International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

Optical Character Recognition for Hindi

Prasanta Pratim Bairagi

Assistant Professor, Department of CSE, Assam down town University, Assam, India

---

Abstract - *Optical Character Recognition is a system which can perform the translation of images from handwritten or printed form to machine-editable form. Devanagari script is used in many Indian languages like Hindi, Nepali, Marathi, Sindhi etc. This script forms the foundation of the language like Hindi which is the national and most widely spoken language in India. In current scenario, there is a huge demand in "storing the information in digital format available in paper documents and then later reusing this information by searching process". In this paper we propose a new method for recognition of printed Hindi characters in Devanagari script. In this project different pre-processing operations like features extraction, segmentations and classification have been studied and implemented in order to design a sophisticated OCR system for Hindi based on Devanagari script. During this research, different related research papers on existing OCR systems have been studied. In this project the main emphasis is given towards the recognitions of the individual consonants and vowels which can be later extended to recognize complex derived letters & words.*

Key Words*: *Optical Character Recognition, Feature Extraction, Segmentation, Hindi Character, Devanagari Script

1. INTRODUCTION

The introduction part is divided into two individual parts. The first part defines about OCR, its types and its uses and the second part defines about Devanagari script, the foundation of Hindi language.

1.1 About OCR

Optical Character Recognition has emerged as a major research area since 1950. Optical Character Recognition is the mechanical or electronic translation of images of handwritten or printed text into machine-editable text [1]. The images are usually captured by a scanner. However, throughout the text, we would be referring to printed text by OCR. Data Entry through OCR is relatively fast, more accuracy, and generally more efficiency than usual keyboard entry. An OCR system enables us to store a book or a magazine article directly into digital form and also make it editable. Development of OCR for Indian script is an active area of research and it also gives great challenges to design an OCR due to the large number of letters in the alphabet, the sophisticated ways in which they combine, and the complicated graphemes they result in. Usually in Devanagari script, there is no separation between the characters written in a text. In this research work different pre-processing operations like conversion of gray scale images to binary images, image rectification and segmentation are considered in order to design this system.

1.2 Types of OCR

Basically, there are three types of OCR. They are briefly discussed below:

- Offline Handwritten Text

The text produced by a person by writing with a pen/ pencil on a paper and then scanned the document to digitalized them is called Offline Handwritten Text.

- Online Handwritten Text

Online handwritten text is the one written directly on a digital platform using different digital device. The output is a sequence of x-y coordinates that express pen position as well as other information such as pressure and speed of writing.

- Machine Printed Text

Machine printed texts are commonly found in printed documents and it is produced by offset processes.

1.3 Uses of OCR

Optical Character Recognition is used to scan different types of documents such as PDF files or images and convert them into editable file.

The OCR system is used for the following purposes:

- Processing Bank cheese
- Documenting library materials into digital format.
- Storing documents in digital form, searching text and extracting data.

1.4 About Devanagari Script

Devanagari script is the foundation of many Indian languages like Hindi, Nepali, Marathi, Sindhi etc and used by more than 300 million people around the world. So Devanagari script plays a very major role in the development of literature and manuscripts. There is so much of literature from the old age manuscripts, Vedas and scriptures and since these are so old so these are not easily accessible to everyone. The need and urge to read these old age scriptures led to the digital conversion of these by scanning the books. For scanning and converting the documents into editable

form OCR system for Devanagari text was introduced. This editable form out of output text can be input to various other systems like it can be synthesized with the voice to hear the enchantment of scriptures etc.

Devanagari script is written in left to right and top to bottom format [2]. It consists of 11vowels and 33 basic consonants. Each vowel except the first one have corresponding modifier using which we can modify a consonant. This line which is available in the upper side of a character is called "Shirorekha". Based on this shirorekha each character is divided into three distinct parts. The portion in the upper side of shirorekha is called upper modifiers, in the middle portion the character is available and in the last portion lower modifiers are available. Moreover, some characters combine to form a new character set called joint characters. Optical Character Recognition for Hindi is comparatively complex due to its rich set of conjuncts. The terminology is partly phoning in that a word written in Devanagari can only be judged in one direction, but not all possible pronunciations can be written perfectly [7].

2. RELATED WORK

The work on developing a character recognition system is initiated by Sinha [3, 4] at Indian Institute of Technology, Kanpur. Till today lots of effort have been devoted to design an OCR for the Devanagari script [5, 6], but no complete OCR for Devanagari is yet available.

Chirag I Patel et al. [7] highlight a method to recognize the characters in a given scanned documents and study the effects of changing the Models using Artificial Neural Network.

Jawahar et al. [8] have proposed a recognition scheme for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts.

Dileep Kumar Patel et al. [9] In this paper, the problem of handwritten character recognition has been solved with multiresolution technique using Discrete wavelet transform (DWT) and Euclidean distance metric (EDM).

3. METHODOLOGY

The algorithm that is used to develop the OCR software for printed Hindi characters is based on the different geometrical features/shapes of Hindi characters. Input image is parsed into many sub parts/images based on these features. Then other properties such as distribution of points/pixels and edges within each sub images are features used to recognize parsed symbol.

The major properties used to segment input character (image) into various sub symbols are- Horizontal lines, Vertical lines, Cross lines, Curves, Loops.

Among all the above properties mostly Horizontal and Vertical lines form an integral part of most Hindi characters.

3.1 Various steps involves in this proposed system

The proposed system includes different steps as follows:

- First take the printed binarized image of a character as an input.
- Extract the pixel information from that image and store them into a suitable memory.
- After successful completion of the 2nd step, try to find out the skeleton of that character based on the pixel information.
- Once the skeleton is available, try to find out the different features or geometrical shapes available in that skeleton.

The feature extraction process contains the following:

  - Detection of Horizontal lines
  - Detection of Vertical lines
  - Detection of Cross lines
  - Detection of Curves
  - Detection of Loops

- Simultaneously we prepare a database where all the features of each and every character are stored.
- Now compare the features found in the input image with the database and check whether the features obtained from that particular character is matches with the stored features list or not. If match found then the next step will be pass the Unicode value of that particular character to the file writer and write the character into a text file.
- Finally we will get the character in an editable format from the image format.

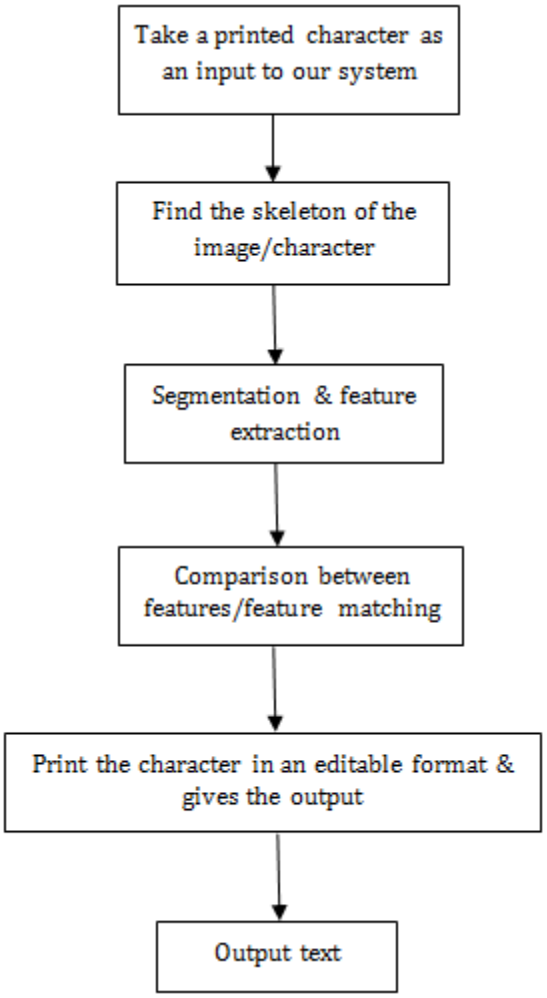

Figure1: Steps involve in this system

3.2. Design of an OCR

Following are the implementation details of the various steps in the proposed algorithm.

- Input file/image format to the OCR

The implemented OCR expects the input image to be in either .bmp or .jpg format. The image should be a binary one. The text image should be written with two possible combination of colour. One is text in black colour and the background should be white or the other one is text in white colour and the background should be black. That is, the image should have only two types of pixel values, 0, for background and 1, for the foreground.

- Binarization

For testing purpose we collected some images of characters and prepare a database of these. Since the developed system is only able to perform its task only on binarized image so we have to perform the binarization operation before the actual task starts. But here the collected images are already binarized so we need not to perform the binarization operations.

- Extracting pixel information

The binary images that are used for testing purposes consist of a white foreground in front of a large black background. The number of pixels in the background far exceeds that of those in the foreground. This means the numbers of 0's will always be at least 5 times the number of 1's. Moreover, smaller number 1s will mean lesser calculations in correlation. The extraction of pixel information is done by analyzing the foreground and background colours and stored the colour information in terms of 0'a and 1's in matrix of the image size.

- Thinning or finding the skeleton of the image

The skeletonization phase is the first one to manipulate the input binarized image and produce polylines that describe the strokes comprising the characters.

Since the algorithm is based on the geometrical and structural properties of the Hindi characters, we think the image to single-pixel width so the contours are brought out more vividly. In this way, the attributes to be studied later will not be affected by the uneven thickness of edges or lines in the symbol. Thinning is a morphological operation that is used to remove selected foreground pixels from binary images. The key here is the selection of the right pixels.

Usually there are three types of pixel present in an image or we can categories the pixels into three categories. These are:

Critical Pixels – Pixels whose removal damages the connectivity of the image. Any pixel which is the lone link between a boundary pixel and the rest image is a Critical Pixel. Its removal will isolate the boundary pixel. Hence it should not be removed.

End Pixels – Pixels whose removal shortens the length of the image. An end pixel is connected to two or less pixels. Remember that we are talking about 8-connectivity here. Different considerations have to be taken for 4-connectivity.

Simple Pixels – Pixels which are neither Critical nor End pixels. These are the ones that can be removed for thinning.

Like the other morphological operation, the behavior of the thinning operation is determined by a Structuring Element. Here in our thinning algorithm we used the eight neighbourhood concept to fine the skeleton of the character. Instead of eliminating one pixel at a time we identify the unwanted pixel of same region and then deleted them at once which decrease the time required to find the skeleton of the image.

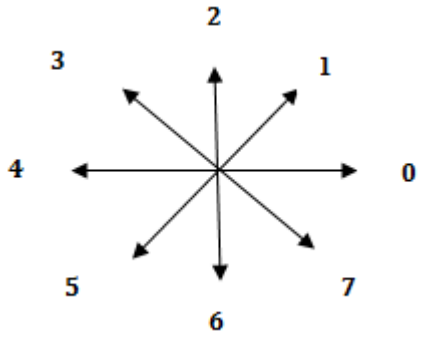

Figure 2: Eight neighbourhood of a pixel

- Detection of lines

After thinning a given alphabet to a single line we try to detect the features i.e. the distinct parts available on that alphabet taking the horizontal (shirorekha) and vertical line as baseline.

For a given input image we move from starting pixel termed as base pixel to the next neighbour pixel to detect the type of line based on some rules.

If the next neighbour pixel is in a left or right direction of the base pixel then the type of line is considered as horizontal line.

If the next neighbour pixel is in an upward or downward direction of the base pixel then the type of line is considered as vertical line.

If the next neighbour pixel is in a left upward or right downward direction of the base pixel then the type of line is considered as a line having negative slope.

If the next neighbour pixel is in left downward or right upward direction of the base pixel then the type of line is considered as a line having positive slope.

- Detection of Loop

Along with the line set we detect loops if available on the given character. If the starting pixel and the ending pixel of a set of line are same then this set of line constitutes a loop.

- Compression of the obtained line segments

Compression is performed to ignore some distortion available in the set of lines constituting the character. Thus we get minimum and necessary line segments which clearly represent that character.

- Detection of Curves

Since most of the characters in Hindi alphabet has a horizontal and vertical line, so we extract these lines first from the obtained line set and from the remaining line set we try to construct loop and curves. Choose any line which is closest to the vertical line and start draw a line from starting point of a line to the ending point of the consecutive line segment. If the sum of length of these line is greater than the length of the end point connecting line by some threshold value then it is considered as a curve. If it intersects any point then reverse the operation to detect common line segment which is belongs to two different parts of that character.

- Identification of individual character

Since most of the alphabets in Hindi have horizontal or vertical line so we find these lines first and then other lines, loops, curves and compare these features with the stored database features to identify the resultant character.

4. RESULTS

The program was rigorously tested on sample images of printed Hindi characters which includes all the vowels and the consonants. The accuracy of this developed software is quite good. Since we can't show all the characters in results so we take a specific character 'PHA' to explain our approaches towards recognized a character.

Step 1: Take the binarized character image as an input.

Figure 3: Input Image

Step 2: Find the skeleton of the character

Figure 4: Skeleton of the image

Step 3: Extract the different features available in that image

- Feature 1: Horizontal line

Figure 5: Horizontal Line

- Feature 2: Vertical line

Figure 6: Vertical Line

- Feature 3: Right side Curve

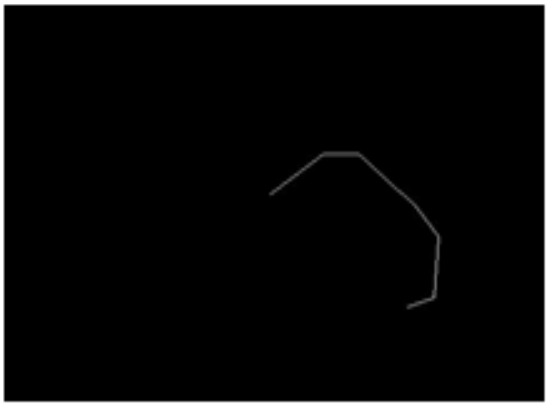

Figure 7: Right Side Curve

- Feature 4: Left side Curve

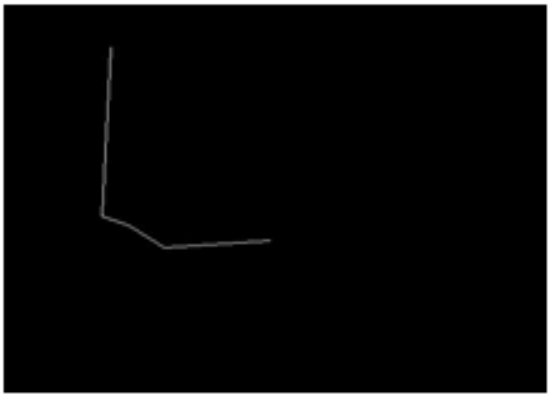

Figure 8: Left side Curve

Step 4: Write the character in editable text format

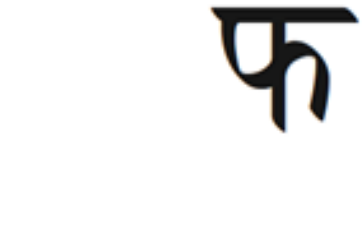

Figure 9: Output text with Unicode value 092B

5. CONCLUSION

In this paper, we have described a system for OCR of printed Hindi characters. The recognition accuracy of the prototype implementation is very promising. During this project it has been clearly noticeable that classification of patterns affects a lot in the accuracy of an OCR system. More the classification, more accurate results can be produced.

6. FUTURE SCOPE

The current thinning or skeleton finding algorithm is depend on the size of the image, which is not a very good approach towards developing a software like OCR. So we can try to overcome this situation by improving the current thinning algorithm which needs more time and effort.

In this project we used a simple form of database to recognize a character. With the current database the character recognition is good but to develop good quality OCR software the database should be fully organized. So some effort and analysis is also required in the database part.

ACKNOWLEDGMENT

The author would like to thank **Prof. (Dr.) L. P. Saikia** (Department of Computer Science Engineering, AdtU) for his constant moral support, encouragement and guidance that helps in correcting mistakes and proceeding further to produce the paper with the required standards.

REFERENCES

[1] S. Mori, et. al.: "Historical Review of OCR Research and Development". Proceeding IEEE, Vol.80, No.7,1992, pp.1029-1058.

[2] R. Plamondon and S. N. Srihari, "On-Line and Off-Line Handwritten Recognition, A Comprehensive Survey", IEEE Pattern Analysis and Machine Intelligence, Vol- 22, No. 1, January 2000.

[3] B. Philip and R. D. Sudhaker Samuel. "An Efficient OCR for Printed Malayalam Text using Novel Segmentation Algorithm and SVM Classifiers" International Journal of Recent Trends in Engineering, Issue. 1, Vol. 1, May 2009

[4] S. Mohanty, H. N. Dasbebartta, "An Efficient Bilingual Optical Character Recognition (English-Oriya) System for Printed documents", IEEE Conference, 13FebruaryN 2009

[5] U. Garain, B. B Chaudhuri, "Segmentation of Touching Character in Printed Devanagari and Bangla Script Using Fuzzy Multifactorial Analysis". IEEE Transaction on System, Man and Cybernetics –Part C: Applicationsand Reviews, Vol.32, No.4, 2002, pp.449-459.

[6] C.V. Jawahar, M.N.S.S.K.P. Kumar, S.S. Ravi Kiran, "A Bilingual OCR for Hindi-Telugu Documents And it's Application. Document Analysis and Recognition". IEEE Proceedings Seventh International Conference on, Vol.1, 2003, pp.408-412.

[7] C. I. Patel, R. Patel, P. Patel "Handwritten Character Recognition using Neural Network", International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011

[8] Sankaran, Naveen, and C. V. Jawahar. "Recognition of printed Devanagari text using BLSTM Neural Network." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

[9] D. K. Patel, T. Som, S. K. Yadav, M. K. Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric", JSIP 2012, 208-214

[10] B. Philip and R. D. Sudhaker Samuel. "A Novel Bilingual OCR for Printed Malayalam-English Text based on Gabor Features and Dominant Singular Values", 2009 IEEE.

[11] V. Bansal, R.M.K. Sinha, "On How to describe Shape Of Devanagari Characters and Use them for Recognition" 5th International Conference on document Analysis and recognition (ICDAR'99), Bangalore India, 1999.

[12] V. Bansal, R.M.K. Sinha, "A Devanagari OCR and A Brief Overview of OCR Research for Indian Script",PROC Symposium on Transaction support System (STRANS 2001), Kanpur, India, 2001.

Author Profile

Mr. Prasanta Pratim Bairagi received his MCA degree from Tezpur University in the year of 2013. He has five (05) years of experience in teaching and currently working as an assistant professor in the Department of Computer Science and Engineering, Assam down town University. His research area of interest includes Image Processing and Wireless Sensor Network.