Volume 3, Issue 6, June 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

OCR System for Complex Printed Kannada Characters

Mr.Nithya.E
Research Scholar, Dept of CSE,
DayanandaSagar College of Engineering,
Bangalore-78, *VTU Belgaum-14, India*

Dr. Ramesh Babu D R
Research Guide, Prof. & HOD, Dept. of CSE,
DayanandaSagar College of Engineering,
Bangalore-78, *VTU Belgaum-14, India*

Abstract— Optical Character Recognition (OCR) is the process of converting the textual image into the machine editable format. This paper proposes an OCR system for Complex printed Kannada Characters. The input to the system would be the scanned image of a page of text that containing complex Kannada characters and the output is a machine editable file. The system first pre-processes the input document containing the complex Kannada characters and converts it into binary form. Then the system extracts the lines from the document image and segments the lines into character and sub-character level pieces. Here histogram technique and connected component method is used for character segmentation and correlation method is used to recognize the characters. Here first we are collecting different sample characters and it is pre-processed and stores it in a file. The input image is segmented to character level pieces and it is compared with sample characters. It returns corresponding target ID. Each target ID has corresponding character class name. Then we are displaying the class name, which is in machine editable format.

Keywords— Correlation method, Connected component method, Histogram Technique.

I. Introduction

Optical Character Recognition is one of the oldest sub fields of pattern recognition with a rich contribution for the recognition of printed documents. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. Though academic research in the field continues, the focus on OCR has shifted to implementation of proven techniques.

When one scans a paper page into a computer, it produces just an image file, a photo of the page. The computer cannot understand the letters on the page, so you cannot search for words or edit it or change the font, as in a word processor. You would use OCR software to convert it into a text or word processor file so that you could do those things. The result is much more flexible and compact than the original page photo. The need for OCR arises in the context of digitizing the documents from the library, which helps in sharing the data through the Internet.

II. Proposed Methodology

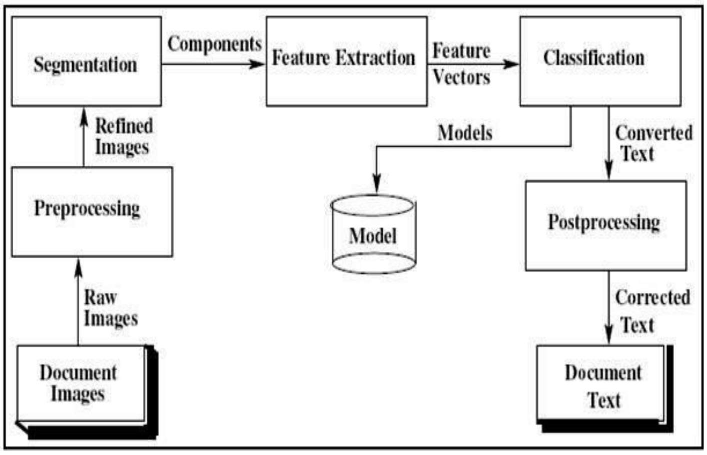

Fig 2.1 Proposed method

The main steps involved in achieving OCR are as follows:

- Preprocessing
- Segmentation
- Character recognition
-

III.Pre-processing

The input to the system is a digital image of the document containing printed complex Kannada text captured by scanning the document using a flatbed scanner or digital camera. The input documents in RBG format. First it is converted to Grayscale image then we are calculating the threshold value of the grayscale image and by using that value we are converting that image to black and white format. At the end we are storing at image in a matrix, say binary form.

IV. Segmentation

A. Line Segmentation

To separate the text lines, from the document image, the horizontal projection profile [1] of the document image is found. The horizontal projection profile is the histogram of the number of ON pixels along every row of the image. White space between text lines is used to segment the text lines. Figure 4.1 shows a sample Kannada document along with its horizontal projection. The projection profile will have valleys of zero height between the text lines. Line segmentation is done at these points.

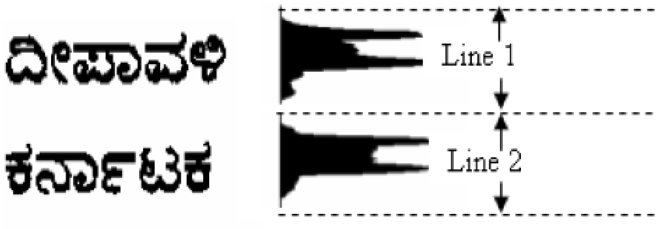

Fig 4.1 Line segmentation

B. Character segmentation

The letters in Kannada are composed by attaching to the glyph of a consonant, the glyphs of the vowel modifiers and the glyphs of the consonant conjuncts. If we considered all the combination, then building the classifier of these numbers of character is very difficult. So our strategy is that we will segment the word into its constituents, i.e. the base consonant, the vowel modifier and the consonant conjunct. It's very difficult to achieve this. If we have good look on the Kannada word, we will see that for extracting glyph of a consonant the glyphs of the vowel modifiers and the glyphs of the consonant conjuncts we can divide the character into two zones [2].

- **Top zone:** Top zone mainly consist of main portion of the character. It includes base consonant or vowels or some vowel modifiers.
- **Bottom zone:** Bottom zone consists of glyphs for the consonant conjuncts.

Here by using connected component method [3] we are first counting the number of consonants, vowels, vathu or vowel modifiers present in the text line. Then we extracting that characters separately and send it for character recognition.

V. Character Recognition

In character recognition phase, it takes each individual character as input. Here we are using correlation method [4] to compare the input image character and stored sample character. Before applying correlation method, the size of input character and stored character size should be same. So we have to resize the input character image. For that purpose we are using Nearest Neighbor Interpolation method. Then Input image is compared with the stored character and it will return corresponding correlation coefficient. The maximum correlation coefficients target ID returned to main application. Each target ID has corresponding class name.

The output after classification has to be transformed into a format, which can be loaded into a Kannada editing package. The method of composition of aksharas in all Kannada typesetting packages is similar. The string representing an akshara is composed from different character class names corresponding to the different components of the akshara as follows: the codes for the base consonant appear first followed by the codes for the consonant conjuncts; the codes for the vowel modifier appear at the last and signify the end of an akshara. Here we are displaying that character class name in open source “Baraha” editor. If it is complex character, letter concatenation takes place before displaying in editor. From “Baraha” editor we will export the character, which will be displaying in notepad.

VI. Application

Some of the major applications of OCR include:

- Library and office automation,
- Form and bank check processing,
- Document reader systems for the visually impaired,
- Postal automation, and
- Database and corpus development for language modeling, text-mining and information retrieval

VII. Results and Discussion

We measured the performance of our system by scanning document that contains different complex Kannada characters. We collected nearly fifty different samples that include vowels, consonants, vathu and vowel modifier. The complex Kannada character means, which is the combination of vowel or consonants or vathu or vowel modifier. The system first segments the document into character level pieces and it is compared with the sample characters. It recognized more than hundred Complex Kannada character and more than hundred complex Kannada words. Here the system gives more than 90% accuracy.

The screen shots of the system shown in the following figures.

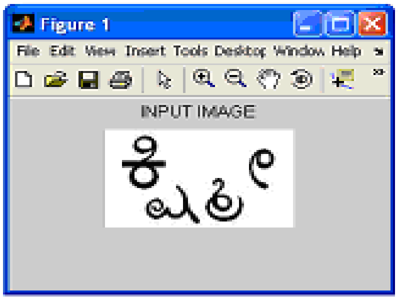

Fig6.1 Input printed complex Kannada Character

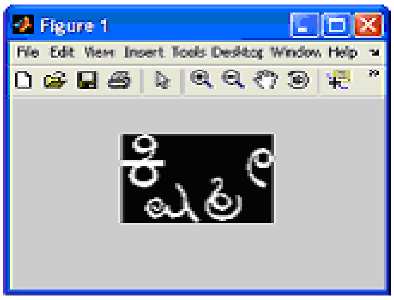

Fig 6.2 Preprocessed Character

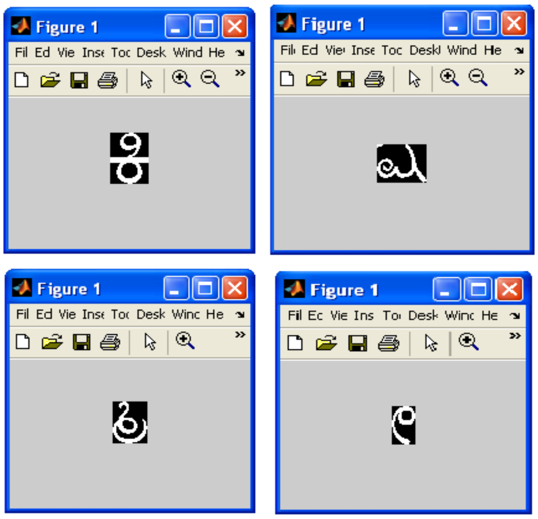

Fig 6.3 Segmented characters

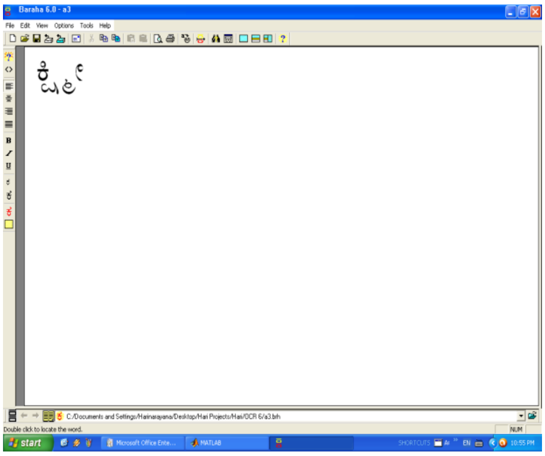

Fig 6.4 Character recognition and displaying

VIII. Conclusion

This paper describes a simple and efficient OCR system for printed text documents in Kannada, a South Indian language. It takes complex Kannada character as input image and converts it into machine editable format. The system is designed to be independent of the font and size of text. At the end, the paper shows some results with the system, which delivers reasonable character recognition accuracy.

REFERENCES

[1] R SANJEEV KUNTE and R D SUDHAKER SAMUEL, “A simple and efficient OCR, for basic symbols in printed Kannada text”, Sadhana Vol. 32, Part 5, October 2007, pp. 521–533. © Printed in India.

[2] T V ASHWIN and P S SASTRY, “A font and size-independent OCR system for printed Kannada documents using support vector machines”, Sadhana Vol. 27, Part 1, February 2002, pp. 35–58. © Printed in India.

[3] Gonzalez R C, Woods R E 1993, “Digital image processing”, (Boston, MA, USA: Addison Wesley Longman Publishing Co. Inc.)

[4] C. Balletti F. Guerra, “Image matching for historical maps comparison”, e-perimetron, Vol. 4, No. 3, 2009[180-186].