Segmentation-Free Character Recognition System

Shreya Mittal
Dept. of Computer Science Engineering.
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22153@bl.students.amrita.edu

Santhwana Santhosh Kumar
Dept. of Computer Science Engineering.
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22149@bl.students.amrita.edu

Satwika Chowdary
Dept. of Computer Science Engineering.
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse22139@bl.students.amrita.edu

***Abstract*—Character recognition systems for complex scripts like Malayalam face challenging problems due to their complicated structure and variety of styles. The varied morphological features of these scripts are often beyond the capacity of traditional segmentation-based techniques. In order to address the issue, we propose a segmentation-free character recognition system in which the input image is first divided into blocks and further segmented into three zones. Features are subsequently extracted in each zone and modeled by HMMs, which permits reliable identification without requiring precise character segmentation. The recognised text is then translated into English to improve its readability for a non-native speaker. Our system performs well on the translation of recognised text and also ensures outstanding recognition accuracy with significant improvements over traditional methods.**

***Index Terms*—Segmentation-Free Character Recognition, Handwritten Document Digitization, Hidden Markov Model (HMM), Feature Extraction, Historical Text Recognition**

I. Introduction

Digitization of historical manuscripts, especially those handwritten in ancient scripts, is one of the most emerging fields of study in digital humanities and computational linguistics. Such documents are usually preserved on delicate materials like palm leaves and are very valuable both historically and culturally.Their handcrafted quality combined with the passage of time made it relatively hard to decipher and translate using traditional techniques. Digitalization of these manuscripts not only gives a guarantee for the safety of their content for future generations but also creates new opportunities for deeper analysis and further research.

During these years, the methods used to digitise handwritten documents have changed enormously. Earlier approaches mostly depended on segmentation-based techniques, where a machine-learning model first extracts individual words or characters from the document before recognising them. Though these methods have turned out to be successful in some applications, they can be difficult to use with complex or deteriorated scripts where segmentation is difficult. Particularly when dealing with old scripts, the step of correctly separating characters from linked or overlapping strokes frequently results in mistakes that spread throughout the recognition process, reducing the overall accuracy.

Due to the fact that segmentation-based systems have their own drawbacks, methods of segmentation-free recognition have been investigated. The following methods avoid character segmentation but focus on recognizing the whole sequences or zones of text. By doing so, they are in a better position to handle irregularities and intricacies of handwritten scripts found mainly in ancient documents. Methods avoiding segmentation are very promising in terms of contextual integrity of the text, often lost when characters are isolated.

There exist a lot of different approaches to segmentation-free recognition and each has certain strengths and challenges. To model and recognise text sequences directly from photos, methods including Hidden Markov Models (HMM), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN) have been used. These models are designed to be trained on patterns and features across different zones of a document image so that more accurate recognition of complex scripts can be obtained. The choice of the approach often depends on the unique qualities of the text and the goals of the digitization project.

In this paper, we propose a segmentation-free character recognition system specially designed for the digitization of ancient handmade documents on palm leaves. Document images will be divided into blocks, which are then further divided into three zones for the extraction of features, modeled using an HMM. This approach overcomes most of the problems due to traditional segmentation-based methods while improving the accuracy of identifying complex handwritten scripts. This system is quite a development in this domain and offers a more reliable solution for the preservation and translation of historical texts into modern languages like English.

The rest of the paper is organized as follows: Section 2 reviews related work; Section 3 details the methodology; Section 4 presents result; Section 5 details analysis; and

Section 6 concludes with future directions.

II. LITERATURE SURVEY

Tong-Hua Su et al. proposed a methodology that acknowledges the problems inherent to the Chinese handwriting recognition [1] as a domain. Another challenge that relate to this is the problem of segmentation where characters overlap or come close to each other as observed in the realistic handwriting. In this context, the authors have surveyed this literature and observed that most methods devised to solve the segmentation problem have aimed at enhancing the segmentation algorithms.

Najwa Altwaijry at el. in their proposed methodology has shown the growing interest of Arabic handwriting recognition due to the unique challenges posed by the Arabic script [2], and that includes its cursive writing nature, context-sensitive shapes of their alphabets, and a wide range of handwriting styles.

"Segmentation-Free Optical Character Recognition for Printed Urdu Text" by Ud Din, et al., describes several techniques and methodologies used in OCR for the Urdu language [3]. The survey provides a a wide range of various methodologies and their suitability in the identification of printed Urdu text.

The method proposed by Parveen Kumar, and Ambalika Sharmaa is called SEGmentation-free Writer Identification (SEG-WI) which is based on Convolutional Neural Network (CNN) free from text segmentation. [4] This model implies the region probability map from unsegmented documents and the decision based on the voting on the top 10% to 50% of the regions. Evaluation of the proposed technique in the IAM, CVL, IFN/ENIT, Kannada, and Devanagari datasets provides better results compared to existing techniques with top-1 accuracy lying consistently in the range of 87. 06% to 100%.

The proposed methodology by Ge Peng focuses on the issues related to the offline handwritten character recognition regarding application of segmentation methods, problems linked to interleaving and touching characters in handwritten manuscripts [5]. It proposes a segmentation-free recognition algorithm leveraging a deep learning network comprising four layers: The first steps include image preprocessing, the use of convolutional neural networks known as CNNs for feature extraction, the second is sequence prediction by bidirectional long short term memory also known as BDLSTM, and finally for the text sequence alignment and classification, connectionist temporal classification also known as CTC. Accompanying the proposed classification approach is also a new method for data length equalizing the given data sets.

The methodology proposed by Fanfei Meng at el., explores the use of AI and machine learning for text recognition and identifies corresponding solutions to the pertaining issues that affect model performance and accuracy [6]. It covers some of the essential topics including data quality and the handling of diversity, scaling up training and inference of large models, handling of multiple languages and font styles, changes in text formatting, HTR, and model explanation. Solutions such as data cleaning, augmentation, and distributed training frameworks are suggested to increase the model's accuracy and stability. On these aspects, the methodology will primarily contribute to the efficiency in the performance, particularly of the recognition of texts, where advancements of accuracy and robustness are to be demonstrated.

The authors Jyoti Ghosh et al., aims at solving the problem of directly applying scene text recognition models for mobile applications, where the latter are constrained in terms of their size [7]. This method puts forward a light-weight Convolutional-Recurrent Neural Network (CRNN) model that can satisfy the demands both in efficiency and accuracy. By substituting the Convolution network with MobileNetV2, the model greatly varies in size concerning parameters and the file size from 1M parameters and 12MB to 0. 5M parameters and 6MB for the file, all the while preserving a similar level of correctness. Text detection and recognition based on the proposed model are then assessed on ICDAR 2013, IIIT 5K, and Total-Text datasets.

The methodology proposed by Salvador Espan˜a-Boquera at el., describes the methodology of recognizing unconstrained offline handwritten texts using hybrid Hidden Markov Model (HMM) and Artificial Neural Network (ANN) models. [8] In this instance, the optical models' structure utilizes the Markov chains, while the emission probability can be estimated with the help of a Multilayer Perceptron (MLP). This work presents new approaches to correction of slopes and removing slants and normalization of sizes of the text images by applying supervised learning techniques.

The methodology propsoed by Tien Do at el., focuses on the issues of scene text detection and recognition, while considering the case when texts are located in complex scenes [9], with different fonts, sizes, styles, languages, and large amount of noises, cluttered backgrounds, etc. The methodology presents the new dataset of in-the-wild signboard images, with 79K text instances at line levels and 79K instances at word levels with the number of images 2104. Consequently, the collected dataset is employed to comprehensively assess the performances of the most advanced/recent SOTA methods in the context of text detection and recognition.

The methodology proposed by Tao Wang at el., addresses the problem of end-to-end text recognition in natural images by harnessing the power of big, deep neural networks combined with recent advances in unsupervised feature learning [**?**]. Unlike conventional systems, this method trains an accurate text detector and character recognizer modules within one framework without the use of hand-engineered features or extensive prior knowledge. These modules are then composed into a full end-to-end, lexicon-driven scene text recognition system using simple off-the-shelf methods. Very promising results of the proposed system on standard benchmarks like Street View Text and ICDAR 2003 datasets prove to be effective and robust.

The authors of the proposed methodology "Evaluating Sequence-to-Sequence Models for Handwritten Text Recognition," Johannes Michael, et al, provide a literature review

of the existing approaches to handwritten text recognition. It mainly discusses sequence-to-sequence model, and contrast it with other types of methods like Hidden Markov Model, [11] as well as the combination of Hidden Markov Model and neural networks. The survey under consideration focuses on the changes from the HMM-based approaches to the more neural-network based ones, stressing the added accuracies and resistance.

The methodology proposed by Batuhan Balci, Dan Saadati, and Dan Shiferaw is the Handwritten Text Recognition using Deep Learning, aims at different methods and techniques to recognize handwritten individual words and convert handwritten text into digital form in this paper [12]. The authors utilized two primary methods: The first method for word features extraction is direct word classification based on Convolutional Neural Networks (CNN); and the second one is based on character segmentation using Long Short Term Memory (LSTM) networks connected with CNN.

The methodology proposed by the authors aims at improving Offline Handwritten Text Recognition (HTR) systems and discussing ways to incorporate HTR into the large-scale Multilingual OCR systems. [13] Some of the main areas highlighted are the issues such as the lack of high quality training images; the online handwriting database is exploited to improve model performance, even when the real images are scarce. The literature survey looks at the evolution from LSTM basic models into a more efficient Neural Network model without Recurrent Links and therefore with higher accuracy and training/imputation parallelism.

The methodology proposed by Gyeonghwan Kim et al. describes an end-to-end method of recognizing text in handwritten page images, which is conceived of as a combination of five functional modules [14]. Some of the literature survey has intensively discussed the improvement on pre-processing of images for the convenient image processing, detection and extraction of text lines and finally smart methods of word segmentation. It also addresses a description of several handwritten word recognition algorithms and the use of languages constraint for text parsing/word recognition.

III. Dataset

The dataset used in this study contains 2,250 samples of the lines extracted from ancient handwritten documents on palm leaves. The document images were preprocessed and segmented into blocks with an average of about 25 blocks per line. Then features were precisely captured from each block, to identify the fine details of the handwritten script.

There are a total of 196 different features describing a wide array of textual and structural characteristics for each sample from the segmented zones of the document. These features will be used to feed our segmentation-free Hidden Markov Model character recognition system. The dataset thoroughly depicts ancient scripts and gives freedom to the model to learn and recognize characters, therefore translating them into English with high accuracy.

IV. Methodology

The approach of the given character recognition system that is free from the segmentation process includes the following steps: Applying the pre-processing on the input image for the character recognition and feature extraction of the characters and then recognizing them by the aid of Hidden Markov Models (HMM). The aim is to bring eye approach strategy, which might be useful to address scripts as of Malayalam and the newly discovered much older scripts. It is relatively stable, and does not crumble in the face of problems that threaten traditional segmentation based approaches; it also offers greater accuracy in recognition.

A. Input Image Acquisition

1) Document Scanning: This is done through getting picture images of manuscripts belonging to the past regimes or even past civilization. These are either scanned or photographed with much careful to make the following of script as clear on the scanned or photographed paper as possible.

2) Preprocessing: This include making the scanned image to be neat by using functions such as the denoise, the contrast and the thresholding. The goal is to produce a clean image in which all "noise" and interferences that could interfere with further processing of the image have been minimized or removed completely.

B. Image Segmentation into Blocks

1) Block Division: The received preprocessed image will be subdivided or partitioned into smaller image pieces better for convenient processing. This step is very crucial for simplification of the recognition process because every block as discussed above and seen from the context below will contain the sequence of characters which can further be treated as another element.

2) Zone Segmentation: Each block in question is further divided in to the three vertical strips namely the top strip and the bottom strip and the strip in the middle. They are based on the usual segmental formation of the Malayalam script in which segments of a character are placed in the the mid zone (consonant), or the top zone or the bottom zone which contains vowels modifying mark.

C. Feature Extraction

Zone-wise Feature Extraction: For each zone for a block certain morphological features are extracted. These features may include:

1) Top Zone: In particular, characters that are displayed in a manner that would indicate they are features such as diacritics or character modifiers.

2) Middle Zone: The properties of the main characters' body and their constitution.

3) Bottom Zone: Even more diacritical marks or tail strokes. Statistical and Structural Features: Edge detection, stroke width, contours shapes and pixel distribution of these features are then calculated within each zone. The features are thought to reflect the individual characteristics of the script's morphology.

D. Modeling Using Hidden Markov Models (HMMs)

1) HMM Training: The extracted features from each zone are used to train Hidden Markov Models. The HMMs are specifically designed to model the sequential nature of the script, allowing for recognition even when characters are connected or overlapping.

2) Sequence Recognition: During recognition, the HMMs evaluate the likelihood of different character sequences based on the observed features from the zones. The model outputs the most probable sequence of characters for each block.

E. Post-Processing and Translation

1) Text Reconstruction: The recognized characters from each block are summed up and then the sequences of the entire full texts are recognized. Any mistake made by the model is rectified with reference to the context it was generated from and with reference to the language models.

2) Translation into English: The recognized text is translated from Malayalam to English using a translation model.This step ensures that the text is accessible to non-native speakers, improving its readability and usability.

F. Validation and Accuracy Measurement

1) Performance Evaluation: The system's performance is evaluated using a dataset of manually transcribed documents. Key metrics include recognition accuracy, translation accuracy, and processing speed.

2) Comparison with Traditional Methods: The results are compared against traditional segmentation-based recognition methods to highlight the improvements in accuracy and reliability.

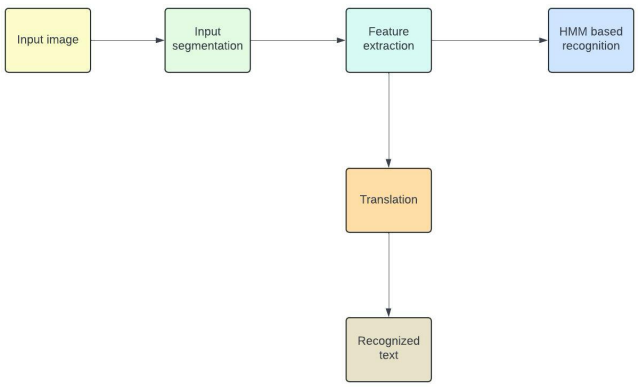

Fig. 1. Flow Diagram

V. Results, Analysis and Discussion

A. Class Separation

The classes in the dataset features are not perfectly separated and hence overlapping. This can be due to the complexity involved in the hand-written scripts and probably their degradation due to the passage of time. It is observable that the classifier gets confused while differentiating a class from another, leading to misclassification.

B. Behaviour of kNN with increasing k

As k increases in the kNN classifier, the model becomes less sensitive to noise, reducing the risk of overfitting. However, large values of k have the risk of having a model that underfits and smoothes out important differences between classes. Optimal performance was observed with moderate values of k and extremes gave overfitting or underfitting.

C. kNN Classifier Performance

kNN classifier performed reasonably well, but not outstandingly. Frequently it was correct with the right k, though its strong reliance on proximity in feature space can be limiting for data sets with overlapping class boundaries. Seeing metrics like accuracy and F1-score, this classifier is not necessarily among the best options against this dataset.

D. Model Fit Analysis

The model exhibited a regular fit, with moderate values of k, where train and test set performance are alike. The lower the values of k, the greater would be the overfitting to data; that is, accuracy would increase on the training set compared to the test set. Similarly, for higher values of k, the model would underfit.

E. Overfitting in kNN

Overfitting occurred at very low k values, such as $k = 1$, where the model became too sensitive to training data, including noise. This resulted in high accuracy during training but poor generalization on the test set. Increasing k reduced overfitting and achieved a better balance in model performance.

References

[1] Tong-Hua Su, Tian-Wen Zhang, De-Jun Guan, Hu-Jie Huang, *"Off-line recognition of realistic Chinese handwriting using segmentation-free strategy"* School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, PR China

[2] Najwa Altwaijry· Isra Al-Turaiki, *"Arabic handwriting recognition system using convolutional neural network"*, eceived: 11 October 2019 / Accepted: 3 June 2020 / Published online: 28 June 2020

[3] Israr Ud Din, Imran Siddiqi, Shehzad Khalid and Tahir Azam *"Segmentation-free optical character recognition for printed Urdu text"*

[4] Parveen Kumar, Ambalika Sharmaa "Segmentation-free writer identification based on convolutional neural network", a Department of Electrical Engineering, Indian Institute of Technology Roorkee, India b Department of CSE, National Institute of Technology Uttarakhand, India

[5] Ge Peng *"Segmentation-Free Recognition Algorithm Based on Deep Learning for Handwritten Text Image"*, School of Big Data, Baoshan University, Baoshan, Yunnan 678000, China (Received 22 October 2023; Revised 15 February 2024; Accepted 15 February 2024; Published online 05 March 2024)

[6] Fanfei Meng* , Branden Ghena *"Research on Text Recognition Methods Based on Artificial Intelligence and Machine Learning"* , . Advances in Computer and Communication, 4(5), 340- 344. DOI: 10.26855/acc.2023.10.014

[7] Jyoti Ghosh · Anjan Kumar Talukdar· Kandarpa Kumar Sarma *"A lightweight natural scene text detection and recognition system"* Received: 20 September 2021 / Revised: 15 January 2022 / Accepted: 23 April 2023 / Published online: 13 June 2023

[8] Salvador Espan˜a-Boquera, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez *"Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models"* , IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 33, NO. 4, APRIL 2011

[9] TIEN DO, THUYEN TRAN, THUA NGUYEN, DUY-DINH LE , (Member, IEEE), AND THANH DUC NGO *"SignboardText: Text Detection and Recognition in In-the-Wild Signboard Images"* , University of Information Technology, Ho Chi Minh City, Vietnam Vietnam National University, Ho Chi Minh City, Vietnam Corresponding author: Thanh Duc Ngo (thanhnd@uit.edu.vn) This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2021-26-02.

[10] Tao Wang David J. Wu Adam Coates Andrew Y. Ng , *"End-to-End Text Recognition with Convolutional Neural Networks"* , Stanford University, 353 Serra Mall, Stanford, CA 94305 {twangcat, dwu4, acoates, ang}@cs.stanford.edu, 21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan

[11] Johannes Michael, Roger Labahn ,Tobias Gruning, Jochen Z ¨ ollner , *" Evaluating Sequence-to-Sequence Models for Handwritten Text Recognition "* Computational Intelligence Technology Lab University of Rostock 18057 Rostock, Germany {johannes.michael,roger.labahn}@uni-rostock.de , , PLANET artificial intelligence GmbH Warnowufer 60 18057 Rostock, Germany {tobias.gruening,jochen.zoellner}@planet.de

[12] Batuhan Balci , Dan Saadati , Dan Shiferaw,*"Handwritten Text Recognition using Deep Learning"* bbalci@stanford.edu dans2@stanford.edu shiferaw@stanford.edu

[13] R. Reeve Ingle, Yasuhisa Fujii, Thomas Deselaers, Jonathan Baccash, Ashok C. Popat *"A Scalable Handwritten Text Recognition System"*, Google Research Mountain View, CA 94043, USA Email: {reeveingle,yasuhisaf,deselaers,jbaccash,popat}@google.com , 2019 International Conference on Document Analysis and Recognition (ICDAR)

[14] Gyeonghwan Kim1, Venu Govindaraju2, Sargur N. Srihari2 *"An architecture for handwritten text recognition systems"* 1 Department of Electronic Engineering, Sogang University, CPO Box 1142, Seoul 100-611, Korea; e-mail: gkim@ccs.sogang.ac.kr 2 CEDAR, State University of New York at Buffalo, 520 Lee Entrance, Amherst, NY 14228–2567, USA Received October 30, 1998 / Revised January 15, 1999