

OPTICAL CHARACTER RECOGNITION FOR MALAYALAM

K Jithesh , K G Sulochana & R Ravindra Kumar
RCILTS - Malayalam

Abstract

Optical Character Recognition (OCR) is the process of automatic conversion of scanned images of machine printed or handwritten documents into computer processable codes. It provides the link between the computer and the vast world of the written text. Most of the current OCR systems are designed for European languages using Roman script. Machine Recognition of Dravidian scripts is difficult because of its complex curved form, larger number of basic elements, and the presence of conjuncts that increases the number of glyphs for recognition by an order of magnitude.

We are presently developing a Text Reading System for Malayalam, combining the techniques of OCR and Speech Synthesis. The OCR module operates on the scanned image of the printed page. Its output is the corresponding sequence of Unicodes. It is implemented as a sequence of three stages - pre-processing & segmentation, recognition and post-processing. Pre-processing consists of noise removal and de-skewing. Segmentation successively classifies the image into regions of paragraph, line, word, and character. The segmented characters are fed to the recognition system. It performs feature extraction followed by a binary decision tree classification to generate the code corresponding to the character. Post processing uses linguistic rules to improve the accuracy of the generated code sequence. This paper lists the approaches and algorithms used for the OCR.. It introduces the salient features of Malayalam script. The overall structure of OCR system is presented. The implementation of the sub-modules are described in detail. The system performance figures are presented and the major issues relating to this are discussed.

Key words : *OCR, Malayalam, Thresholding, Segmentation, Feature extraction, Classification, Linguistic rules.*

1. Introduction

Optical Character Recognition is the process of scanning in text from printed material into a text document on the computer. Normal scanning techniques simply create an image of the document, while OCR actually recognizes the print and stores it in an editable text-based format. Thus an OCR system enables one to take a book or a magazine article, feed it directly into an electronic computer file, and then edit the file using a standard Word Processor.

At present keyboarding is the most common method of inputting data into the computer, which is a very complex and labour intensive process. Increased productivity by reducing human intervention and the ability to efficiently store text are two major selling points for an OCR system. An OCR system together with a Speech Synthesizer can be used as a text reading system. Such a system would be useful to read out texts/ information to the illiterate and visually challenged population, constituting a significant portion of society. Other application areas include postal departments, banks and publication industry.

A good OCR system should be capable of recognising characters in different fonts, styles and sizes. The early systems could only process one or two sets of characters in fixed type and size. Modern character recognition methodologies, which use sophisticated techniques, have enabled the recognition of complex characters and symbols. Now many OCR systems are available for European languages using Roman script. The computer recognition of Japanese characters was considered to be a very difficult task, because of the huge size of the character set and the complexity and similarity of the Kanji characters. Presently many Chinese and Japanese OCR systems are available with fairly good accuracy rate.

The primary aim of this paper is to list the approaches and algorithms used for developing the Malayalam OCR by the Resource Centre for Indian Language Technology Solutions – Malayalam (RCILTS-Malayalam), at CDAC, Trivandrum.

2. Malayalam Language and Script

More than 96% of people in Kerala and Lakshadweep Islands use Malayalam, which is the official language of both. The source of the present day Malayalam script is to be traced to one specific style of writing known as *vaTTezhuttu* (round writing). *vaTTezhuttu* was known by different other names, such as *gajavativu* (elephant shape) *cerapandyalipi* *vaTTezhuttu* probably referred to the fact that the letters in this style of writing were rounded in shape. This style later came to be known as *kolezhuttu*. The most significant attempt to reform the Malayalam script so as to make it better suited for easier learning by children, fast type-writing and printing has taken place in 1981. The chief advantage of the new system is the considerable reduction in the number of special letters representing less frequently used conjunct consonants and combinations of vowels with consonants. Instead of 500 different letters required for handling the traditional script, the present system, which is equally efficient, needs only 90 signs. But some of the publications in Malayalam still use a mixture of both. The Malayalam OCR should cater to both old and new scripts alike.

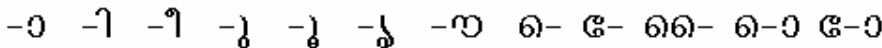
- There are 13 vowels, 36 consonants and 5 pure consonants in the Malayalam alphabet.

അ ആ ഇ ഊ ഉ ഊ ഋ ഡി ഡി ഡെ ഒ ഓ ഔ

ക ഖ ഗ ഘ ങ
ച ഛ ജ ഝ ഞ
ട ള ഡ ഡ ണ
ത മ ദ ധ ന
പ ഫ ബ ഭ മ
യ ര ല വ
ശ ഷ സ ഹ
ള ഴ ഴ

രീ നീ ശീ ണീ തീ

- The dependent vowels in Malayalam are represented by separate symbols. There are 12 vowel signs. The vowel signs can occur agglutinated with consonants or it can occur independently

-  and Vakar. These are formed when the Malayalam letters YA, RA/RRA, LA and VA occur at the end of a consonant / consonant cluster.

ക ള ല

- A crescent mark called Chandrakkala is the vowel omission sign in Malayalam. Anuswaram and visargam are two other signs present in the Language.

ഃ ള

- Two types of compound (conjuncts) characters occur in Malayalam - one vertically compounded and the other horizontally compounded.

ക ള സ ഷ സ ബ ജ ല വ സ സ സ
 ത ത ധ ധ ധ ത ത സ ത ധ മ ജ
 ജ ജ ജ ക ക ക ക ള ള ള ള ള
 ധ ധ ധ ധ ധ ധ ധ ധ ധ ധ ധ ധ
 ഷ ഷ ഷ ഷ ഷ ഷ ഷ ഷ ഷ ഷ ഷ ഷ
 സ സ സ സ സ സ സ സ സ സ സ

ച ട പ ബ ണ ജ ശ സ ഗ യ
 റ ള വ ട ക ള ള ള ള ള ള
 ള സ സ ക ബ ബ ച ന്നു

ക ള ള ത ത ത ത ത ത ത ത ത
 ന വ ധ ധ ധ ധ ധ ധ ധ ധ ധ ധ
 ത ത ത ത ത ത ത ത ത ത ത ത
 ധ ത ത ത ത ത ത ത ത ത ത ത

3. The Malayalam OCR System

The three major blocks of the CDAC OCR system are:

1. Pre-processing Block
2. Recognition Block
3. Post-processing Block

Block diagram shown in Fig.1

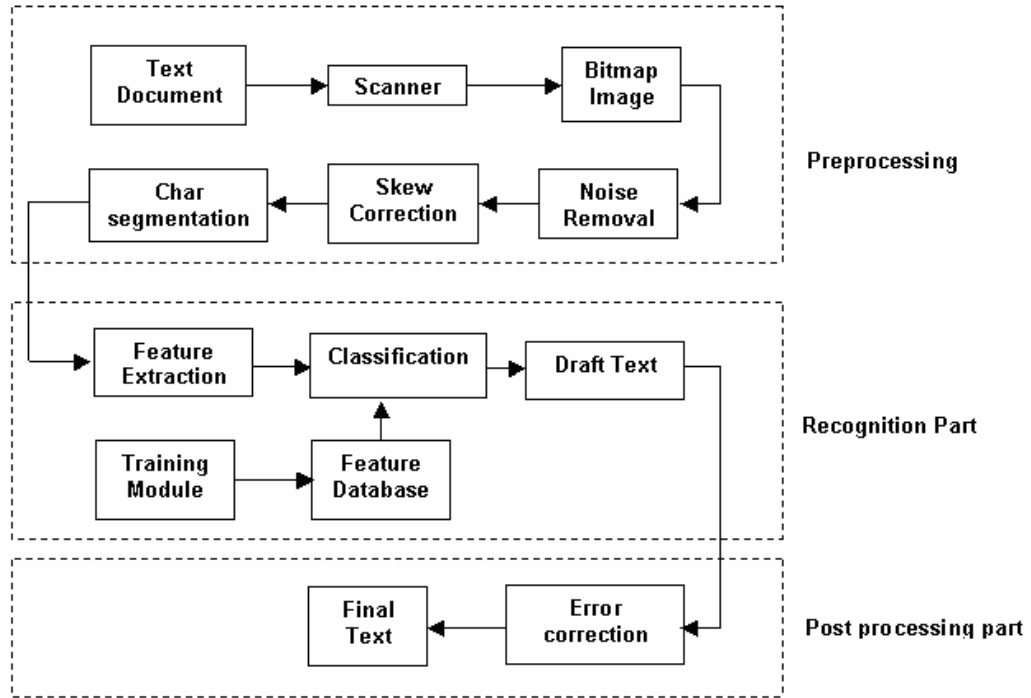


Fig. 1

3.1 Pre-processing

In pre-processing, the image scanner optically captures the text images to be recognized. The bitmap file thus obtained is operated upon by a series of algorithms and transforms to obtain the individual characters segmented. We also try to compensate for poor quality documents and images and apply correction for image skew.

3.1.1 Gray scale to Binary Image conversion and Noise removal

The scanned images in gray tone are converted into two-tone (binary) images using a histogram based thresholding approach. The simplest property that pixels in a region can share is intensity. So, a natural way to separate light and dark regions is through thresholding. Thresholding creates binary images from gray-level ones by turning all pixels below some threshold to zero and all pixels about that threshold to one.

If $g(x, y)$ is a thresholded version of $f(x, y)$ at some global threshold T ,

$$g(x, y) = \begin{cases} 1 & \text{if } f(x, y) \geq T \\ 0 & \text{otherwise} \end{cases}$$

Two types of thresholding methods are available one the Global thresholding (one threshold for entire image) and the other Local thresholding (different threshold for different image regions). The algorithm used here is the Otsu's algorithm. Otsu's algorithm chooses the optimal threshold by maximizing between class variance of pixel intensity. The Sahoo et al. study on global thresholding, concluded that Otsu's method was one of the best threshold selection methods for general real world images. We also tried some other thresholding methods, but the Otsu's method gave a better result. After thresholding the image may contain small holes in dark areas, small

notches and bumps in straight edge segments and isolated black pixels. For correcting these problems we used the logical smoothing approach.

3.1.2 Skew Detection and Correction.

The skew introduced (if any) during the scanning process is detected and corrected in this step. There are several methods for detecting skew in a page; some are based on the projection profile of the document, some rely on detecting connected components and finding the average angles connecting their centroids and some others use the Hough Transform. Projection profile based technique is one of the popular skew estimation techniques and we have used this method. The method involves projecting the page at several angles, and determining the variance in the number of black pixels per projected line. The projection parallel to the true alignment of the lines will likely have the maximum variance, since when parallel, each given ray projected through the image will hit either almost no black pixels or many black pixels. After estimating the skew angle the skew is corrected by rotating the image against the estimated skew angle.

3.1.3 Text Segmentation

Text analysis techniques are applied to segment the text image into lines, words and characters. For segmentation we used the histogram-based technique.

Line, word and character segmentation

To segment individual lines of a document image a horizontal projection profile is built by counting the number of black pixels in each horizontal row of pixels. The position between two consecutive lines where the histogram value is least is selected as the boundary of a line. Fig. 2 shows a document image and its horizontal histogram.

After line segmentation a vertical projection profile is built for each segmented line. Here the black pixels are counted from each pixel columns, if count is less than the threshold value it is considered as zero. If the projection profile contains n consecutive zeros then the mid point of that run is considered as the boundary of a word. Similarly we can segment words into characters. The value of n will vary for character and word segmentation.

Fig.2

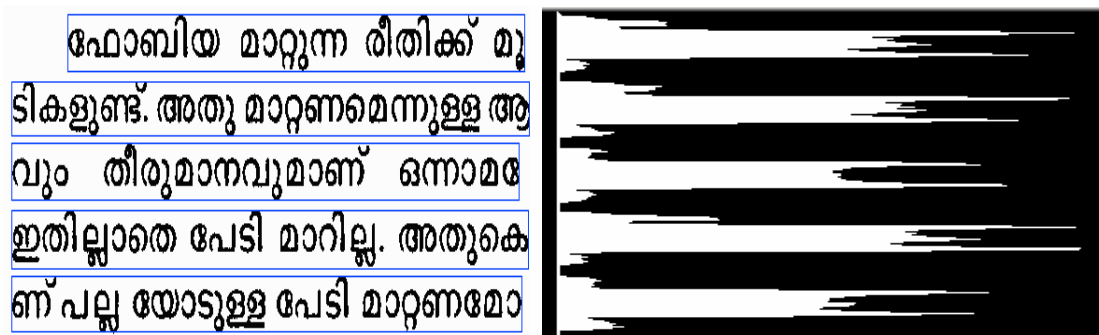


Fig.3



The above picture shows the vertical histogram of a text line. Certain dependent vowel/consonant signs will overlap the independent vowel/consonant with which it occurs and hence they will be segmented as a single unit. For such clustered characters we applied the component-labeling algorithm to separates the components.

3.2 Character Recognition

This is the major module in the OCR system. There are two basic methods used for character recognition namely Matrix Matching and Feature Extraction. Of these two methods, matrix matching is the simpler and more common. Matrix Matching compares what the OCR scanner sees as a character with a library of character matrices or templates. When an image matches one of these prescribed matrices of dots within a given level of similarity, the computer replaces the image with the corresponding character code. Matrix matching works best when the OCR encounters a limited font size or styles.

Feature Extraction method of character recognition, also known as Intelligent Character Recognition (ICR), or Topological Feature Analysis, is based on features specific to the character. Here the system looks for general features such as closed loops, vertical lines, end points, etc. to identify individual characters. This method is much more versatile than matrix matching. Our Malayalam-OCR System is based on feature extraction method.

The character recognition module extracts features from the segmented characters and these features are input to a binary tree classifier. After passing through different stages of the classifier, the character is identified and corresponding character code is assigned.

3.2.1 Character Grouping.

In a Malayalam text document more than 60% of characters are basic characters or horizontal compound characters. To differentiate the above group of characters from others, all character sub images are put into different bins depending on their height. The bin with maximum number of characters is identified, and the height of this group of characters will be set as the normal character height. Characters whose height is greater than normal height are considered as vertical compound characters or dependent vowel signs. The Characters with less than the normal height (Chandrakkala, full stop, Coma, etc.) are contained in another bin.

3.2.2 Feature Extraction and classification

Feature extraction can be considered as finding a set of vectors, which effectively represent the information content of a character. We have selected some features based on topology, stroke, endpoint and run number for the classification of Malayalam characters. These are identified after a careful study of Malayalam writing system. The topological and

stroke based features are used for the initial classification of characters. They are selected with following considerations

- Character structure.
- Reliability of the feature
- Font and size independence and
- Speed of feature extraction

The selected features are classified into three different groups, Type-1, Type-2 and Type-3. Of these, Type-1 and Type-2, which are independent of font size and style, are the main features that help identification of the characters.

Type-1 Features

- i) Vertical bars
Vertical Left Bar, Vertical Right Bar, Vertical Mid bar, Left Part Bar and Right Part Bar
- ii) Horizontal bars
Horizontal Bottom bar

Type-2 Features

- i) Loops.
Number of loops, Loop height and Loop position.
- ii) Aspect ratio.

Type-3 Features

- i) End points
Number and location of end points.
- ii) Run Number based feature.

The classification module uses the features extracted from the character sub images. Two popular classification methods are available one binary tree classifier and the other shortest distance classifier. We used the binary tree classification method. The binary tree classifier is a hierarchically based classifier, which compares the data with a range of properly selected features. The selection of features is done by an assessment of the separability of the classes. Therefore each decision tree or set of features should be designed carefully. The advantages of the decision tree classifier are that computing time is less than the maximum likelihood classifier and by comparison the statistical errors are avoided. However the disadvantage is that the accuracy depends fully on the design of the decision tree and the selected features. Using the extracted features the binary tree will classify characters into small subsets. Most leaf nodes of the classification tree contain two or three characters. At the leaf nodes we used the Run number based feature to identify characters.

All recognized characters, which are labeled with corresponding character codes, are converted to UNICODE, ISCII or ISFOC coded text depending on the user inputs.

3.3 Post-processing part

The post-processing module corrects the mistakes, which occurred during the recognition stage. Linguistic rules are applied to the recognised text to reduce classification errors. For example, certain characters never occur at the beginning of a word and if found so, they are remapped appropriately. Similarly, dependent vowel

signs can occur only with consonants or consonant conjuncts; if found along with vowels or soft consonants, they are remapped into consonants/conjuncts similar in shape to the vowel sign. Independent vowels occur only at the beginning of a word and if found anywhere else, they will be mapped into a consonant or ligature having similar shape.

4. Performance of the OCR

Our Malayalam OCR recognises 50 characters per second and gives an accuracy of 97% for good quality printed documents. The specifications and performance of the system is given below.

Skew detection and correction : - 5 to +5 degree
Supported image formats : BMP, TIFF
Image scan resolution : 300dpi and above.
Document Type : Single-font single size.

Supporting Fonts:

Fonts Names : CDAC Fonts (ML-TTKarthika, MLW-TTKarthika) Mathrubhumi Font, Manorama Font, Fonts used by DC Books
Font Size : 12-20
Font Styles : Normal, BOLD
Supported Code format : ISCII/ISFOC
Supported output format : RTF/HTML/ACI/TXT

Character recognition accuracy (%)

Document Type	Good quality Paper	Bad quality Paper
Computer Printed Document	97%	94%
Magazine	92%	90%
Newspaper	85%	82%
Books	95%	93%

Number of samples tested > 500

5. Conclusion

The Feature Extraction methodology adopted for character recognition has been able to provide very good accuracy for a wide range of font sizes and styles. The maximum accuracy obtained is 97%. The major factor hindering the improvement of accuracy is the similarity in character shapes and features of certain Malayalam characters. Further refinement of the system is possible by training the OCR Engine to handle commonly encountered errors. Incorporating a lookup table of commonly used words will also enable the system to correct wrong recognitions. Presently we are working in this area.

6. Acknowledgements

We are indebted to the TDIL Programme of DIT particularly Dr.Om Vikas, Senior Director and Head, TDIL programme and Dr.P.K.Chaturvedi, Director for their inspiration and nurturing approach to Indian Languages. We would like to express our sincere thanks to Dr.V.R.Prabodhachandran Nair, eminent linguist and Dr.Usha Nambudiripad, our consultant linguist for their valuable help in the identification of characters in old Malayalam script.

7. References

- (1) "Digital Image Processing" By Rafael C Gonzalez, Richard E Woods.
- (2) "Digital Image Processing and Analysis" B.Chanda , D.Dutta Majunder.
- (3) "A Complete Bangala OCR System" Pattern Recognition Vol.31.PP 531-549 B.B.Chaudahri and U.Pal.
- (4) "A Complete Oriya OCR System" Pattern Recognition Vol.27.PP 23-34 B.B.Chaudahri and U.Pal and M Mitra.
- (5) A Fast Algorithm for Multilevel Thresholding, Ping-Sung Liao, Tse-Shen and Pau-Choo Chung, Journal of Information Science and Engineering 17, 713-727 (2001).
- (6) "A Gurumukhi OCR System " By G.S Lehal and Chandan Singh
- (7) "Skew Detection and Correction" By Katherine Marsden.
- (8) "Malayalam for Beginners" By Dr.V.R.Prabodhachandran Nayar.