A Review of Different Approaches Used for Devanagari Character Recognition

Mayank Sahai
Student, Department of CSE and IT,
ITM University,
Gurgaon, India

Neha Sahu
Assistant Professor,
Department of CSE and IT,
ITM University, Gurgaon, India

ABSTRACT
Handwriting Recognition System helps converting hand written text documents to digital form. Hand written recognition systems are different for different languages depending upon the script's complexity. Extensive research have been performed on various languages across the globe including English but Devanagari Script has been left far behind due to its complex nature. This paper enlightens some of the popular research performed in recognizing Devanagari script. It also summarizes various advantage and scope of using different methodology including Bounding Box technique, Ostu's algorithm, neural networks and many more.

Keywords
Devanagari Character Recognition, OCR, Classification, Feature Extraction

1. INTRODUCTION
Handwriting Recognition is one of the most engrossing and challenging research areas in the field of image processing and pattern recognition. Handwriting recognition System is a system by which a computer system can recognize the characters and the symbols written by hand in natural language like English, Devanagari, and Gurumukhi etc. The HWR system is basically of two types: Online HWR System and Offline HWR System. In Online HWR system, the text to be recognized is given as input to the system using a stylus or digitizer. Then the data signals undergo some filtration process, the data signal is then normalized to normal size and the slant and slope is corrected. After normalization, the text is divided into segments, and each segment is classified and labelled. Then using a search algorithm the most appropriate path is sent back to the user as output. In Offline HWR system, the text to be recognized is given as input in the form of scanned text, camera pictures etc. The Optical Character Recognition (OCR) is a type of Offline HWR system. In OCR, the input data is segmented into pieces using different algorithms. After the data is segmented into pieces, the text is further segmented into words or characters and sent to the recognition system. In the engine, the skeletonization and preprocessing is applied on the segmented text. Then different classifiers are applied which extract certain features and creates a character hypothesis list. Then a search algorithm is used to search the most appropriate path in conjunction with language models and sent as output to the user.

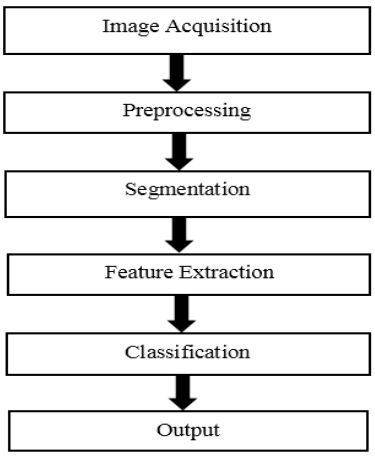

Figure 1: Steps of HWR system

The flowchart in Figure. 1 shows the steps that are performed during the recognition. The first step is the image acquisition in which the image consisting of the words to be recognized is input to the system. Then the image is preprocessed by performing the operations such as noise removal, normalization, skew detection and correction, greyscale and binarization etc. After preprocessing, the word in the image is segmented into individual characters for recognition. Then the key features are extracted from the image and these key features include height, width, density, loop, lines, stems and other character traits. This step is known as feature extraction. In the classification step, the methodologies of pattern recognition are used for assigning an unknown sample to a predefined class. Then the character is compared with the character in the trained system and an output is obtained in form of character.

2. CHARACTERISTICS OF DEVANAGARI SCRIPT
In India, there are officially eighteen languages and Devanagari is one of them. The Devanagari script is used for writing Sanskrit and other Indian other languages. It is written from left to right, lacks distinct letter cases and is recognizable by the horizontal line running along the tops of the letters that links them together known as "Shirorekha" or headline. It consist of 11 vowels and 33 consonants. Vowels can be written as independent letters or by using them above, below, before or after the consonant they belong to. When the vowels are written in this way they are known as modifiers and the characters so formed are known as conjuncts. Two or more consonants can be combined together to form compound characters.

Table 1: Vowels And Corresponding Modifiers [2]

Vowels:	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
Modifiers:		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ

Table 2: Consonants [2]

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट
ठ	ड	ढ	ण	त	थ	द	ध	न	प	फ
ब	भ	म	य	र	ल	व	श	ष	स	ह

Table 3: Half Form Consonants With Vertical Bar [2]

क्	ख्	ग्	घ्		च्		ज्	झ्	ञ्	
			ण्	त्	थ्		ध्	न्	प्	फ्
ह्	भ्	म्	य्		ल्	व्	श्	ष्	स्	

Table 4: Example Of Combination Of Half-Consonants And Consonants [2]

क क्क	क ल्क्ल	घ न्घ्न	च ञ्च्ञ	श च्श्व	त न्त्न	प त्प्त	प ल्प्ल
व व्व	भ न्भ्न	म ल्म्ल	ल ल्ल्ल	श न्श्न	श व्श्व	श ल्श्ल	स न्स्न

Table 5: Example Of Special Combination Of Half-Consonant And Consonant [2]

क ष क्ष	ज ञ ज्ञ	ट ट ट्ट	ट ठ ट्ठ	त र त्र	द द द्द
द ध द्ध	द व द्व	द व र द्व्र	श र श्र	द भ द्भ	द य द्य

Table 6: Special Symbols [2]

क़	ख़	ग़	ज़	फ़	ड़	ढ़	ॅ	ॉ	ः	ँ	ऽ	़

In Devanagari script, there are four imaginary lines that are drawn for a word, Headline which is also the header line, Baseline where the characters complete without modifiers, Upperline which the line above the Headline after above modifiers, and Lowerline which after the below modifiers. The text is partitioned into three zone: Upper zone between Headline and Upperline, Middle zone between Headline and Baseline, and the Lower zone between Baseline and Lowerline.

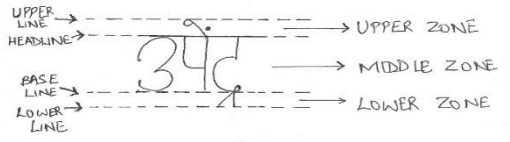

Figure 2: Different zones of Devanagari text

3. COMPLEXITY OF DEVANAGARI SCRIPT

While segmenting the Devanagari character, there are many problems that occurs due to complexity of the script. Following are some of the complexities of the script:

i. Problem of broken character

ii. Problem of overlapped characters

iii. Problem of touching characters

iv. Problem of skewed characters

v. Problem of irregular intensity with character

vi. Problems due to presence of upper and lower modifiers

vii. Problems while separating the word and "Shirorekha"

viii. Problems of recognition special symbols

The HWR system should be developed such that it deals with all these complexities and recognizes the word properly.

4. IMAGE ACQUISITION

Image acquisition is the first step of the HWR system in which the image consisting of handwritten Devanagari characters are input to the system. Either a scanner or the digital camera can be used to take the image of the document consisting of the Devanagari script. The quality of the camera and the scanner effects the quality of the image.

5. PRE-PROCESSING

After the image is input to the system, the image is preprocessed by applying some steps which will help to improve the quality of the image for the recognition process. Some of the steps of the preprocessing are the following:

5.1. Grayscale conversion

The image input to the system can be in rgb format which is a three dimensional image, but segmentation and recognition process can be applied only on a two dimensional image, so the colored image is converted into grayscale image. [6, 7, 8]

5.2.Binarization

Binarization is a process of converting a grayscale image into binary form i.e. 0's and 1's representation, in which 0 represents black color and 1 represents white. Binarization separates the background and the foreground objects which is useful in segmentation. One of the most widely use method for Binarization is Ostu's algorithm. This method uses a threshold value to minimize the intra-class variance between the background and the foreground objects. [6, 7, 8]

5.3. Skew correction

Different people have different writing style. It may be possible that the handwritten text in a document will not be in a straight line, this requires for the skew detection and correction. This is because if the skew will not be corrected then there will be problem in segmentation. First the skew is detected and then corrected. To detect a skew, the angle of the page is detected first and then the angle of lines are compared with it and the skew is detected. Then the skew correction algorithm is applied and the skew is removed. [6, 7, 8]

6. SEGMENTATION

Segmentation is the process of breaking the connected word into individual character for the recognition process. The text in the document is first segmented into lines and then into words and finally into characters. To segment the Devanagari words into characters, the first step if to remove the Headline or the Shirorekha and then segment the word into individual characters. Bounding box technique [6], graph search [5] and Ostu's threshold technique [7] are some of the techniques used for segmentation.

7. FEATURE EXTRACTION

The process of extracting the useful information from raw data to minimize the intra-class pattern variability and maximize the inter-class pattern variability is known as feature extraction. Features are extracted from the segmented words so that to differentiate between classes. Some of the methods for feature extraction include: zone and count metric based system [7], feature extraction using MDRNN [4], statistical features [6] etc.

8. CLASSIFICATION AND RECOGNITION

After the feature have been extracted, some classification method is employed and the handwritten segmented character is recognized. Based on the features extracted the word is classified. Some of the methods commonly used for classification in the HWR system include: Neural Network [6, 7, 8], Support Vector Machine, Hidden Markov Mode etc.

9. RESULT

The following table compares the different approaches used by different authors in their paper for segmentation, feature extraction and classification. The table also compares the recognition accuracy of the systems proposed by different authors.

Table 8: Comparison of methods and accuracy of systems proposed by different authors

S. No	Method Proposed by	Segmentation	Feature Extraction	Classifier	Data Size	Accuracy obtained
1.	Neha Sahu et al. [6]	Bounding box technique	Height, width, density, loops etc	Neural Network	Unspecified	75.6 %
2.	Felipe Mendonca Gouveia et al. [4]	Graph search	Extracted using MDRNN	MDRNN-LSTM	5,685	87.7 %
3.	Gaurav Jaiswal [7]	Ostu's threshold technique	Zone and count metric based system	Neural Network	1174	75%

10. CONCLUSION

In the fast growing era of technology, there has been a drastic increase in the research field of Devanagari Character recognition system. The recognition of the Devanagari character is a difficult task due to the "Shirorekha", upper and lower modifiers, and also due to the complexity of the characters. The errors relates to improper, skewed, broken letter, zig-zag letter images should be considered and removed to obtain a better accuracy during recognition. The features extracted should improve the process of recognition. Most of the research have used the concept of neural networks for classification, but there are many other techniques which can be used for classification.

The main complexity of the Devanagari Characters is the Shirorekha extraction and after extraction, recognition of the word properly. But since only few research have been reported in this area, so different techniques can be applied for Shirorekha extraction and recognition of word properly.

Devanagari characters also consist of lower and upper modifiers, but since this increases the complexity of the characters so there are very less innovation related to it. Different techniques should be applied on the Devanagari characters consisting of modifiers and the system should recognize the word correctly and properly with modifiers also.

From the survey, it have been noted that the major problems in the recognition of Devanagari character include Shirorekha extraction, broken characters, skewed characters, errors generated by the scanner while scanning the images and many more. Proper techniques should be applied to handle all such errors and create an efficient and effective hand written recognition system.

11. FUTURE SCOPE

Devanagari script is the basis of many other scripts in the Indian language. There are many historic and important documents present in India which needs to be digitized using the recognition system. Different authors are working in this field and there is huge scope in this area. As most of the authors focused only on the fixed sized words, so one of the most important area is the recognition of random sized words.

The main complexity of Devanagari words is the presence of modifiers. Only few research have been successfully on it but the recognition accuracy is very so. So, authors should also work on the Devanagari words consisting of lower and upper modifiers and improve the system accuracy by recognizing the word properly.

12. REFERENCES

[1]. Hoyamoon S. M. Beigi : An Overview of Handwriting Recognition, Proceedings of the 1st Annual Conference on Technological Advancements in Developing Countries, Columbia University, July 24-25, pp. 30-46, New York (1993)

[2]. Vikas J Dongre and Vijay H Mankar: A Review of Research on Devnagari Character Recognition, International Journal of Computer Applications Vol. 12– No.2, (November 2010)

[3]. Nur Sukinah Aziz and Mohd Nizam Saad: Redesigning the User Interface of Handwriting Recognition System for Preschool Children, 2nd International Conference on Education Technology and Computer (ICETC), (2010)

[4]. Mustafa Ali Abuzaraida, Akram M. Zeki and Ahmed M. Zeki Problems of Writing on Digital Surfaces in Online Handwriting Recognition Systems, 5th International Conference on Information and Communication Technology for the Muslim World, (2013)

[5]. Felipe Mendonca Gouveia, Byron Leite Dantas Bezerra, Cleber Zanchettin and Jo˜ao Raul Jardim Meneses: Handwriting recognition system for mobile accessibility to the visually impaired people, IEEE International Conference on Systems, Man, and Cybernetics, (October, 2014)

[6]. Neha Sahu and Nitin Kali Raman, "An Efficient Handwritten Devanagari Character Recognition System Using Neural Network, IEEE, pp. 173-177,(2013)

[7]. Gaurav Jaiswal: Handwritten Devanagari Character Recognition Model Using Neural Network, International Journal of Engineering Development and Research(IJEDR), Vol. 2, Issue 1, (2014)

[8]. Richa Patil and Varunakshi Bhojane: Character recognition of Devanagari characters using Artificial Neural Network, International Journal of Computational Engineering Research (IJCER), Vol. 05, Issue 02, (February, 2015)

[9]. Kunal Shah, Jaideep Singh, Prashant Pushkarna, Hasnain Kurawadwala and Abhishek Alate: A New Approach for Segmentation of Devanagari Characters, Global Research Analysis(GRA), Vol. 2, Issue : 4, (April 2013)

[10]. R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, Umapada Pal: Offline Recognition of Devanagari Script: A Survey, IEEE Transaction on Systems, Man and Cybernetics, Vol. 41, No. 6, (November 2011)

[11]. R. J. Ramteke, S. C. Mehrotra: Feature Extraction Based on Moment Invariants for Handwriting Recognition, IEEE,(2006)

[12]. Chavan S. V., Kale K. V. Kazi M. M., Rode Y. S.: Recognition of Handwritten Devanagari Compound Character a Moment Feature Based Approach, International Journal of Machine Intelligence, Vol. 5, pp: 421-425, (April 2013)

[13]. Ankita Karia, Sonali Sharma, Reevon Rodrigues, Maitreya Save" Character Recognition (Devanagari Script), International Journal of Engineering Research and Applications, Vol. 5, pp. 109-114,(April 2015)