Recognition of Handwritten Telugu documents using HMM

A Venkata Srinivasa Rao[1], V Kavya[2], K Hari Charan[3], P Nikhil Babu[4], Ch N S Manojna[5]
[1]*Department of ECE, SASI Institute of Technology & Engineering, Tadepalligudem, W.G.Dist, AP, India*
[2,3,4,5]*UG Students, Department of ECE, SASI Institute of Technology & Engineering, Tadepalligudem, W.G.Dist, AP, India*

***Abstract*— Indian script recognition is a difficult task. The creation of comprehensive OCR systems for Indian language scripts is still in its early stages. Recently, full OCR systems for Bangla and Devanagari scripts were created. Character touching and overlap are two important issues that Telugu script recognition research should overcome. Telugu character segmentation is a challenging task for character recognition. In this paper, we make an attempt at segmentation and recognition of handwritten Telugu script using the Drop-Fall and HMM models.**

***Index Terms*—Segmentation, Telugu scripts, Recognition, Drop-Fall model, Hidden Markov model.**

I. INTRODUCTION

India is a multilingual Asian nation with a vast array of old written systems. For many of these scripts to take on a commercial form, OCR development is still needed. Preprocessing and binarization are considered to be crucial stages in the basic OCR technique. Character segmentation is frequently used in automated document processing systems as the next step after a line segmentation engine. The effectiveness of line segmentation has a direct impact on the precision of character segmentation and recognition. Numerous methods have been introduced. In this study, we proposed using hidden Markov models (HMMs) to recognise handwritten text.

The statistical techniques of hidden Markov modeling were first presented and studied in the late 1960s.For a variety of applications; they have proven to be incredibly helpful. Prior to being used to solve handwritten word recognition issues, this method was used to solve speech recognition issues.

Researchers tried to adapt this method to handwritten word recognition applications because speech recognition and the recognition of handwritten words share many similarities. HMM has been effectively applied for speech recognition, character recognition, and mobile communication methods since the 1980s. Moreover, it has been quickly adopted in the disciplines of fault diagnosis and bioinformatics.

The process of separating foreground pixels from background pixels is known as thresholding. The Otsu's method, proposed by Nobuyuki Otsu in 1979, is one of the various approaches to attaining optimal thresholding. Because it is both straightforward and efficient, this global thresholding selection method, which is based on spatial clustering, is widely used. To determine the threshold value where the weighted variation between the foreground and background pixels is the least, Otsu's method uses a variance-based technique. The important thing is to measure the distribution of background and foreground pixels while iterating over all conceivable threshold settings. Locate the threshold at which the dispersion is the smallest.

The Drop-Fall Algorithm is based on finding the point from which the merged portions should be separated by moving a marble on either side of the touching letters. Drop-fall algorithms are based on the principle that if one were to drop an imaginary marble from the top of the first letter and make the cut where the marble falls, one might make a fairly optimal "cut" between two connected characters. Despite appearing straightforward, the algorithm has turned out to be really helpful.

Telugu is a member of the family of syllabic alphabets and is derived from the Brahmi script. There are 36 consonants and 18 vowels, of which 35 consonants and 13 vowels are frequently used. A syllabic unit can consist of a vowel (V), a consonant (C), or any combination of the two. Among the combinations are CV, CC, CCV, and CCCV.

II. LITERATURE REVIEW

B. Hari Kumar et al. [1] implemented the handwritten scripts with advanced delicacy. Convolutional neural networks (CNN) for the recognition of handwritten and machine-published characters have to be increased. Information on character segmentation and recognition is processed quickly. Text is swiftly and continuously typed in large amounts. A paper-based form is frequently replaced with an electronic version that is easy to store or send through email. It is less expensive than hiring someone to manually enter a significant volume of textbook data.

Shoda Bhavya et al.'s [2] proposed model produces findings that favour the hunk partitioning technique as the length of the image text increases, whether it is applied directly to the picture of each word or to a succession of patches taken from each image. We discovered that normalising the handwritten text images increased the proposed recognition model's accuracy. There are also numerous tables with partial findings that enable us to assess the relative contributions of each suggested modification to the outcomes of the new model.

G.R. Hemanth et al. [3] proposed an adaptive technique for offline paragraph recognition by pre-processing and training the dataset sequentially using CNN and RNN. The input paragraph images are separated into line images and then processed further into word images using OpenCV contour algorithms. The NN model layers are then fed these word images for recognition. The RNN layers go over the output of the CNN layers in more detail. The CTC is provided with the output of the RNN layers in order to decode the output text. The outcomes show the promise of using CNN and RNN in tandem to continuously improve accuracy.

Thapani Hengsanankun et al. [5] proposed Thai word segmentation using a five-state left-to-right Hidden Markov model; a probabilistic technique of Thai word segmentation utilising HMM is proposed. The observation symbols are represented by Thai language components of speech, which are also utilised to create the HMM's unidentified word pattern. The eight classes that make up the classes of the unknown word are used to build the five-state left-to-right hmms. Prior to using the HMMs to categorise unknown words, the string-matching technique is used to identify overlapping terms.

AVS Rao et al. [6] proposed a binarization model for segmentation; noisy documents are cleaned with the help of the Modified IGT algorithm and then segmented using a conventional profile mechanism. Analysis of the efficiency suggests optimization of the present model in the domain of vowels, consonants, CV sounds, and other (grammar and punctuation marks, etc.), considering them inclusively and/or exclusively.

Irfan Ahmad et al. [8] propose a sub-character HMM-based text, where sub-character HMMs make use of similar patterns occurring in different characters of the Arabic script. Researchers studied sub-character HMM-based text recognition in this paper, where the fundamental HMM units are sub-characters rather than characters or character-shapes. Sub-character HMMs take advantage of the Arabic script's well-known tendency for similar patterns to appear in various characters or in various contexts for the same character.

AVS Rao [9] proposed a segmentation technique for an ancient Telugu document image in which the horizontal profile pattern is convolved with a Gaussian kernel. It is only applicable for noisy images. Document images are perceived as background and foreground information containers. Background information reflects the characteristics of noise, whereas foreground information is embedded with text, with lower intensity values tending towards black pixels. The characteristics of noise are analysed in the horizontal and vertical profiles. Line segmentation is carried out by convolving a horizontal profile with a Gaussian kernel of order 1 and sigma 3. However, individual character segmentation from the script line is found to be ineffective with this approach due to the high degree of non-uniform intensities. An extensive analysis is carried out to identify the relationship between the maxima and minima of the vertical profile. A thresholding approach is adopted while segmenting the individual characters in the script line. An efficiency of 73.58% is observed with this approach without losing any information in the ancient document image.

AVS Rao et al. [10] proposed Drop-Fall Algorithm to improve the performance of segmentation of touching handwritten Telugucharacters. The proposed technique produces better results when handwritten Telugu characters are segmented (created with software). This drop-fall is especially helpful for

cutting the related parts at various points where touching characters are present. As a future development, the same technique will be tried on Telugu characters that are handwritten and mechanically printed.

Florence Luthy et al. [12] proposed three different recognizers based on hidden Markov models, as well as the author's independent experiments, which are reported in this paper.

V. Jagadeesh Babu et al.'s [13] proposed online handwritten symbol recognition system for Telugu is based on HMM and uses a combination of time-domain and frequency-domain features.

III. METHODOLOGY

An effective methodology for the recognition of Telugu handwritten documents is provided, along with information on how the documents were cleaned and the relevant phases of the processing method. Flow chart for the proposed model given in Fig-1.

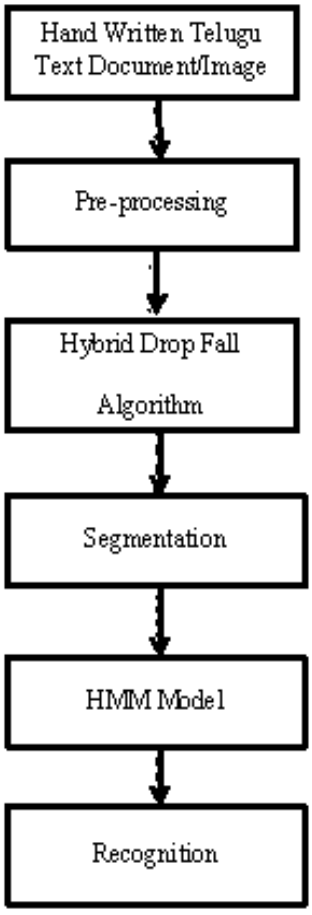


Fig-1.Flow Chart of the Proposed Model

In this work, initially 15 handwritten Telugu text documents were collected from the NET.

Binarization is the initial stage of preparing the document for future processing. The chosen thresholding techniques are either global or local, depending on how much the original document has degraded. The thresholding phenomenon is an easy-to-use technique for separating the object information from the cluster of pixels that are most likely linked to background information. In this article, the ground and background clusters are separated using the OTSU method (global thresholding). The use of this global thresholding enhances the contrast between the characters and their surroundings. Images with isolated characters that are in black and white (binarized) are the outcome.

The drop-fall algorithm is based on the principle that a cut should be made where a hypothetical marble lands in between two connected characters. According to this straightforward explanation of the method, the main problem that needs to be solved in order to implement it is where to drop the marble from. This is crucial because if the process begins in the incorrect location, the "marble" could easily slide down the first or second digit's left or right sides and become utterly useless. There are various ways to choose where to begin the drop-falling procedure. It makes sense to begin the scene as near to where the two characters are as possible.

Segmentation plays a major role in document image analysis. The segmentation of Telugu script into meaningful units is somewhat difficult because of the cursive nature of the script. Segmentation of handwritten text documents into individual characters or digits is an important phase in document analysis, character recognition, and many other areas. All text and image segmentation can be achieved at three levels:

1. Line Segmentation
2. Word Segmentation
3. Character Segmentation

The Hidden Markov Model (HMM) [5], a statistical model that uses a Markov process with hidden and unknown parameters, was first put forth by Baum L.E. (Baum and Petrie, 1966). The hidden parameters in this model are found using the observed parameters.

Then, additional analysis is conducted using these parameters. A form of the Markov chain is the HMM. Although its state cannot be seen directly, the vector series can be used to determine it. HMM has been effectively applied for speech recognition, character recognition, and mobile communication methods since the 1980s. Additionally, it has been swiftly embraced in the disciplines of fault diagnosis and bioinformatics.

The main principle of HMM is that observed events are connected to states through the probability distribution rather than having a one-to-one correlation with states. It is a stochastic process that explains state transitions as well as a stochastic process that characterises the statistical correspondence between states and observed values. The basic stochastic process is a Markov chain. Observers can only view the observed value; they are unable to view the states. The existence of states and their properties are determined by a stochastic process. Consequently, it is known as a "hidden" Markov model.

A Markov model is a stochastic method for randomly changing systems that possess the Markov property. This means that at any given time, the next state is only dependent on the current state. HMM is based on the statistical Markov model. In HMM, the system's states are hidden (unknown), but whenever the system is in a certain state, it emits an observable or visible output.

Let X_n and Y_n be discrete-time stochastic processes and $n \geq 1$. The pair (X_n, Y_n) is a *hidden Markov model* if

- X_n is a Markov process whose behavior is not directly observable ("hidden");
- $P(Y_n \in A | X_1 = x_1, \dots, X_n = x_n) = P(Y_n \in A | X_n = x_n)$ for every $n \geq 1 x_1, \dots, x_n$.

Let X_n and Y_t be continuous-time stochastic processes. The pair (X_t, Y_t) is a *hidden Markov model* if

- X_t is a Markov process whose behavior is not directly observable ("hidden");
- $P\left(Y_{t_0} \in A | \{X_t \in B_t\}_{t \leq t_0}\right) = P\left(Y_{t_0} \in A | X_{t_0} \in B_{t_0}\right)$ for every t_0 .

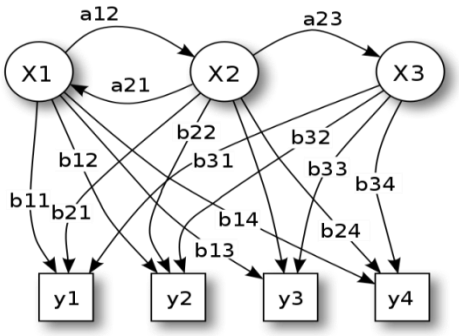


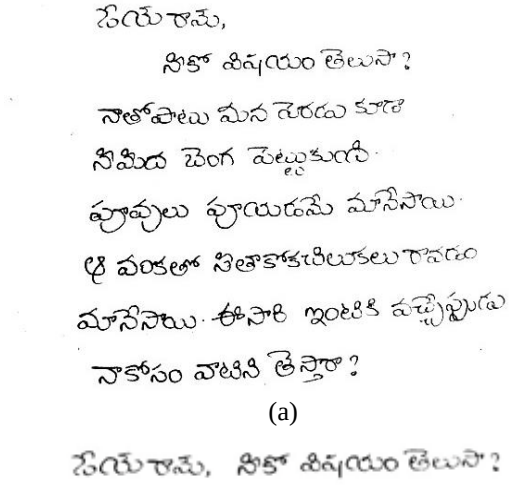
Fig-2. Probabilistic parameters of a hidden Markov model (example)

X — states
y — possible observations
a — state transition probabilities
b — output probabilities

IV. RESULTS & DISCUSSIONS

The proposed approach was examined using a variety of Telugu samples collected from the NET. The input images are pre-processed using the Otsu's binarization technique. The Drop Fall algorithm is used to segment the pre-processed image. The Drop Fall algorithm's results are further used as datasets.

Meanwhile, we took another input image and segmented it into individual characters or digits, which is a necessary phase in character recognition. The HMM approach is used for character recognition, as previously stated. A typical image is shown in Fig 3(a). Figures 3(b), 3(c), and 3(d) show the results of implementing the defined algorithm for line, word, and character segmentation.



(a)

(b)

(c)

(d)

Fig-3.(a) Input Image (b) Line Segmented Image (c) Word Segmented Image (d) Character Segmented Image

V. CONCLUSIONS

In the present work, character recognition in Telugu text document images is proposed with the inclusion of segmentation technique. The strategies proposed for increasing segmentation system performance also have a wider impact on the accuracy of handwriting recognition. The segmentation of Telugu characters that have been touched up by hand and improved using software yields superior results when utilising the proposed Drop-Fall and HMM techniques. This Drop-Fall is particularly helpful for cutting the related components at various points where touching characters are present. But we used limited number of characters to train the HMM model for recognition purpose. In this paper we made an attempt to work with HMM model for recognition of characters irrespective of results.

REFERENCE

[1] B Hari Kumar, R Likhita, P Reshma, U Mounika, Sk Hazra Begum,"*Character Segmentation and Recognition Using Bharathi Script* ", Proceedings of Journal of the Asiatic Society of Mumbai, ISSN:0972-0766, Vol. XCV, No.1(II), 2022.

[2] Shoda Bhavya, Shaik Chandini, Suvarnaganti Deepika, Parvathaneni Kiranmayi, "*Handwritten Character Recognition Using CNN* ", Proceedings of International Journal of Advance Research and Innovative, Issue-3, Vol-8 , 2022

[3] G.R. Hemanth, M. Jayasree, S. Keerthi Venii, P. Akshaya, and R. Saranya, "*Cnn-Rnn Based Handwritten Text Recognition*", Proceedings of ICTACT Journal on Soft Computing, ISSUE: 01, Vol. 12, 2021.

[4] Vijaya Krishna Sonthi, Dr. S. Nagarajan, Dr. N. Krishnaraj, "*Automated Telugu Printed and Handwritten Character Recognition in Single Image using Aquila Optimizer based Deep Learning Model* ", Proceedings of (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, 2021.

[5] Thapani Hengsanan kun, Atchara Namburi, " *Improving Thai Word Segmentation Using HMM: A Case Study of Sentiment Analysis"*, Proceedings of 24th International Computer Science And Engineering Conference (ICSEC), 2020.

[6] N VenkataRao, A S C S Sastry, Dr A S N Chakravarthy, A V Srinivasa Rao, "*Analysis of canonical character segmentation technique for ancient Telugu text documents*", Proceedings of Journal of Theoretical and Applied Information Technology, Vol.82. No.2, 2015.

[7] Namrata Dave, "*Segmentation methods for handwritten Character Recognition*", Proceedings of Journal of Signal processing, Image Processing and Pattern Recognition, Vol.8.No.4, 2015 .

[8] Irfan Ahmad, Leonard Rothacker, Sabri A. Mahmoud, "*Novel Sub-Character HMM Models for Arabic Text Recognition*", Proceedings of 12th International Conference on Document Analysis and Recognition (ICDAR), 2013.

[9] Srinivasa Rao A V, "*Segmentation of Ancient Telugu Text Documents*", Proceedings of I . J .Image, Graphics and Signal Processing (IGSP), 2012.

[10] Srinivasa Rao A V, D R Sandeep, V B Sandeep, S Dhanam Jaya, "*Segmentation of Touching Hand written Telugu Characters by Drop Fall Algorithm*", Proceedings of International Journal of Computers &Technology, 2012.

[11] A Venkata Srinivasa Rao, Madasu Subharao, Nekkanti Venkata Rao, A S C S Sastry, L Pratap Reddy, "*Segmentation of Touching Handwritten Numerals and Alphabets*", Proceeding of Second International Conference on Computer and Electrical Engineering , 2009 .

[12] Florence Luthy, Tamas Varga, and Horst Bunke, "*Using Hidden Markov Models as a Tool for Handwritten Text Line Segmentation*", Proceedings of the ninth International Conference on Document Analysis and Recognition (ICDAR), 2007.

[13] V Jagadeesh Babu, L Prasanth, R Raghunath Sharma, "*HMM- Based Online Handwriting Recognition System for Telugu Symbols*", Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR), 2007.