Segmentation-Free Character Recognition System

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

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

I. Introduction

Character recognition with complex scripts like Malayalam is tricky due to the complicated structures of the characters and their contextual dependencies. Traditional approaches for segmentation result in inaccuracies with overlapping and touching characters which make digitization harder and retard the access of regional languages. To address this, we discuss segmentation-free recognition systems as a potential solution. The approach is to divide the image into blocks, segment them into three zones, and then extract the features and model them using Hidden Markov Models. The system would help in identifying Malayalam text and translate it into English which will not only enhance the digital accessibility but also the process of communication. In our work, novelty in the zone-based feature extraction and HMM modeling has been introduced, and that will improve the recognition accuracy to a greater extent. The effort also goes inline with the UN SDG, particularly in promoting an inclusive and equitable quality education and supporting the linguistic diversity.

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 local extrema of text contours are classified with MLPs for the slope correction and the size normalization, whereas the slant is adjusted nonuniformly with ANNs. These methods are evaluated utilizing the IAM database and they prove to yield recognition rate that is equal to the highest in literature.

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.

III. METHODOLOGY

A. Data Portrayal

Within the application of client information division for the IRCTC stage, we start by collecting information from clients collaboration with the site. This information incorporates different properties such as client socioeconomics, booking history, and inclinations. Once collected, the information experiences pre-processing, where it is changed over into a lucid organize, such as CSV records, to encourage investigation. Amid this pre-processing organize, we carefully extricate the important highlights that will be valuable for our investigation whereas disposing of any superfluous data. This step is vital to guarantee that our demonstrate centers on the most critical information focuses. After highlight extraction, the information is encoded to get ready it for input into the division show. Following, we perform similitude scoring, which permits us to degree the degree of likeness between diverse clients based on their characteristics and behaviors. This scoring makes a difference in classifying the information into unmistakable classes and names, viably gathering clients with comparative characteristics. By visualizing these trends and designs, ready to pick up profitable bits of knowledge into client behavior on the IRCTC stage, eventually upgrading the in general client encounter and advising focused on promoting strategies.

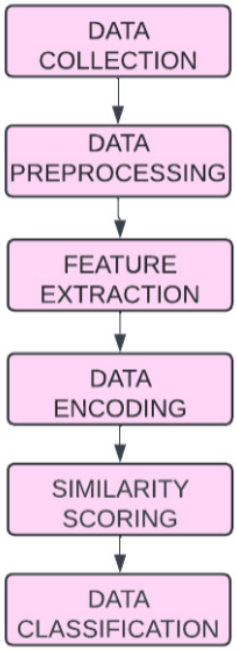

Fig. 1. Data Flow

B. Framework Portrayal

The client inputs his data within the site utilizing the client interface given by IRCTC. The back-end benefit of the application stores the data given by the client in a capacity unit. The back-end moreover forms a few demands by the client and within the handle stores extra data on the client. We get this information from the back-end utilizing API call focuses or utilizing information recovery. The information gotten goes through the information stream prepare. The prepared information is utilized by the division show to classify into different classes and labels. these patterns and designs, we are able pick up profitable experiences into client behavior on the IRCTC stage, eventually improving the by and large client encounter and educating focused on promoting procedures.

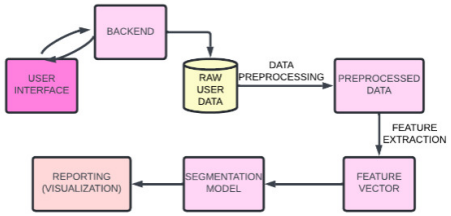

Fig. 2. System architecture

C. Define Parameters

Parameters are values extracted from data to define its characteristics. When analyzing user data from IRCTC, parameters such as'mean price','sample mean', 'probability of making a loss', and 'conditional probability' are used to extract insights from the data.

IV. Results, Analysis and Discussion

A. The Importance of the Rank of an Observation Matrix in the Model Building for Classification

The rank of an observation matrix is quite important in the classification because it shows how much independent information enters the learning stage. It is necessary to make a better difference in the classes using the model. This is important because higher ranks are a guarantee in the implementation of accurate character recognition by the HMMs.

B. Regression vs. Classification Tasks

One major difference is that regression predicts continuous values—things like stock prices—while classification gives discreet labels, such as character recognition. The former aims to minimize the prediction error, while the latter tries to maximize accuracy in placing inputs into categories.

C. Suggestions for Building a Predictive System for Stock Data

Utilize time series analytics, combined with relevant features—that is, historical prices and volumes—to make sound predictions on stock prices and changes. The developed machine learning ARIMA/LSTM work shall be good enough to capture the trends and present a prediction closer to the targets through careful feature engineering and validation.

References

[1] Tong-Hua Su, Tian-Wen Zhang, De-Jun Guan, Hu-Jie Huang, *"Offline 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