International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | April-2016 www.irjet.net p-ISSN: 2395-0072

Development of Intelligent Optical Handwritten Character Recognition System for Devanagari Script

Abhishek Sutar[1], Naveen Menon[2], Priyanka Potdar[3], Priyanka Patil[4], Komal Karande[5], Dr. Vijay Ghorpade[6]

[1-5]*Student, Dept. of Computer Science of Engineering, DYPCET, Kolhapur, Maharashtra, India*
[6]*Guide, Dept. of Computer Science of Engineering, DYPCET, Kolhapur, Maharashtra, India*

Abstract - *Some Indian languages like Hindi, Sanskrit and Marathi are composed from Devanagari script. This makes clear that it is the most widely used script in India. Most ancient documents are written in Devanagari script. The script has 16 vowels and 36 consonants. Unlike, English language Devanagari script does not have any upper or lower case characters. Instead, it has a horizontal line at the top called as shirorekha which joins the characters to form a word. Thus, the report gives a brief review about the development of Optical Handwritten Character Recognition for Devanagari script, the techniques & algorithms which are designed for examining & recognizing the Devanagari script. Further the report focuses on the key principles of the system such as text analysis and segmentation, matrix matching and recognition of single and compound characters. To keep up accuracy in recognizing the characters, the system uses some neural network algorithms.*

Key Words: **Devanagari, Matrix, OCR, OHCR, Recognize, Segmentation**

1. INTRODUCTION

Optical Character Recognition (OCR) is defined as a technique that scans an input image and converts it into an editable format. There are different tools available in the market which scans only printed documents and some of them are able to scan handwritten documents. But when it is concern with the precision of the system's output most existing systems fails to maintain enough accuracy as required. In today's world, there is a tremendous growth in the field of digitization. In government sectors and rural areas, there is a huge demand of such systems especially for converting the handwritten documents into editable format. The proposed system takes this charge by converting both the printed as well as the handwritten documents. For this process, there are various techniques and algorithms which have been introduced in earlier research. Among all those techniques, we have used some text analysis and segmentation techniques, matrix matching practice and some predefined procedures. The system assures to deliver much higher accuracy than already existing systems. The accuracy in character recognition of document totally depends on the user handwriting styles. It becomes complex for the system to examine and classify curving and thin letters. Marathi language words are categorized into strips as shown in fig 1.1. The Shirorekha is said as the header strip of the word which connects all characters in a single word. In standard Marathi writing, the end of shirorekha is considered as the last character of a particular word. The core strip also called as middle core of the word, consists of the consonants and vowels. The complex letters consists of the two words as a single character. Top strip and bottom strip having the complex letter characters. The basic features of Marathi language is discussed above and the further system development is purely done on this basis.

Fig – 1: Layers in Marathi word

2. LITERATURE SURVEY

As the Devanagari script [3] is most popular and widely used script in India, there is a huge demand for such systems that accurately and precisely recognize the Devanagari script. A variation of software's and tools are offered, that perform such tasks. Many papers have been introduced and a lot of efforts are being carried out in this research area. Still there is a need of software system that strictly emphasizes on the key problems of existing systems and overcome recognizing [7] problems by introducing some new techniques and features giving solution to this problem.

3. SYSTEM ARCHITECTURE

Our proposed system is based upon one time user interaction at only the first stage for giving the input for the system. Standard input is given by the cropping operation of the image in which user can choose the part or whole image which consists of the handwritten or printed script in it. Initially user provides input to the OHCR system the further operations are carried out by the system. No user interaction is required until the output gets generated. OHCR system works in sequential manner starting from initial stage to last stage. The overall system architecture is separated into modules. Major modules are:

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

The overall system architecture is shown in the below figure.

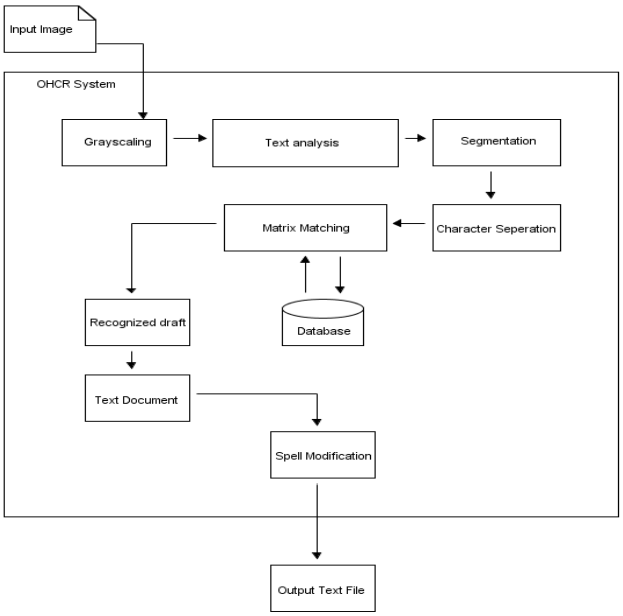

Fig – 2: OHCR system architecture

4. MODULES IN SYSTEM

4.1 Module 1 – Preprocessing part

A. Binarization

In Binarization, first the original image gets converted into gray-scale image by using threshold value and later converted into binarized image. Thresholding is the modest technique of image segmentation. It substitutes each pixel in an image with a black pixel, if the image intensity is less than some fixed constant or white pixel if the image intensity is larger than that constant.

Otsu's binarization process:
Input: Image
Procedure
Start:

1. Given image is first converted into gray-scale by analyzing the intensity of each pixel with respect to RGB.
2. Generate a new pixel which is in the form of grayscale and write it to output image.
3. Calculate threshold value for detecting a pixel whether it is intensity of 255(black) or 0(white)

If (Pixel less than threshold)
Set new pixel to 0 intensity
Else
Black

End
Output: Binarized image

B. Test Analysis and Segmentation

Text analysis techniques are applied to section the text image into lines and characters so that the place of each character is identified. The segmentation phase consists of three parts: line segmentation, word segmentation and character segmentation. These segmentation units are executed one after another. That is, when line segmentation completes, the output of this phase is given to word segmentation phase where every word is separated from each image file. Finally, the word segmentation is taken as the input for character segmentation. Here, individual characters are separated and stored into a different output file.

C. Line Segmentation

When the Binarization phase is completed the line segmentation technique is applied to separate the complete image into multiple uniform size images. The height of each segmented image is assigned to 70 pixels where the width is assigned dynamically. The output from line segmentation is stored into different files. These files are later served as the input for word segmentation phase.

D. Word Segmentation and Character Segmentation

In word segmentation, the basic idea is to isolate each word and store it in a new file. In this case, we have used a method of breaking each word when the consecutive

white pixels are encountered. These recognized words are then stored into a different file. As soon as the word gets recognized by the system, it is immediately send to the character segmentation phase. Here, each output file of word segmentation forms the input of character segmentation phase. The first recognized word is then separated by the system when it encounters the existence of consecutive white pixels. Each separated characters are finally stored in different output files. This process continues until all words are recognized by the system.

4.2 Module 2 – Recognition part

A. Training

In the training process, each solo alphabet is generated as a single image. These alphabets are binarized and provided to the system for training. The system will down sample the image and generate a 45*45 matrix. The matrix for each alphabet is stored into a single matrix DB. Once the matrix for all the alphabets is stored in the matrix DB, this database is used for recognition. The matrix DB contains matrix for every simple letter in the Devanagari script.

Training process:

Input: Binarized simple char

Procedure:

Start: 1. Down sampling: It is a process of adjusting input image into a suitable matrix.

2. The image is binarized.

If (pixel color is white)

Set matrix value to 0

Else

Set matrix value to 1

End

Output: Standard matrix of each input character.

B. Recognition

The letters produced from the character segmentation process is provided for recognition as input, one by one. For recognition, the system uses Kohenen algorithm. Kohenen algorithm is one for the basic forms of self-organizing neural network algorithms. The input is provided to the algorithm, the algorithm chooses the winning neuron, the one which corresponds with the input vector in the best way. During recognition, the algorithm performs down sampling of the image. The down sampling process fits the input image into a 45*45 matrix. The algorithm generates a matrix for each input image; this matrix is matched with the matrix available in the matrix DB. The matrix that matches closest to the matrix for the input image is declared as the winning matrix. Thus, the letter is recognized.

Recognition process:

Input: Output from file from character separation.

Procedure:

Start:

1. Load the matrix DB
2. Down sampling and matrix generation of input image.
3. Generated matrix is compared with matrix DB.
4. The matrix from matrix DB that matches closest to the input matrix is selected as winner.
5. The letter that corresponds to the winning matrix is displayed.

Output: Recognized character.

4.3 Module 3 – Preprocessing part

A. Recognize draft

The initial stage after the recognition phase is the draft recognition which implements neural graph technology. When the system detects a template in the matrix DB, template is checked as per the 1's and 0's matrix. The input pattern matrix is automatically generated in standard size and successfully stored in output text file. The system detects neuron to fire which approximately matching the pattern. As result it creates a raw draft for this recognition for each character in document.

B. Spell modifier

The additional and optional feature is the spell check which automatically redirects to a word which is occurs in system eventually. In the worst case of input spell modification suggest for words which are difficult to recognize to the system.

5. CONCLUSIONS

The proposed system overcomes the restriction of existing system by providing higher accuracy in word recognition. The proposed system focuses on handwritten text compared to the existing system that works on printed text. The system uses neural network algorithms for recognition of characters.

REFERENCES

A. Paper

[1] Chandranath Adak, Bidyut B. Chaudhuri, An Approach of Strike-though Text Identification from Handwritten Documents, Dept. of Computer Science &Engg. University of Kalyani, West Bengal, 2014.

[2] Yogendra Bagoriya, Nisha Sharma, Font Type Identification of Hindi Printed Document, CDAC Noida, GGSIPU, Delhi, 2014

[3] Shruti Agarwal, Dr. Naveen Hemarjani, Offline Handwritten Character Recognition with Devnagari Script, Computer Science and Engineering Department, Suresh GyanVihar University Jagatpur, Jaipur, India, 2013

[4] Dhanashree Joshi, Sarika Pansare, Combination of multiple image features along with KNN classifier for classification of Marathi Barakhadi, Dept. of Computer Engg. Sinhgad Academy of Engineering, Pune, 2015.

B. BOOKS

[5] "Java-A Begineer's Guide,3rd Edition(2005)"by Herbert Schildt.

[6] "An Implementation of OCR system Based on Skeleton Matching"by Ning LI.

[7] "An Invitation to Image Analysis and Pattern Recognition" by fred A. Hamprecht.