HMM based – Character Recognition for offline Handwritten Indic Scripts

Sandhya N
Dayananda Sagar College of Engineering,
Visvesvaraya Technological University
Kumarswamy Layout, Bangalore

Rahul Bant
Dayananda Sagar College of Engineering,
Visvesvaraya Technological University
Kumarswamy Layout, Bangalore

Krishnan R
Dayananda Sagar College of Engineering,
Visvesvaraya Technological University
Kumarswamy Layout, Bangalore

Ramesh Babu D R
Dayananda Sagar College of Engineering,
Visvesvaraya Technological University
Kumarswamy Layout, Bangalore

***Abstract*— Handwritten character recognition is always an advanced area of research in the field of image processing and pattern recognition. But most of the work done in handwritten character recognition is for English, Arabic, Chinese, Korean, and European Languages and a very few works have been reported for some Indian Languages. Recognition of handwritten characters in Indic scripts is a complex task due to the similarities between characters under different writing styles and large character sets. A character-based off-line handwritten recognition system is proposed here using Hidden Markov Models for two major Indic scripts namely – Kannada and Tamil which uses shape features and Singular, Value, Decompositions coefficients. The method employed involves stages namely, preprocessing, segmentation, feature extraction, training and recognition. Characters are modeled using Hidden Markov Models. The word models are built by concatenating character models. The recognition accuracies obtained for characters is 76% for Kannada and 70% for Tamil. Whereas the recognition accuracies obtained for words is very less which is 40% for Kannada and 30% for Tamil.**

Keywords— Hidden Markov Model, Handwritten character recognition, Word recognition, Kannada, Tamil words

I. INTRODUCTION

Character Recognition which is an area of Pattern Matching is the mechanical or electronic conversion of text into machine-encoded text. The text may be Handwritten or Printed. Handwritten Character Recognition (HCR) is the ability of a computer to receive and interpret intelligible handwritten input from sources such as touch-screens, stylus and other devices. Printed Character Recognition is the ability of a computer to receive and interpret intelligible input from sources such as paper documents, photographs etc.

Handwritten recognition is of two types: offline and online. Off-line Handwritten recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. Online Handwritten Recognition involves the automatic conversion of text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching.

In this paper we use Hidden Markov Model for recognizing Kannada and Tamil characters. The Kannada alphabet is an abugida of the Brahmic family, used primarily to write the Kannada language, one of the Dravidian languages of southern India. Tamil script belongs to the family of syllabic alphabets and consists of symbols for vowels and consonants.

The paper is organized as follows: Section 2 explains the related work. Section 3 presents the HMM- based recognition system. Section 4 discusses the results and section 5 gives the conclusion.

II. RELATED WORK

Prior work on handwritten character recognition for Indic scripts namely Kannada and Tamil are very limited. This section gives the details about the survey carried out for the offline handwritten character recognition.

In [1] an unconstrained Kannada Handwritten Text Database (KHTD) is introduced. To provide a framework for other researches, recent text-line segmentation results on this dataset are also reported. [2] presents a survey of applications of OCR in different fields and further experimentation for three important applications such as Captcha, Institutional Repository and Optical Music Character Recognition. In [3] the comparison study between the various algorithms based on binarization algorithms is done and proposes methodologies for the validation of binarization algorithms. In [4] it has been tried to achieve automatic recognition of handwritten Devanagari Script by using various algorithms. [5] describes that the performance of OCR algorithms and systems is based on the recognition of isolated characters.

The paper [6] proposes a Hidden Markov Model (HMM) for recognition of handwritten Devanagari texts. The classification accuracy is 87.71% and 82.89% for training and test sets respectively. [7] describes the simplest way of character recognition, based on matching the stored patterns or prototypes against the character or word to be recognized. For improved classification Deformable Templates and Elastic Matching are used for recognition task. A system for offline recognition of cursive handwritten Tamil characters is presented in [8]. [9] addresses the problem of document binarization as a pre-processing step for optical character recognition (OCR) for the purpose of keyword search of historical printed documents. [10] proposes an offline handwritten character recognition method of converting handwritten text into machine process-able format using Principal Component Analysis (PCA) for handwritten Gurumukhi characters.

III. HMM-BASED RECOGNITION SYSTEM

The HMM based handwritten recognition system for Indic Scripts is as shown in figure 1.

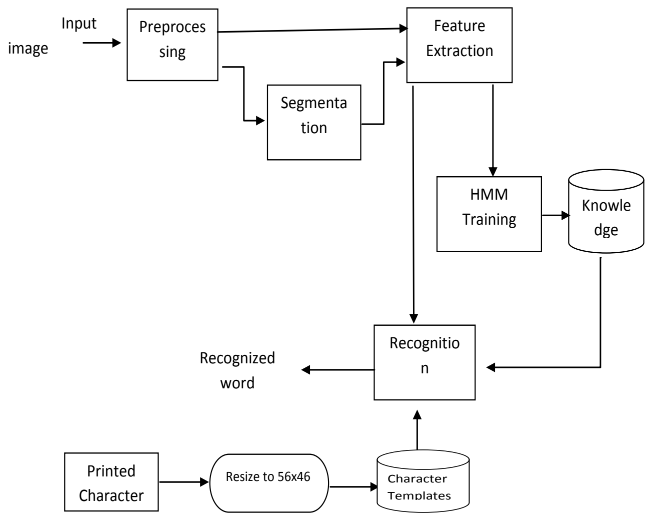

Fig 1: System Architecture for Offline Handwritten Word Recognition for Indic Scripts

A. *Template Creation*

First the printed characters of the Indic scripts – Kannada and Tamil are selected. Each printed character is resized to an image of size 56x46. Once the characters are resized they are stored in the character template database.

B. *Preprocessing*

The pre-processing is a series of operations performed on the scanned input image. It essentially enhances the image rendering it suitable for segmentation and feature extraction. The various tasks performed on the image in pre-processing stage are:

- ➢ Noise Removal
- ➢ Binarization
- ➢ Dilation
- ➢ Normalization

- *Noise Removal*

When the document is scanned, the scanned images might be contaminated by additive noise and these low quality images will affect the next step of document processing. Therefore, a pre-processing step is required to improve the quality of images before sending them to subsequent stages of document processing. Due to the noise there can be the disconnected line segment, large gaps between the lines etc. So it is very essential to remove all of these errors so that the information can be retrieved in the best way. One such additive noise is called as “Salt and Pepper Noise”. The black points and white points sprinkled all over an image, typically looks like salt and pepper, which can be found in almost all documents. Figure 2 shows the noise removal from the input image.

- *Binarization*

Binarization of gray-scale character images is a crucial step in offline character recognition. Good binarization facilitates segmentation and recognition of characters. Binarization process converts a gray scale image into a binary image.

- *Dilation*

The function of the dilation is that it takes the binary image as input and converts it into the dilated image i.e. it enlarges the binary image.

- *Normalization*

Normalization is required as the size of the character varies from person to person and even with the same person from time to time. It is the process of converting the random sized image into standard sized image. The input character image is normalized to a fixed size of 56x46.

C. *Feature Extraction*

In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate. The extracted features are used to build the HMM character model. The following features are extracted:

- ➢ Aspect Ratio
- ➢ Height Fraction
- ➢ Width Fraction
- ➢ Singular, Value, Decompositions (svd) co-efficient

- *Aspect Ratio*

Aspect ratio of a stroke is defined as the ratio of height of the stroke to the width of the stroke. Figure 2 shows the aspect ratio for the Kannada and Tamil character. The ratio is as given in equation (1).

$$\text{Aspect Ratio} = \frac{\text{Height(Stroke)}}{\text{Width(Stroke)}} \qquad (1)$$

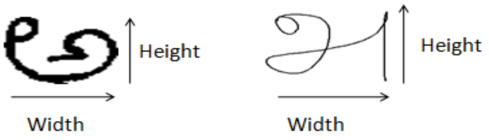

Fig 2: Aspect Ratio for Kannada and Tamil character

- *Height Fraction*

Height fraction captures the vertical location of the stroke with respect to the character and is given in equation (2).

$$\text{Height\_Fraction} = 1 - \left(\frac{\bar{y}(\text{stroke})}{y_{max}(\text{character})}\right) \qquad (2)$$

where $\bar{y}(\text{stroke})$ is the average y value of the stroke and $y_{max}(\text{character})$ corresponds to the y value of the lower most point in the character.

- *Width Fraction*

Width fraction captures the extent to which the stroke spans horizontally across the character and is given in equation (3).

$$\text{Width fraction} = \frac{\text{Width(stroke)}}{\text{Width(character)}} \qquad (3)$$

where width(stroke) and width(character) correspond to the bounding box width of the stroke and the characters.

- *Singular Value Decompositions (svd)*

[U,S,V] = svd (X) produces a diagonal matrix S of the same dimension as X, with nonnegative diagonal elements in decreasing order, and unitary matrices U and V so that X = U*S*V.

D. Segmentation

The segmentation of characters can make the difference between very good and very poor results from a recognition process. The goal of the segmentation is to partition the multiple character images into regions, each containing an isolated complete character.

E. Training

Training is carried out using the Baum-Welch re-estimation procedure. The features extracted from the symbols are used to train each symbol. For modeling a character using HMM, a simple left-to-right topology with no state skipping was adopted.

F. Recognition

In recognition the input characters are tested. The input characters are preprocessed segmented and its features are extracted. The features extracted during training are used for recognition. The output character image from the HMM is identified and then matched with the character templates stored in the character template knowledge base which gives the final recognition result.

IV. Results and Discussion

The performance of the classifier model is evaluated using the different parameters, and graph is plotted to check the performance validation.

The classifier model was trained with a total of 100 training samples for Kannada and 61 samples for Tamil and tested with few characters and words for Kannada and Tamil. Number of true positives (TP) and false negatives (FN) are noted down for these characters and performance measure accuracy is calculated using the equation (4). Table 1 shows the accuracies with character images for an average of 10 characters. Table 2 shows the accuracies with word images.

$$\text{Accuracy} = \text{TP} / (\text{TP+FN}) \qquad (4)$$

TABLE I. RECOGNITION ACCURACY FOR CHARACTERS

x axis	y axis	
	Tamil	Kannada
10	70%	76%

In the performance analysis graph x axis indicates characters and y axis indicates recognition accuracy for Kannada and Tamil. Figure 3 shows the performance graph for Kannada and Tamil characters.

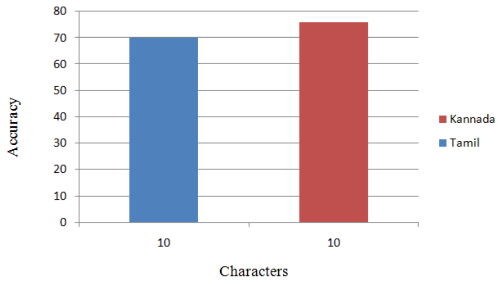

Fig 3 Performance analysis Graph for characters

TABLE II. RECOGNITION ACCURACY FOR WORD

x axis	y axis	
	Tamil	Kannada
5	30%	-
10		40%

V. CONCLUSION AND FUTURE WORK

HMM based offline handwritten character recognition has been proposed for the two major Indic scripts namely – Kannada and Tamil. In feature extraction three shape features and Singular, Value, Decompositions (S, V, D) co-efficients have been extracted. These features are used to build the character models using Hidden Markov Model. The characters are trained using the Baum-Welch algorithm. The model has been tested considering various characters and words from the database. The recognition rate achieved for characters is 76% for Kannada and 70% for Tamil. Whereas the recognition rate achieved for words is very less which is 40% for Kannada and 30% for Tamil.

By adding additional relevant features and combining different classifier models the accuracy can be increased. Some features specific to the mostly confusing characters can be used, to increase the recognition rate. The recognition accuracy for Kannada and Tamil words are very less. Therefore increasing accuracy for words can be carried out as future work.

REFERENCES

[1] Alaei A, Nagabhushan P, Pal U, "A Benchmark Kannada Handwritten Document Dataset and Its Segmentation," Document Analysis and Recognition (ICDAR), 2011 International Conference on , vol., no., pp.141,145, 18-21 Sept. 2011.

[2] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin, "A Survey of OCR Applications", International Journal of Machine Learning and Computing, Vol. 2, No. 3,pp.314-318, June 2012.

[3] Aroop Mukherjee, "Enhancement of Image Resolution by Binarization", International Journal of Computer Applications (0975 – 8887), Vol 10, No.10, November 2010.

[4] Ashwin S Ramteke and Milind E Rane, "A Survey on Offline Recognition of Handwritten Devanagari Script", International Journal of Scientific & Engineering Research Volume 3, Issue 5, May-2012.

[5] B S Saritha and S Hemanth", An Efficient Hidden Markov Model for Offline Handwritten Numeral Recognition", InterJRI Computer Science and Networking, Vol. 1, Issue 1, July 2009.

[6] Bikash Shaw and Swapan Kumar Parui, "Offline Handwritten Devanagari Character Recognition: An HMM Based Approach", Proceedings of the 2nd International Conference on Pattern Recognition and Machine Intelligence, pp. 528-535, 2007.

[7] Dr. Pankaj Agarwal, "Handwritten Character Recognition Using Kohonen Network", International Journal Computer Science and Technology, vol. 2(3), pp.112-115, 2011.

[8] Kannan, R.J. Prabhakar, R. Suresh, "Off-line Cursive Handwritten Tamil Character Recognition", Security Technology, International Conference on SECTECH '08, pp.159,164, 13-15, Dec. 2008.

[9] Maya R. Gupta, Nathaniel P. Jacobson, Eric K. Garcia "OCR Binarization And Image Pre-Processing For Searching Historical Documents", Pattern Recognition, Vol. 40 Issue 2, pp. 389-397, February, 2007.

[10] Munish Kumar, R. K. Sharma and M. K. Jindal, "Offline Handwritten Gurmukhi Character Recognition: Study of Different Feature-Classifier Combinations" , ACM 978-1-4503-1797, 2012.