An Efficient Character Recognition System for Handwritten Malayalam Characters Based on Intensity Variations

Abdul Rahiman M and Rajasree M S

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. Most people have no problem in reading any kind of light prints or heavy prints, upside down prints, prints of different fonts and styles, handwritten whether it is neatly or sloppily written. But Computers may find difficultly in deciphering many kinds of printed characters which is of different fonts and styles or handwritten characters. To evolve a panacea to this problem human brains have been indulging in various research activities. This paper is a humble attempt for the recognition of handwritten Malayalam *(a South Indian Language)* characters. In our study we have classified the connected characters into 3 categories. 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 enscribed 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 94%.**

Index Terms—**Malayalam, Optical character recognition, Feature extraction, Connected character; Intensity variations, HLH patterns.**

I. Introduction

Optical Character Recognition (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. 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 exits 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.

II. Malayalam Script

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 13 vowels and 37 consonants. The complete character set of Malayalam is depicted in Fig 1. 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

Manuscript received September 5, 2010; revised February 7, 2011

Abdul Rahiman M, Research Scholar, Karpagam University, Coimbatore. & Asst Professor, Computer Science & Engg, LBS Ins of Tech for Women Trivandrum, Ke rala , India.

Rajasree M S, Professor & Head, Department of Computer Science & Engg, Govt College of Engg Trivandrum, Kerala, India.

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 below figure 1 consists of vowels, consonants and vowel signs.

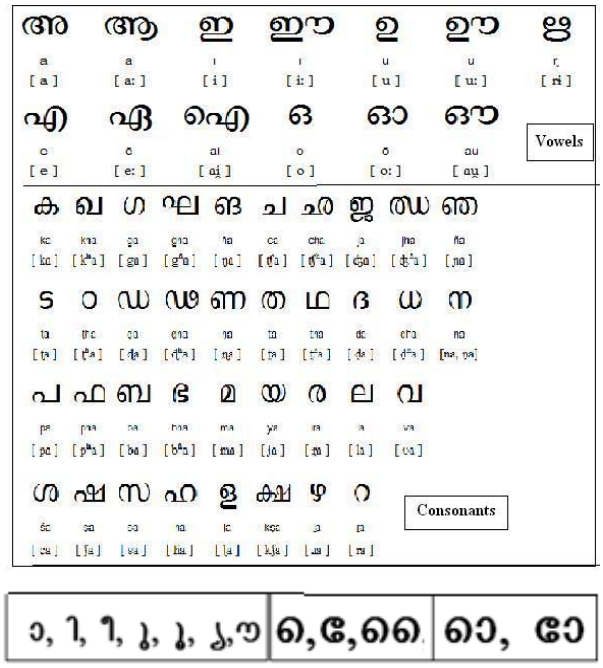

Fig 1: Vowels, Consonants & Vowel Signs

Fig 2 Combinational characters in Malayalam.

There exist a number of combinational characters which is depicted in figure 2. Most of the connected characters appear in Old script. An illustration of the character named Ka in Malayalam is shown in figure 3 where all the forms of vowels and vowel signs are associated with it.

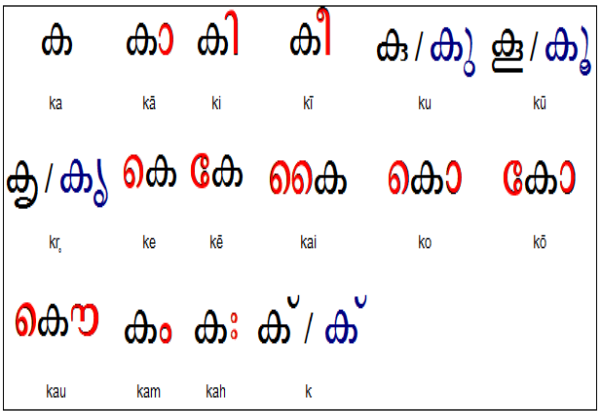

Fig 3: Vowel signs associated with a character ‘Ka”.

III. Literature Survey On HCR Malayalam

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

IV. Proposed Approach

This method employs recognition of isolated and

combinational handwritten characters in a noiseless environment. The basic principle is to identify specific terminologies in each character and extend the same to a set of characters in order to achieve accurate results with very low complexity algorithms. The separation of letters is shown in figure 4 which uses intensity variations for segregating the line and character from the scanned image. This work separate the entire character set in to three different classes. Ra type characters, Pa type characters and Special symbols. This classification is based on the shape and appearance of the character. This shape feature is extracted to recognize the letter.

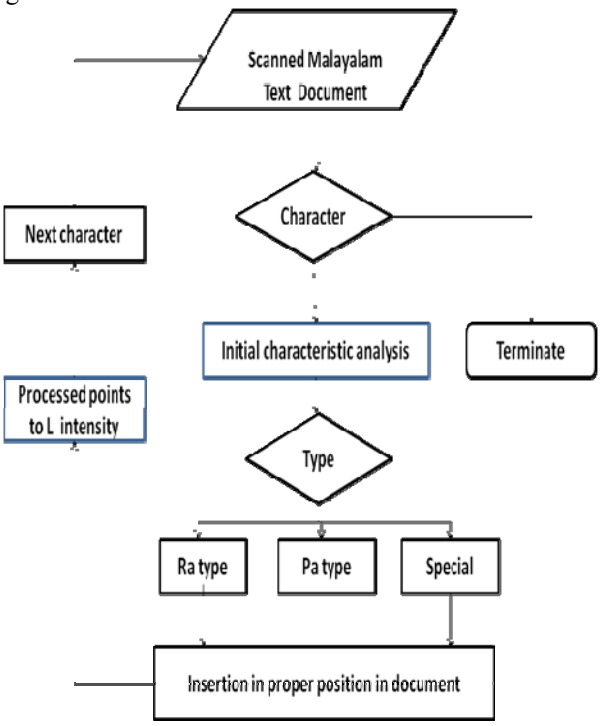

Fig 4. Character Separation Technique.

V. Character Separation

In order to apply the algorithm for recognition of handwritten characters, segregation of scanned images is of prime importance. Rather than adopting the normal projection profile methodology, an alternative technique to identify foreground and background colors of the scanned image was incorporated. This is used to authenticate the letters written in colors other than the identified background color. The H-L notation has been adopted to represent background points and valid character path. The steps followed to segregate and to separate the individual letters from the scanned input image can be postulated as follows.

Step 1: Process the image from the top left, one segment at a time from right extreme to left extreme.

Step 2: A pixel with H intensity is authenticated as a valid point in a character.

Step 3: Horizontal, vertical and diagonal comparisons are devised to identify the constituents of the character

Step 4: The character, hence isolated, may be restructured to a window with special emphasis on boundary points.

Step 5: The dynamic widow is then processed and various checks are applied to dentify the character.

Step 6: Once identification is completed, the character is inserted in the correct position in the sequence identified on comparison with the position coordinates.

Step 7: The processed window is isolated and the rest of the document is scanned for the next character.

Step 8: On encountering the next significant intensity, the above steps are applied and process is repeated until the entire document is processed.

In the case of combinational characters, 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 5. 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.

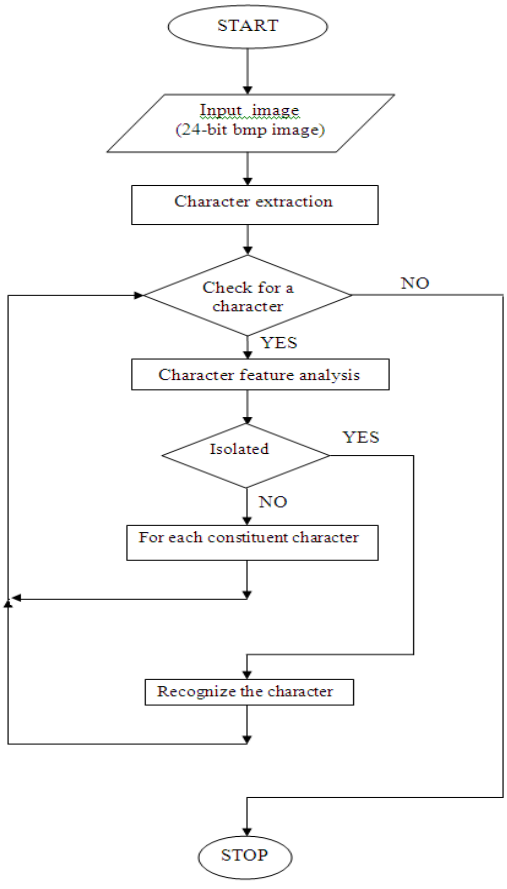

Fig 5. Flow chart of Combinational HCR System

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.

VI. Feature Extration Technique

Once the segregation is accomplished, the feature extraction process is initialized. The length and breadth of each character can be calculated by manipulating the HLH intensity values of the segregated image, which in turn, is stored in a dynamic window matrix. Inferences are arrived at on the basis of the sequence pattern procured on horizontally processing the dynamic matrix. Furthermore, the pattern with highest probability is identified. The matrix is then processed for vertical as well as straight line patterns. Consider the character depicted as in figure 6. The intensity pattern HLH can be observed in the letter and hence infer that two vertical pillars exist on processing the image horizontally. This work separate the entire character set in to three different classes. Ra type characters, Pa type characters and Special symbols. This classification is based on the shape and appearance of the character. This shape feature is extracted to recognize the letter.

Fig 6. Horizontal HLH Patterns.

VII. Recognition Strategy

On analyzing different input samples, it is observed that when the image is processed horizontally, a HLH occurrence as a subset of the sequence pattern is observed. Hence the character is inferred to be "ra" type. After completing the horizontal processing, the character is vertically processed as many times as many recurrences were observed. A HL pattern near the top leads to the inference that there is a probable letter path. The characters are classified as "Pa" type on identifying a vertical or a horizontal line and require special sequence identification checks to recognize the different "ra" type patterns within it. This includes characters which has the characteristics of either "Pa" and "Pa" type characters or both. Special characters require horizontal, vertical and a few diagonal sequence checks to identify the character. Algorithm for recognizing a character is shown below.

1) Identify the characterstic window
2) Apply horizontal check and find the high probablistic reccurning times of HLH patterns
3) Identify and classify the character into "ra", "pa" or special type characters.
4) The view port window is segmented corresponding to recurrences and each segment is further investigated.
5) Vertical processing is applied to each segment to determine the specific intensity sequences.
6) Diagonal cross-sectional analysis may be applied for special characters.

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.

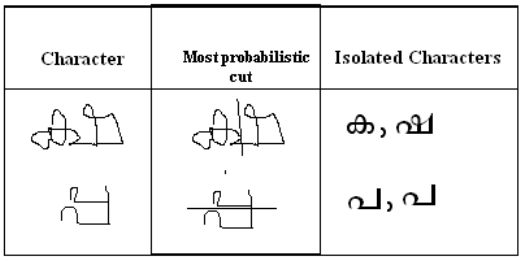

Fig 7. Horizontal & Vertical divisions.

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 go to step10.
Step 7 : Vertical recurrence analysis is carried out. If successful recognition go to step 10.
Step 8 : Special recurrence algorithms are undertaken.
Step 9 : The character stands unidentified.
Step 10 : Stop.

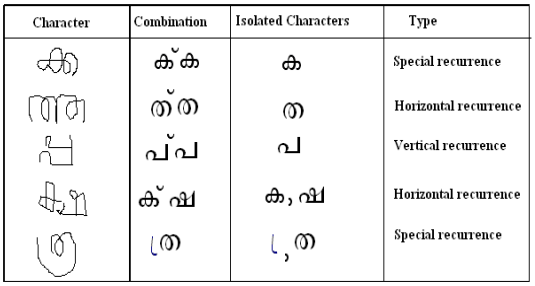

Fig 8. Character combination types.

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. Fig 8illustrates these in detail.

VIII. Conclussion

We conducted the experiment using different lines of text from multiple sources. The samples are mostly from school children who write the new script of Malayalam language. We also tried the samples from standard Malayalam fonts available. The documents were scanned using HP DeskJet F4288 scanner at high resolutions. We also 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 2490 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 and 2. 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 92% has been achieved.

TABLE 1. Experimental Results of Isolated Characters.

Input Document	**Character Analysis**						**Complete Document**	
	Ra Type		***Pa Type***		***Special Type***			
Set	Total Characters	Correctly Recognized	Total Characters	Correctly Recognized	Total Characters	Correctly Recognized	Total Characters	Correctly Recognized
1	164	162	155	148	135	127	454	437
2	202	191	156	148	135	133	493	472
3	199	188	159	154	129	126	487	468
4	155	149	138	131	134	129	427	409
	720	690	608	581	533	515	1861	1786
tion Success	**95.83 %**		**95.55 %**		**96.6 %**		**95.96 %**	

TABLE 2. Performance Analysis of Handwritten Recognition

Input Document	Character Analysis						Complete Document	
	Horizontal Recurrence		Vertical Recurrence		Special Recurrence			
Set	Total Characters	Correctly Recognized	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%**	

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. 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, Aewathy Shajan, Amala Elizabeth and M S Rajasree, " Isolated Handwritten Malayalam Character recognition based on HLH intensity patterns", Proc of International Conf on Machine leraning 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 Conf erence 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.

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 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).

Dr Rajasree M S is the 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 received her M.Tech from NIT Calicut and PhD from IIT Madras. She served as the Principal of Lal Bhahadhur Shastri Institute of Technology for Women, Poojappura Trivandrum Kerala.

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 Universit