Devanagari OCR Using Projection Profile Segmentation Method

Mr. Akshay J. Kant[1], Dr. Mrs. Arati J. Vyavahare[2]

[1]Student, E&TC Engineering, Modern College of Engineering, Maharashtra, India
[2]Professor, Dept. of E&TC Engineering, Modern College of Engineering, Maharashtra, India

---***---

Abstract -*Optical character recognition (OCR) is electronic translation of scanned images of printed or handwritten text into machine editable text. In India, for documentation Devanagari script is used by more than 300 million people. There are lots of research done on character recognition of Devanagari script. In OCR there are main two types of documents processed. Typed text that is machine printed and handwritten text. Character recognition of machine printed text is easy as compared to handwritten text. Because in typed text the font size and text quality is good as compared to handwritten text. In handwritten the style of writing vary from person to person, so it is very difficult to recognize the characters. There are so many languages covered under the Devanagari script like Sanskrit, Hindi, Marathi, Nepali, etc. In the proposed work the main focus is on Marathi character recognition. The process of OCR is first input image which is handwritten text is scanned and pre-processed. The pre-processed image is segmented into lines, lines into words and words into characters. The features are extracted from segmented characters using classifiers. Finally post processing where recognized image is converted into editable text.*

Key Words:Devanagari, OCR, Handwritten.

1.INTRODUCTION

Since the evolution of digital computer the interaction between machine learning and human computer is most challenging research field. In optical character recognition (OCR), the input image which is handwritten text is pre-processed first. In this stage the input image is converted into grayscale image and then to binary image using different method like Otsu method. If any noises are present in image then it get remove in this stage. We have to process image clean and clear for next step. In next step i.e. segmentation the image is broken into characters. The image is set of lines, words and characters. In segmentation step each line, word and characters are separated. This step is very important. As if we get proper character segmented image then it helps to recognize image properly. The next classification step, in this the segmented characters are matched with stored database characters set. The matching process like feature based, edge based, and pattern based, etc. Final step is post processing, in this the matched image is processed to output character in text file using ASCII code.

2.FEATURES OF DEVNAGARI SCRIPT

In India there are so many languages which are used in daily basis like Hindi, Marathi, Sanskrit, etc. In most of the language Devanagari script is used. Devanagari script has 11 vowels and 33 consonants. These are basic characters. The Devanagari script is written in left to write format.

The word is divided into three parts. The upper part is called header part. All characters have horizontal line in upper part known as header line or Shirorekha. The middle part which contains actual body of characters. And the lower part. The following diagram shows all the three part of the word.

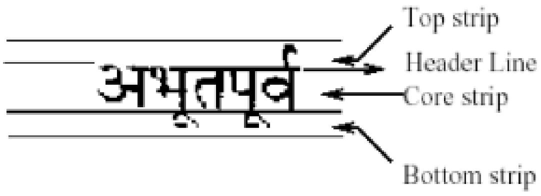

Fig -1: Three zones of Devanagari word

3. PROPOSED OCR SYSTEM

Following steps have been followed in the design of proposed OCR system:

- Preprocessing
- Segmentation
- Feature Extraction
- Classification.

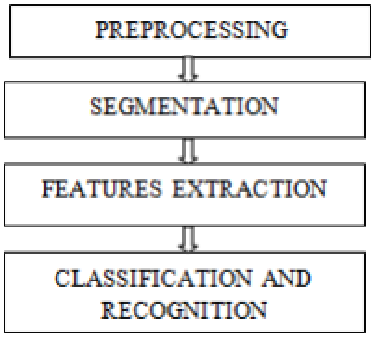

Fig -2: Steps of OCR system

3.1Pre-processing

The handwritten document is scanned as input image. The input image is converted into gray scale image. The gray scale image is converted into binary image using Otsu's thresholding method. In Otsu's thresholding method the threshold value is selected using Otsu's algorithm. In binary image the value below threshold value is converted 1 and above threshold it is 0. The image is then inverted to represent background pixel by 0 and object which are characters represent by 1.

The image noise is reduced by Gaussian filter to get noiseless image. Various morphological processing techniques are carried out on image to get clear and enhanced image.

The handwritten characters are tilted or skewed. We need proper characters for segmentation and classification. It should be eliminated because it can reduce the accuracy of the subsequent processes.

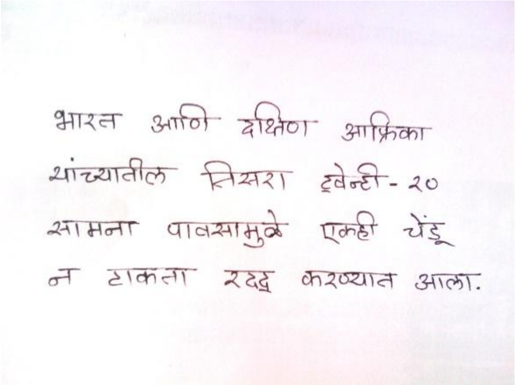
भारत आणि दक्षिण आफ्रिका
यांच्यातील तिसरा ट्वेन्टी-२०
सामना पावसामुळे एकही चेंडू
न टाकता रद्द करण्यात आला.

Fig -3: Input image

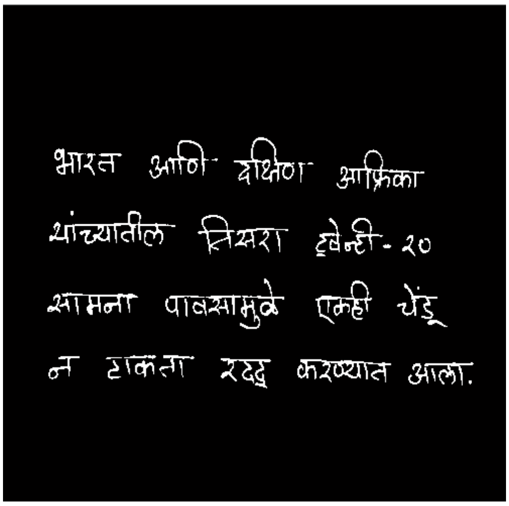
भारत आणि दक्षिण आफ्रिका
यांच्यातील तिसरा ट्वेन्टी-२०
सामना पावसामुळे एकही चेंडू
न टाकता रद्द करण्यात आला.

Fig -4: Pre-processed image

3.2. Segmentation

After image is pre-processed the next step is segmentation. In segmentation there are three steps 1. Line detection and segmentation. 2. Word detection and segmentation and 3. Character detection and segmentation.

In this experiment we are using projection profile method for segmentation.

3.2.1 Line Segmentation

The horizontal projection profile method is used to calculate sum of all white pixels on every row. It gives corresponding histogram of that image. The steps of line segmentation are as follows

1. Construct the horizontal histogram of image.
2. Find the threshold value which is gap between corresponding two histogram.
3. Separate each histogram by threshold value and save it
4. We get segmented lines from image.

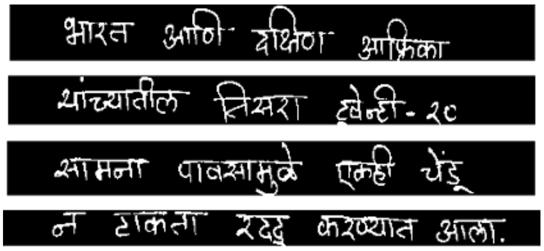
भारत आणि दक्षिण आफ्रिका
यांच्यातील तिसरा ट्वेन्टी-२०
सामना पावसामुळे एकही चेंडू
न टाकता रद्द करण्यात आला.

Fig -5: Segmented lines

3.2.2 Word Segmentation

In word segmentation we use vertical projection profile method to calculate sum of all white pixels. Plot the histogram of computed white pixels. The steps of word segmentation are as follows:

1. Construct the vertical histogram of the image.
2. Find the threshold value
3. separate corresponding histogram from each other and save it
4. The saved images are segmented words from line
5. Repeat process for each line

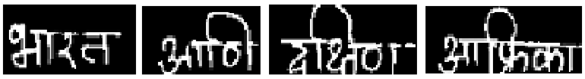

Fig -6: Segmented words

3.2.3 Character Segmentation

In character segmentation we use both horizontal and vertical projection profile method to separate characters from words. The steps of character segmentation are as follows:

1. Construct the horizontal histogram of the image.
2. Plot the histogram of computed white pixel.
3. Find the header line or shirorekha and remove it.
4. Construct the vertical histogram
5. Find the threshold value and separate histogram
6. Save the segmented image which are characters

Fig -7: Segmented characters

3.3Feature Extraction

In future extraction, we used Support Vector machine (SVM). In this SVM extract the set of feature from the database characters. It is very important step in developing a classification system.

3.4 Classification

In classification the extracted feature of image is matched with database feature of character. We used Support Vector Machine (SVM). SVMs is mostly used to solve various real world problems of uncertainty in Data Based Systems. We used a multiclassifer of k classes to classify the input character image. About 85% recognition rate is achieved.

4. CONCLUSIONS

In this paper, we mainly focused on segmentation of characters. As it is very difficult to segment joint character. At few points segmentation is good but at some point it was not up to the expectation. The unwanted segmented image provides unwanted recognized character. This reduced the overall accuracy of character recognition. This may be because of the shape and size of the characters. Segmentation is very crucial step as style of writing vary from one person to another or same person writes in different style over period of time. There is still lots of research need to be done on segmentation part. There is only few work done on script identification. Script identification is also very important. All these issues can be dealt in the future for handwritten documents in Devanagari script.

REFERENCES

[1]R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal, 'Offline Recognition of Devanagari Script: A Survey' IEEE Transactions On Systems, Man, And Cybernetics part C: Applications And Reviews, Vol. 41, No. 6, November 2011.

[2] Vijay Kumar, Pankaj K. Sengar, 'Segmentation of Printed Text in Devanagari Script and Gurmukhi Script', International Journal of Computer Applications, Volume 3 – No.8, June 2010.

[3] Smeet D. Thakur, Prof. Smita S. Sikchi, 'Offline Recognition of Image for content Based Retrieval' International Journal of Latest Trends in Engineering and Technology, Vol. 3,Issue2,November 2013.

[4] Mrs.Vinaya. S. Tapkir, Mrs.Sushma.D.Shelke, 'OCR For Handwritten Marathi Script' International Journal of Scientific & Engineering Research Volume 3, Issue 8, August-2012.

[5] Ratnashil N Khobragade, Dr. Nitin A. Koli , Mahendra S Makesar,- A Survey on Recognition of Devanagari Script, International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013.

[6] AshutoshAggarwal, Rajneesh Rani, RenuDhir, 'Handwritten Devanagari Character Recognition Using Gradient Features' International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 5, May 2012.

[7] Meng Shi, Tetsushi Wakabayashi, WataruOhyama, Fumitaka Kimura, 'Comparative Study on Mirror Image Learning (MIL) and GLVQ' Pattern Recognition, 2002. Proceedings. 16th International Conference on Volume 2, 2002.

[8] VedguptSaraf, D.S. Rao, 'Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency', International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-4, April 2013.

[9] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) 1979.