

An HCR System for Combinational Malayalam Handwritten Characters based on HLH Patterns

Abdul Rahiman M
Karpagam University
Coimbatore

Aswathy Shajan, Amala
IBM
Chennai

Rajasree M S
Govt College of Engg
Trivandrum

ABSTRACT

An efficient and robust algorithm for recognition of handwritten Malayalam characters is proposed in this study. Malayalam, also known as Kairali, is one of the four major Dravidian languages of Southern India. It consists of basically 15 vowels and 36 consonants. The existence of lot of their combinations and connected characters, the recognition poses a gargantuan challenge in front of us. Till now Malayalam lacks an efficient OCR which meets all conditions. The intertwined characters are very complex owing to the non-adherent styles in which they may be presented. Here we propose an algorithm which uses the inveterate characteristic features to recognize these characters with perceptive accuracy by utilizing the intensity variations in the way in which they may be written. This algorithm recognizes the antediluvian script of Malayalam characters which are connected in nature. Here the input is a 24-bit bmp image which can be encribed using the Light pen. The output is editable version of the recognized Malayalam characters. In our study we have classified the connected characters into 3 categories. The algorithm is tested for 3 sets of samples ranging 402 letters in noiseless environment and produces accuracy of 92%.

General Terms

Algorithms

Keywords

Malayalam; Optical Character recognition; Feature Extraction; Connected character; Intensity Variations; HLH Patterns.

1. INTRODUCTION

Optical Character Recognition (OCR) is the process of translation of images of typewritten or handwritten text into machine editable text, which 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. 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 exists for Malayalam language. This paper is intended to provide an efficient method for the development of OCR system for handwritten Malayalam characters which are connected in nature. In [6] we 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.

2. MALAYALAM LANGUAGE

Malayalam is the Official language for the State of Kerala, the southernmost part of India. This language is derived from the Grantha script, which is the descendant of Ancient Brahmi. The character set consists of 51 letters which includes 15 vowels and 36 consonants. The complete character set of Malayalam is depicted in Fig 1(a-c). The set also consists of 12 vowel signs. These vowels are called as dependent vowels as they are validated unless present in some combination with a consonant or a conjunct.

The Malayalam script exhibits no inherent symmetry and thus making the recognition task very tedious. The new script of Malayalam language is marked by isolated characters. The old script on the other hand is marked by the combination of these characters in different forms. As a consequence of the disparity, irregularity and the diversity in the ways in which the connected characters are presented, an algorithm which is totally independent on the size yet concentrates on the characteristic features is chosen. The old character set of Malayalam is heavily complex. As a result of the difficulties of printing Malayalam, a simplified or reformed version of the script was introduced during the 1970s and 1980s. The main change involved writing consonants and diacritics separately rather than as complex characters. These changes are not applied consistently applied so the modern script is often mixture of traditional and simplified characters. Here we propose a methodology to identify these complex characters ie. the connected characters and print them in the reformed style.

The complete character set of Malayalam script shown in figure 1 consists of vowels, consonants and vowel signs.

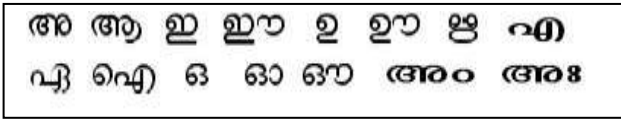


Figure 1 (a): Vowels.

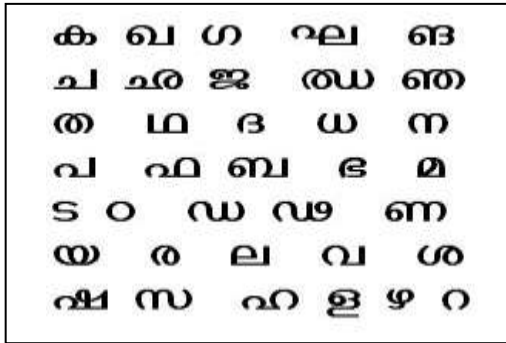


Figure 1 (b): Consonants.

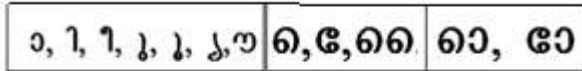


Figure 1 (c): Vowel signs.

Most of the connected characters appear in Old script. The difference between the Traditional orthography and Reformed Orthography is illustrated in Figure 2.



Figure 2: Difference in Old & New script

An illustration of the connected characters which were heavily used in the ancient Malayalam script is shown in Figure 3. Although now we usually use modern script there are certain letters for which the ancient connected characters are used especially when the text is handwritten. This emphasizes the need for an efficient algorithm which may correctly recognize the various ways of expressing the same sequence with meaning.

3. EXISTING RECONGNITION TECHNIQUES

It is hard to say that a complete Malayalam OCR exists which meets all conditions. Malayalam OCR lacks an efficient algorithm. Even in the field of printed characters there are little advancements for this language.



Figure 3: Combinational characters in Malayalam.

Even though the administrative language of Kerala is Malayalam, only a few works were reported in this area. Government of Kerala has now taken initiative for the development of this language and scope of development in this area is promising.

The first OCR system was developed by Centre for Development of Advanced Computing [7] (C-DAC) Thiruvananthapuram, a Government of India institution. It uses Otsu's algorithm for binarization and Projection profile method used for skew detection and correction of image. The recognition phase linguistic rules are applied. An accuracy of 97% is reported in this method. Another system is reported by M Abdul Rahiman and M S Rajasree [8] which uses wavelet based feature extraction and neural network based recognition. Bindu Philip and R D Sudhakara Samuel [9] proposed an OCR for Malayalam using column stochastic image matrix. In [10] Neeba N V and C V Jawahar proposed a method of recognition of Malayalam Character from books.

The recognition of handwritten Malayalam character is still in the stage of infancy. Only a little research is going on in this area. Our earlier work [6] in the field of handwritten Malayalam character recognition provided a new method for isolated characters. HLH intensity patterns were used for the recognition of characters and an accuracy of 86 percentages was achieved. Another work was reported by G Raju [11] in which the daubechie wavelets (db4) were used for recognition. Lajish V L, Suneesh T K K and Narayanan N K [12] proposed a system which is based on statistical classification. Artificial Neural Networks are applied for recognition of Handwritten Malayalam characters in the work done by Lajish V L [13]. The area of handwritten Malayalam character is still promising and offers a plethora of opportunities for research.

4. SYSTEM DESCRIPTION

Here we propose a way to identify the isolated as well as the complex connected script of Malayalam language in a noiseless environment. The flow chart of the system to recognize the characters is illustrated in figure 4.

Here in our study we start with the assumption that to find an isolated character. On successful the corresponding character is recognized and reported. In case of a connected character, we take the recognition process a step higher and will try to segregate the character into its corresponding counterpart and analyze each segment individually.

Initially we study the image and our first step will be to separate the characters assuming the connected character as a single character using any of the character separation algorithms and enclosing it in a matrix which we will be analyzing in the later

part of our study. If check for an isolated character fails then we will analyze the pattern in the pattern analysis phase where we will classify the connected characters into three modes.

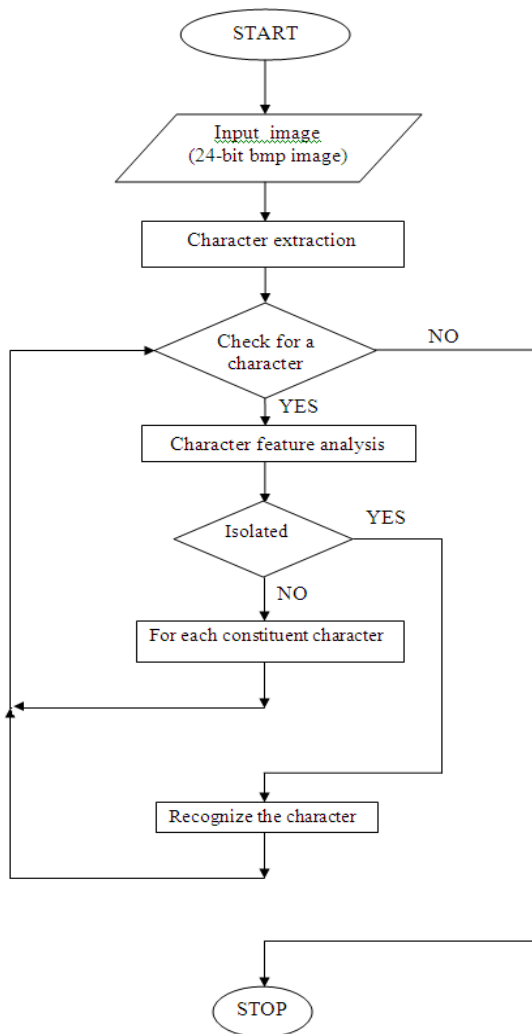


Figure 4: Flow chart of the system.

5. FEATURE EXTRACTION TECHNIQUE

In order to use this algorithm initially we assume the character has 2 intensity variations. The H-L notation has been adopted to represent background points and valid character path. We segregate the individual letters into a matrix which can be processed further using any of the character separation algorithms which is shown in figure 5.

Now we will go for the isolated character recognition algorithm. In case of connected character we try to separate them into one of the following 3 categories. The horizontal recurrence, vertical recurrence or special recurrence. In horizontal recurrence two characters are combined horizontally, in vertical recurrence vertically and in special recurrence the characters may have no inherent characters of the combined characters but still a combination of the original letters. Figure 6 illustrates these in detail.

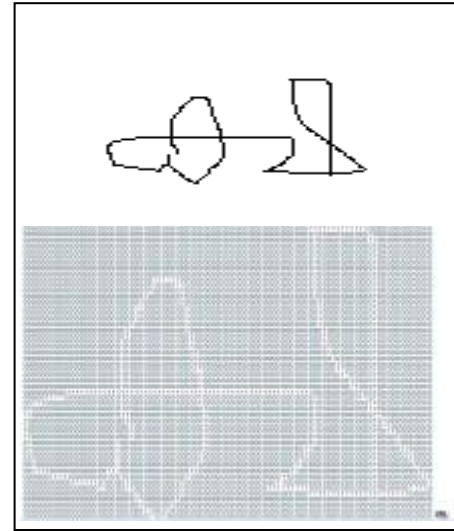


Figure 5: Matrix formed utilizing intensity variations.

Character	Combination	Isolated Characters	Type
കക	ക്ക	ക	Special recurrence
തത	ത്ത	ത	Horizontal recurrence
പപ	പ്പ	പ	Vertical recurrence
കഷ	ക്ഷ	ക, ഷ	Horizontal recurrence
ര	ര	ര, ത	Special recurrence

Figure 6: Character combination types.

6. RECOGNITION PROCEDURE

Here we have go in for the recognition based on the connected characters into the above 3 categories. The isolated character recognition specified in the algorithm is based on the HLH intensity patterns. After extracting the character into a matrix, if it is a normal isolated character it is recognized. In the case of vertical connected characters the letter gets horizontally partitioned based of most probabilistic occurrence of the high intensity. The division of character into various sized small elements are carried out and each small part is further analyzed to find if it makes up to an isolated letter. When we succeed in getting both the combination of letters in a pre expected manner and both are found to be liable the particular connected character is written in the modern style applicable for it. In the case of a horizontal recurrence vertical partitioning takes place and the particular letter sequence is identified. In special recurrence characters we will use the HLH intensity patterns to understand the characteristics and special vertical checks and horizontal checks are applied on the character as a whole and on the parts and the correct letter sequence gets identified. Figure 7 depicts the horizontal and vertical checks.

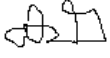
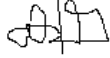


Character	Most probabilistic cut	Isolated Characters
		ക, ഷ
		പ, ഡ

Figure 7. Horizontal & Vertical divisions

7. THE ALGORITHM

The algorithm for the system is shown below. Here check for isolated characters are performed and horizontal and vertical checks are done. Based on this the character is placed in any of the three categories.

- Step 1 : Extract the character into a matrix.
- Step 2 : Check for isolated character occurrence.
- Step 3 : If true display the correct character.
and go to step 10.
- Step 4 : Check the length by width ratio of the matrix.
- Step 5 : If length < width then goes to step 7.
- Step 6 : Horizontal recurrence analysis is carried out .If
successful recognition step10.
- Step 7 : Vertical recurrence analysis is carried out. If
successful recognition step 10.
- Step 8 : Special recurrence algorithms are undertaken.
- Step 9 : The character stands unidentified.
- Step 10 : Stop.

8. EXPERIMENTAL INVESTIGATION

We conducted the experiment using different lines of text from multiple sources. We used a light pen to write on the paint to create a 24 bit bmp image which was given as the input. Handwritten characters with different styles and of different persons are used as database for study. A total of 629 handwritten characters are used for the experimental purpose. A set of specific connected characters were chosen from the various disciplines and the experiments were conducted. The output was an editable form of text in printable format using the modern script of Malayalam language. The experimental results are tabulated in Table 1. The successes in recognition of the vertical and special type recurrence characters are very much higher than the horizontal occurrence. However the overall efficiency of 91.41% has been achieved.

9. CONCLUSION

In this paper we have proposed an algorithm that can effectively recognize the combinational characters in Malayalam language. This is developed as an extension to the study of isolated handwritten [6] Malayalam characters. Experiments were conducted with different handwritten characters of different individuals. An accuracy of 92% has been achieved.

Our paper emphasizes on the inherent characteristic features of the various Malayalam letters and also the form in which they may be presented. This system used Light pen device to input the handwritten characters. Further studies are being carried out for the extension of this work to the scanned handwritten images. This can be achieved including pre processing works to this algorithm. Online recognition of handwritten characters is also experimented as an extension of this work. This system is further combined with the previous algorithm to recognize both the new and old script in Malayalam.

10. REFERENCES

- [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. Seethalakshmi., 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, Aewathy 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, pp 1140-44, March 2009.
- [9] Bindu Philip and R D Sudhakara Samuel, “ A Malayalam OCR system using column stochastic image matrix approach”, Proc of International Conf on Recent Technologies in communication and computing, Kottayam, December 2009.
- [10] Neeba N V and C V Jawahar, “ Recognition of books by verification and retraining”, Proc of International Conference on Pattern Recognition, Florida, December 2008
- [11] 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.
- [12] 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.
- [13] 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

Table 1: Performance Analysis of Handwritten Recognition

Input Document	Character Analysis						Complete Document	
	Horizontal Recurrence		Vertical Recurrence		Special Recurrence		Total Characters	Correctly Recognized
Set	Total Characters	Correctly Recognized	Total Characters	Correctly Recognized	Total Characters	Correctly Recognized		
1	68	64	92	85	38	34	198	183
2	45	41	66	60	95	85	206	186
3	123	112	48	44	54	50	225	206
Total	236	217	206	189	187	169	629	575
Recognition Success	91.2%		91.74%		90.37%		91.41%	

Abdul Rahiman M is currently working as Asst Professor in the Department of Computer Science & Engineering in LBS Institute of Technology for Women, Trivandrum, Kerala State, India. He did his M.Tech degree in Computer Science from Kerala University in Computer Science with specialization in Digital Image Computing and also undergone MBA degree in Systems and Post Graduate Diploma in Human Resource Management from Kerala University. He has many publications in various Journals and International conference proceedings. He is doing his PhD in Karpagam University, Coimbatore in the area of pattern recognition. He is a Life Member of Indian Society for Technical Education (ISTE) and Member of International Association of Computer Science & Information Technology (IACSIT).

Aswathy Shajan & Amala Elizabeth Thomas did graduation in Computer Science Engineering from Kerala University is working with IBM Chennai as Software Engineers.

Dr Rajasree M S, is a Professor in Computer Science Engineering and the Head of the Department of Computer Science and Engineering in Government College of Engineering, Trivandrum, Kerala, INDIA. She did her MTech from NIT Calicut and PhD from IIT Madras. She is serving in many professional bodies and has many publications in several International Proceedings and reputed Journals. She is guiding many research scholars in various Universities in India. She is a member of ISTE and Chairperson of Board of Studies, Kerala University.