OCR for printed Kannada text to Machine editable format using Database approach

B.M. SAGAR[1], Dr. SHOBHA G[2], Dr. RAMAKANTH KUMAR P[3]
Information Science[1], Computer Science[2], Computer Science[3]
Visvesvaraya Technological University[1, 2, 3]
Lecturer[1], Professor[2], Professor[3], R.V.C.E, Bangalore-59, Karnataka, INDIA
sagar.bm@gmail.com[1], shobhatilak@rediffmail.com[2], pramakanth_2000@yahoo.com[3]

Abstract: - This paper describes an Optical Character Recognition (OCR) system for printed text documents in Kannada, a South Indian language. The proposed OCR system for the recognition of printed Kannada text, which can handle all types of Kannada characters. The system first extracts image of Kannada scripts, then from the image to line segmentation then segments the words into sub-character level pieces. For character recognition we have used database approach. The level of accuracy reached to 100%.

Key-words: - Optical Character Recognition, Segmentation, Kannada Scripts

1. Introduction

Optical character recognition (OCR) refers to reading text from paper and translating the images into a form that the computer can manipulate. OCR systems have been effectively developed for the recognition of printed characters of non-Indian languages. Until quite recently, the focus of this endeavor has been on characters of English Language. Such systems are also available for many European languages as well as some of the Asian languages such as Japanese, Chinese, etc. However, there are not many reported efforts at developing OCR systems for Indian languages especially for a South Indian language like Kannada.

Section 3 describes work done on Kannada character recognition. Section 4 describes the segmentation process of line, word and character. Section 5 describes the proposed system for the Kannada character recognition. Section 6 describes the method of character recognition with the increased efficiency. Section 7 describes the experimental results and then conclusion and future work.

1.1 Introduction to Kannada Scripts

Kannada is one of the South Indian languages. The Kannada character set is almost identical to that of other Indian languages. It is written horizontally from left to right and the concept of lower and upper case is absent. **[1]**

Kannada language has 16 vowels and 34 consonants as the basic alphabet of the language. The number of written symbols, however, is far more than the 50 characters, because different characters can be combined to form compound characters (ottaksharas).

2. Background Study

Due to the impact and the advancements in the Information Technology, nowadays more emphasis is given in Karnataka to use Kannada at all levels and hence the use of Kannada in computer systems is also a necessity. Therefore, efficient OCR systems for Kannada are one of the present day requirements. Currently there are many OCR systems available for handling printed English documents with reasonable levels of accuracy **[1].** It is difficult to find OCR systems for Kannada with the increased accuracy. Few researchers are worked on Kannada character recognition with novel set of features for the recognition problem which are computationally simple to extract. The recognition achieved by employing a number of 2-class classifiers based on the

Support Vector Machine (SVM) method. The recognition is independent of the font and size of the printed text and the system is seen to deliver reasonable performance **[3].**
Another researcher who worked on Kannada character recognition with Hu's invariant moments and Zernike moments. Those are used in the system to extract the features of printed Kannada characters. Neural classifiers have been effectively used for the classification of characters based on moment features. An encouraging recognition rate of 96·8% has been obtained **[1].**
For a printed Tamil character Recognition the ability of the neural network been used. Neural Network is also applied to Text block identification. Recognition varied from 94 to 97% for a particular type of font. **[6]**

In our system we have used database approach for the character recognition. Section 5 describes the method of character recognition with the increased efficiency.

3. Segmentation Process

Due to the peculiarities of the Kannada script, the following segmentation scheme is proposed where lines are segmented then words and finally characters. These are then put together to the effect of recognition of individual aksharas or characters.
As Kannada is a non-cursive script, the individual characters in a word are isolated. Spacing between the characters can be used for segmentation.

3.1 Line Segmentation

Line segmentation is the process of identifying lines in a given image.
Steps for the line Segmentation is as follows

1. Scan the BMP image horizontally to find first ON pixel and remember that y coordinate as y1.
2. Continue scanning the BMP image then we would find lots of ON pixel since the characters would have started.
3. Finally we get the first OFF pixel and remember that y coordinate as y2.
4. y1 to y2 is the line.
5. Repeat the above steps till the end of the image.

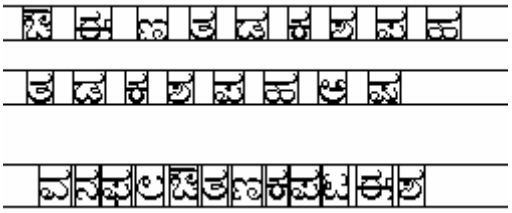


Figure: 3.1 Shows the line and character segmentation

3.2 Word Segmentation

As we know that there is a distance between one word to another word. We use that concept for word segmentation. After the line segmentation scan the image vertically for word segmentation.
Steps for the word Segmentation is as follows
1. Scan the BMP image vertically for the recognized line segment, to find first ON pixel and remember that x coordinate as x1. Treat this as starting coordinate for the word.
2. Continue scanning the BMP image then we would find lots of ON pixel since the word would have started.
3. Finally we get the successive five (this is assumed word distance) OFF pixel column and remember that x coordinate as x2.
4. x1 to x2 is the word.
5. Repeat the above steps till the end of the line segment.
6. Repeat the above steps for all the recognized line segments.

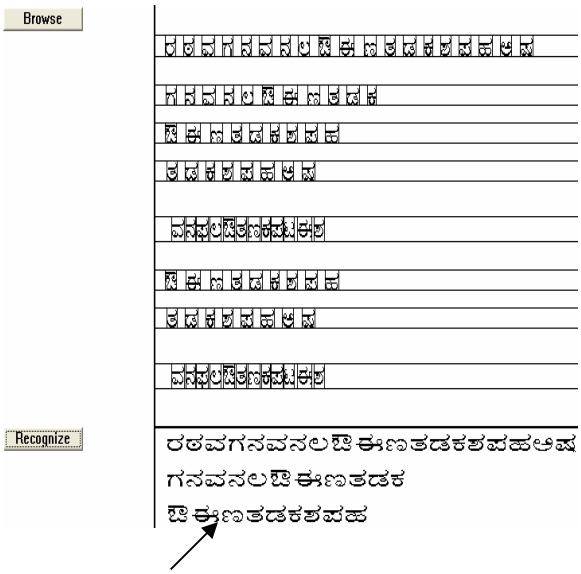
3.3 Character Segmentation

1. Scan the BMP image vertically for the recognized word segment, to find first ON pixel and remember that x coordinate as x1. Treat this as starting coordinate for the character.
2. Continue scanning the BMP image then we would find lots of ON pixel since the characters would have started.
3. Finally we get the OFF pixel column and remember that x coordinate as x2.
4. x1 to x2 is the character.

5. Repeat the above steps till the end of the word segment, line segment.
6. Repeat the above steps for all the recognized line segments.

4. Proposed system

The OCR's task is to identify the characters of Kannada script and the word processor provides an interface for viewing and editing documents in Kannada. Figure 4.1 shows the details. In this work, the sequence of operations carried out is as follows. A page of Kannada text is scanned. The image format used is the bmp format. The input to the system is a scanned image file in BMP format of pure Kannada document. The document is then segmented into lines and each line into individual characters. The documented is scanned and a line in the image file is extracted. The extracted line is given as input to the Character Segmentation. Within each line the characters are segmented one by one. The extracted character that is still to be recognized is given as input to the Character Recognizing Module.



Output

Figure: 4.1 shows interface for viewing and editing documents in Kannada.

5. Character Recognition

After we got the character by character segmentation we store the character image in a structure. This character as to be identified for the pre defined character set.
There will be preliminary data will be stored for all the kannada characters for a identified font and size. This data contains the following information

1. Character ascii value
2. Character name
3. Character BMP image
4. Character width and length
5. Total number of ON pixel in the image.

For every recognized Character above mentioned information will be captured. The recognized character information will be compared with the pre defined data which we have stored in the system.
As we are using the same font and size for the recognition there will be exact one unique match for the character. This will identify us the name of the character.
If the size of the character varies it will be scaled to the known standard and then recognizing process will be done.

6. Experimental Results

Figure 4.1 shows the input to the system and once we say *recognize* we get the output at the bottom.
Since we are using database approach for the character recognition we get 100% accuracy. But the limitation for this approach is that for each character we need to have details like Character ASCII value, Character name, Character BMP image, Character width, length and total number of ON pixel in the image. This takes lot of space as well as lot of computation involved in recognizing the character. But we get 100% accuracy.

8. Conclusion & future work

In this paper, we have presented a database approach for recognizing Kannada characters.
Kannada is widely used language in South India. Lots of applications need Kannada

OCR which can give 100% accuracy. The database approach shows the required accuracy but with the above said limitation. Using Neural Network, Support Vector Machine recognition work can be carried out but not with the required accuracy. But we can make use of dictionary approach to increase the accuracy.

Reference:

[1] R SANJEEV KUNTE and R D SUDHAKER SAMUEL "A simple and efficient optical character recognition system for basic symbols in printed Kannada text" by

[2] "Hidden Markov Models for Online Handwritten Tamil Word Recognition" Bharath A, Sriganesh Madhvanath, HP Laboratories India HPL-2007-108, July 6, 2007

[3] T V ASHWIN and P S SASTRY "A font and size-independent OCR system for printed Kannada documents using support vector machines", Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India

[4] Rohana K. Rajapakse, A. Ruvan Weerasinghe "A Neural Network based character recognition system for Sinhala Script" , Department of Statistics and Computer Science, University of Colombo

[5] SEETHALAKSHMI R "Optical Character Recognition for printed Tamil text using Unicode", Thanjavur, Tamil Nadu

[6] K.H. Aparna and V.S. Chakravarthy "A Complete OCR system Development of Tamil Magazine Documents" Dept of Electrical Engineering, IIT Madras.

About the authors

B.M.Sagar, Lecturer of Department of Information Science and Engineering. He obtained his Master's Degree in Computer Science & Engineering from VTU and B.E. in Computer Science & Engineering from VTU. His research interests are Pattern Recognition. He has guided more than 25 under graduate projects. He has presented and published papers at national conference / International Conference.

Dr. Shobha G., Professor of Computer Science & Engg. She has been awarded Ph.D for her thesis titled "Knowledge Discovery in Transactional Database Systems" from Mangalore University, Mangalore. She obtained her M.S. degree in Software Systems from BITS, Pillani and BE in Computer Science from Gulbarga University. Her research interests are Data Mining, DBMS, and Operating Systems & Networking. She has guided more than 30 undergraduate and 09 post graduate projects.

Dr. Ramakanta Kumar, P was awarded Doctorate from Mangalore University, has teaching experience of around 14 years in academics and Industry. His area of research is on Artificial Intelligence, Pattern recognition. He has to his credits 03 National Journals, 02 International Journals, 12 Conferences and 15 Research Publications. He is guiding 04 MTech students and 03 Phd students.