venture into Telugu script identification in bilingual environments. This paper encapsulates a comprehensive exploration of script identification techniques, emphasizing the uniqueness and challenges in discerning the Telugu script within bilingual contexts. The research delves into the intricacies of the Telugu script, how it differs from other scripts, and the difficulties in identifying it when it is used alongside other scripts.

H. Bilingual Script Identification of Printed Text Image [8]

Significant strides have been made in the realm of optical character recognition (OCR) in recent times. However, the task of identifying scripts in bilingual or multilingual text is still considered a considerable challenge. Our exploration involves the examination of current methodologies and the introduction of a fresh approach to the identification of English and Punjabi scripts at the line level, with special attention given to headline and character density features. The substantial achievements in OCR for monolingual printed and handwritten text images are acknowledged at the beginning of our journey. However, the surge in bilingual and multilingual contexts necessitates the distinction between scripts before applying them to individual OCR systems. The complexity of the task is acknowledged, and the need for script identification techniques at various levels—paragraph, line, and word is emphasized. Existing methods dealing with European and oriental scripts, as well as specific languages like English, Hindi, and Kannada, are examined. Contributions from Ambekar et al., Jindal and Hemrajani, Prakash et al., Gupta et al., and Mohanty and Bebartta showcase a diverse range of script identification approaches. However, the proposed work arises from the limitations of these methods in handling bilingual or multilingual text images. The proposed methodology focuses on the identification of English and Punjabi scripts using headline and character density features. The system is trained with specific fonts for each script, achieving commendable accuracy across various font sizes. The step-by-step process includes image binarization, line segmentation, script identification, word and character segmentation, feature extraction, and classification. Noteworthy features such as histogram projection profiles and the number of holes is employed for efficient character classification. Results from extensive experiments showcase high script identification accuracy for both English and Punjabi scripts. The successful implementation of our proposed method is emphasized in the conclusion, and future research directions are suggested. The future scope involves extending the system to recognize other language scripts, adapting the approach for handwritten text, and making further improvements in accuracy and efficiency. Our exploration underscores the evolving nature of OCR systems and ongoing efforts to enhance their capabilities in handling diverse linguistic contexts.

I. Handwritten Character Recognition of Modi Script using Convolutional Neural Network Based Feature Extraction Method and Support Vector Machine Classifier [9]

The paper by Solley Joseph and Jossy George on “Hand written Character Recognition of MODI Script using Convolutional Neural Network Based Feature Extraction Method and Support Vector Machine Classifier” presents a compelling and innovative approach to addressing the challenge of recognizing characters in the ancient MODI script. The authors introduce a two-step methodology, utilizing a Convolutional Neural Network (CNN) autoencoder for feature extraction and a Support Vector Machine (SVM) for classification, achieving a noteworthy accuracy of 99.3%. The incorporation of on-the- fly data augmentation adds robustness to the dataset. While the paper effectively highlights its strengths, including a clear presentation and high accuracy, it could benefit from a more detailed comparison with existing methods, a thorough explanation of the CNN autoencoder architecture, and insights into failure cases. Nevertheless, this research significantly advancesthe field of handwritten character recognition.

J. PNN and Deep Learning Based Character Recognition System for TULU Manuscripts [10]

This paper presents a groundbreaking approach to offline handwritten Tulu character recognition, focusing on degraded images extracted from palm leaf manuscripts and paper documents. The research employs a multifaceted methodology integrating adaptive thresholding, noise removal, skeletonization, zone-wise gradient direction values, wavelet transform, and the synergy of probabilistic neural network (PNN) and deep convolutional neural network (Deep CNN) models for preprocessing, feature extraction, and classification. Results indicate exceptional recognition efficiencies, with the Deep CNN model achieving 97.05% for Tulu characters from paper documents, 98.12% for Tulu numerals, and 88.07% for Tulu palm leaf characters. Comparative analyses highlight the system’s superiority over existing methods, establishing its effectiveness in preserving and digitizing ancient Tulu manuscripts. The paper concludes by outlining promising future directions for advancing the proposed recognition system. This research significantly contributes to the field of handwritten character recognition, providing a robust solution for the challengesposed by degraded Tulu images.

III. CONCLUSION

In conclusion, the project on Tulu Lipi identification represents a significant step toward preserving and understanding the Tulu script using advanced computational methods. The utilization of a convolutional neural network (CNN) implemented in Pytorch demonstrates the potential for automated recognition of Tulu characters. The dataset is carefully pre- processed, incorporating data augmentation techniques for the training set to enhance model generalization. The training process involves monitoring training and validation losses over epochs, with the best-performing model saved for future use.

The evaluation phase assesses the model's performance on a separate test set, providing insights into its generalization capabilities. The inference code demonstrates the application of the pretrained model to make predictions on new Tulu script images, showcasing the robustness of the developed system. Visualizations of test images and their predicted labels serve as a transparent means to verify the correctness of the preprocessing steps and provide a tangible view of the model'srecognition capabilities.

This project not only contributes to the technological advancement in handwritten character recognition but also has broader implications for cultural preservation and linguistic research. By bridging traditional scripts like Tulu Lipi with modern computational methods, the project opens avenues for the automated analysis and understanding of historical scripts. The combination of deep learning techniques and meticulous data processing highlights the potential of AI in the field of script recognition and promotes the integration of technology in preserving linguistic diversity. Overall, the project on Tulu Lipi identification stands as a testament to the intersection of heritage preservation and cutting-edge technology.

REFERENCES

1. Gaygole, Aditi Rojatkar, Dinesh. (2019). A Review on Methods of Script Identification for Printed and Handwritten Documents. 1-6. 10.1109/i-PACT44901.2019.8960184.
2. Jawahar, C.V. (2015). Script and Language Identification for Document Images and Scene Texts.
3. Bhunia, A.K., Konwer, A., Bhowmick, A., Bhunia, A.K., Roy, P.P., Pal, U. (2018). Script Identification in Natural Scene Image and Video Frame using Attention based Convolutional-LSTM Network. ArXiv, abs/1801.00470.
4. Rumaan Bashir, Kaiser J. Giri, Javaid Iqbal Bhat. (2016). A Clas- sification Of Script Identification Systems at International Journal of Engineering Sciences, Issue June, Vol. 18.
5. Manimozhi and M. challa, "An Efficient Translation of Tulu to Kannada South Indian Scripts using Optical Character Recognition," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 952- 957, doi: 10.1109/ICCMC51019.2021.9418225.
6. Science, Computer Jarrah, Moath Al-Ayyoub, Mahmoud Jararweh, Yaser. (2021). Text detection and script identification in natural scene images using deep learning. Computers Electrical Engineering. 91. 107043. 10.1016/j.compeleceng.2021.107043.
7. Banoth, R. Dhir, R. (2016). Identification of Telugu script in a bilingual document image. International Journal of Scientific Progress and Research, 20(2), 107-113. 1
8. Kaur, I., Mahajan, S. (2015). Bilingual Script Identification of Printed Text Image.
9. S. Joseph and J. George, "Handwritten Character Recognition of MODI Script using Convolutional Neural Network Based Feature Extraction Method and Support Vector Machine Classifier," 2020 IEEE 5th Inter- national Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2020, pp. 32-36,
doi: 10.1109/ICSIP49896.2020.9339435.
10. Savitha, C., Antony, P.J. (2019). PNN and Deep Learning Based Character Recognition System for Tulu Manuscripts 1855.

CITATION
Midhun V., Suraksha, Abhinav V., Suhail A.N., & Anoop B K. (2024). A Comprehensive Review on BARAVU-Tulu Lipi Identification. In Global Journal of Research in Engineering & Computer Sciences (Vol. 4, Number 2, pp. 105–109). https://doi.org/10.5281/zenodo.10969139