A SURVEY ON MALAYALAM OCR MODULES

[1]Joslin Johnson, [2]Catherine Davis, [3]Ashly Raphel, [4]Asst Prof. Soumya Varma
[1,2,3,4] Department of Computer Science & Engineering, Sahrdaya College Of Engineering & Technology

ABSTRACT: People start learning to read and write during the early stage of education. As years pass by they may have acquired good reading and writing skills. It may not be difficult for them to read any kind of either printed or handwritten characters. But Computers may find difficulty in deciphering many kinds of printed characters which is of different fonts and styles or handwritten characters. Malayalam OCR is a complex task owing to the various character scripts available and more importantly the difference in ways in which the characters are written. The dimensions are never the same and may be never mapped onto a square grid unlike English characters. This survey paper provides the details of different Malayalam ocr modules and their techniques for identifying and recognizing the malayalam old scripts and converting it to new Malayalam script.

KEYWORDS: Malayalam, Handwritten characters, Old script to new script, identification of Malayalam script, Optical Character Recognition.

INTRODUCTION

OCR is one of the most challenging areas of image processing and pattern recognition. OCR plays a vital role in creating digital library expanded. It is highly essential and unavoidable while dealing with Indian languages for which there has been little digital access.Only few approaches had been devised for handwritten Malayalam documents which include wavelet Transforms, Kohonen Networks and Projection Profiles. Since little attempts have been made to develop OCR that could recognize handwritten Malayalam documents, this area needs further more developments and the researches are still going on this field. A lot of techniques of pattern recognition such as Template Matching, Neural Networks, Syntactical Analysis, Hidden Markov Models, Bayesian Theory, etc have been exhumed to develop robust OCRs for different languages. The current system has efficient and inexpensive OCR packages which are commercially available for the recognition of printed and handwritten documents. Among those we have enough facilities for languages such as English [1], Chinese [2] etc. When considering the Indian languages, many attempts are made to develop the OCR system for Devanagari, Oriya, Tamil [3], Telugu [4], and Kannada [5] etc. While taking Malayalam into consideration an effective method of recognition is still promising. The recognition of handwritten character recognition poses a great challenge to researchers. Even now a lot of problems in this area are still to be addressed. Handwritten character recognition (HCR) system is so complex with the variety of character structure and distorted and broken characters and personal independence.

It is hard to say that handwritten recognition exits for Malayalam language. In [6] has proposed an algorithm for the recognition of isolated handwritten Malayalam characters which used the HLH intensity patterns for the feature extraction technique. The input used in the present work is the image input given by the Light pen device. The characters are written through Light pen device and it is converted into 24 bit bmp image. The output is an editable computer file which is the equivalent character written by the user.

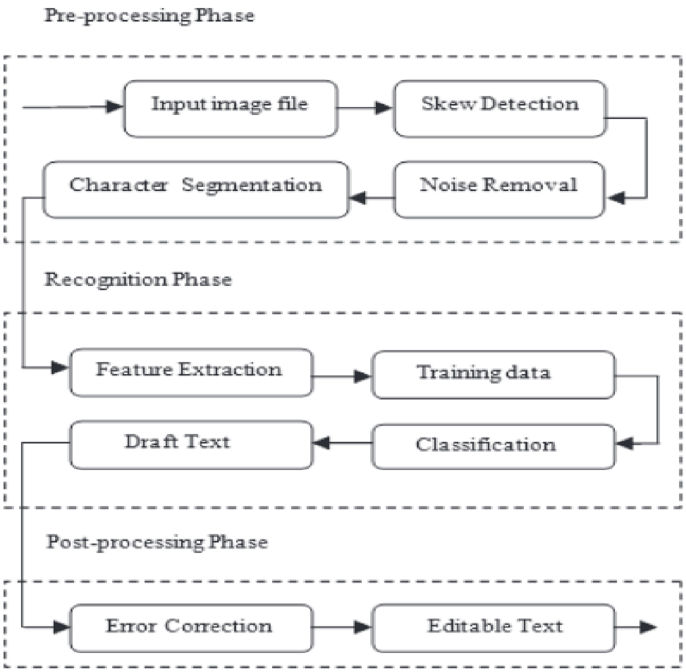

Figure 1: Block Diagram

METHODOLOGY

Malayalam is a South Indian language - which is the principal language of the State of Kerala, spoken by about 36 million people in the world. The Malayalam script is a Brahmic script used commonly to write the Malayalam language. Like many other Indic scripts, Malayalam follows a writing system that is partially alphabetic and partially syllable-based. The Malayalam script uses both old and new script for depicting characters.

Due to the complexity of the Malayalam character set, an efficient method for the recognition for handwritten characters has not been proposed till now. Based on Otsu's algorithm for binarization an OCR system was devised by Centre for Development of Advanced Computing [7] (CDAC) Thiruvananthapuram, Kerala, a Government of India Institution. In this system, projection profile method is used for skew detection and correction of image; and in the recognition phase linguistic rules are applied. An accuracy of 97% was reported in this method. Using wavelet based feature extraction and neural network based recognition, a new work was reported by M Abdul Rahiman and Rajasree [8]. Another work was reported by G Raju, [9]in which the daubechie wavelets (db4) were used for recognition. Another OCR system was proposed by Lajish V L, Suneesh T K and Narayanan N K [10] [11] which was based on statistical classification. Most recently, a method for the recognition of Isolated Handwritten Malayalam Character using HLH Intensity Patterns was devised by M Abdul Rahiman, G Manoj Kumar and M S Rajasree [12].

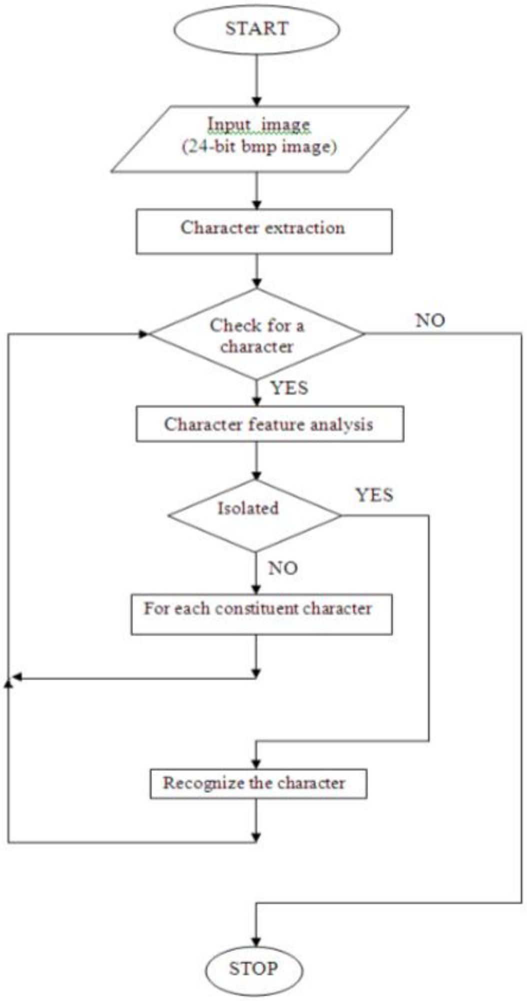

Figure 2: Flow chart of Combinational HCR System

Work	Input	System	Segmentation	Features	Classifier
Abdul Rahiman M	Inscribed using Light Pen	HCR	Identifying background and foreground color intensity of image	Number of horizontal and vertical pillars	Dynamic matrix
Gowri Shankar V	Super Pen	Online HCR	Not mentioned	18 shape features based on the direction	Soft matching of strings

Bindu Phillip	Scanned Image	Bilingual Malayalam English PCR	Classical Projection Profile	Average Gap, Singular Values, Frequency of transition	Support Vector Machine
Amrita Sampath	Generated using Stylus pen	Online HCR	-	Freeman code representation of direction information	Neural network using back propagation
G. Raju	Scanned document	Offline HCR	Not mentioned	Count of zero crossing	Feed forward Neural Network
Bindu S Moni	Scanned document	Offline HCR	Not mentioned	Direction of pixels with respect to neighboring pixels	Modified Quadratic Discriminant function

TABLE 1: A comparative study of the existing OCR modules

CONCLUSION

This project helps to convert all handwritten Malayalam old scripts (hard copies) to its editable form (soft copies) and they can be used in the future. This is more useful in Government offices where a number of documents have to be handled. In such offices maintaining a soft copy of the documents is more organized than keeping a hardcopy, especially for old documents. Therefore using this system we can keep soft copies of all these documents and hence the problem of damaged old documents can be avoided. In this project, we are trying to identify and convert old scripts of Malayalam language to its present age new scripts using OCR module. It involves the following steps. First the image of the Malayalam old scripts is given as input to the system. Then noise is removed from the scanned image. Then the image is converted to text which is the new Malayalam script.

REFERENCE

[1] D. Trier, A K Jain and T Taxt, “Feature Extraction methods for Character Recognition – A Survey”, Pattern Recognition, Vol 29, pp 641-662,1996.

[2] S N Srihari,X Yang and G R Ball, “ Offline Chinese Handwriting Recognition: an assessment of current Technology”, Front. Computer Science, China, Vol. 1 (2), pp 137-155, 2007.

[3] R. Seetha lakshmi., T.R. Sreeranjani , T. Balachandar, Abnikant Singh, Markandey Singh, Ritwaj Ratan, and Sarvesh Kumar, “Optical Character Recognition for printed Tamil text using Unicode”, Journal of Zhejiang University SCI 6A(11) , pp.1297-1305, 2005.

[4] C. V. Lakshmi and C Patvardhan, “ A multi-font OCR system for printed Telugu text”, Proc. of Language engineering conference LEC, Hyderabad, pp.7-17, 2002.

[5] T. V. Ashwin and P. S. Sastry, “ A font and size independent OCR system for printed Kannada documents using support vector machines”, Saadhana, Vol. 27, Part 1, pp. 35–58,February 2002

[6] M Abdul Rahiman, Aswathy Shajan, Amala Elizabeth and M S Rajasree, “ Isolated Handwritten Malayalam Character recognition based on HLH intensity patterns”, Proc of International Conf on Machine learning and computing, ICMLC 2009, Banglore, NOV 2009.

[7] Journal of Language Technology, Viswabharat@tdil, July 2003.

[8] M Abdul Rahiman and M S Rajasree, “Printed Malayalam Character Recognition Using Back propagation Neural Networks”, Proc.of IEEE International Advance Computing Conference (IACC 2009),Patiala, March 2009.

[9] G Raju” Recognition of unconstrained handwritten Malayalam characters using zero crossings of wavelet coefficients”, Proc. of International Conference on Advanced Computing and Communications, ADCOM, pp 217-221, Dec 2006.

[10] Lajish V L,Suneesh T K K and Narayanan N K, “ Recognition of Isolated handwritten images using Kolmogorov-Smirnov Statistical classifier and K –nearest neighbor classifier”, Proc. Of International Conference on Cognition and Recognition, Mandya, Karnataka, December, 2005.

[11] Lajish V L, “ Handwritten Character Recognition using perpetual Fuzzy zoning and Class modular Neural Networks”, Proc. of fourth International Conf on Innovations in IT, 2007.

[12] M Abdul Rahiman and M S Rajasree, “Isolated Handwritten Malayalam Character Recognition using HLH Intensity Patterns”, Second International Conference on Machine Learning and Computing, Banglore,2010.