Handwritten Devanagari Character Recognition System: A Review

Sonal Khare
Department of Computer Science and Engineering, Krishna Institute of Engineering and Technology, Ghaziabad, India

Jaiveer Singh
Department of Computer Science and Engineering, Krishna Institute of Engineering and Technology, Ghaziabad, India

ABSTRACT
India is a multi-lingual country consisting of eleven different scripts. Hindi is third most widely used language after English and Chinese. Hindi, the national language of India, is written in the Devanagari script. Devanagari script is also used for other languages such as Sanskrit, Marathi and Nepali. In India, Devanagari script is used by more than 500 million people for documentation. Devanagari script should be given special attention so that analysis of ancient Indian literature can be effectively done. Lot of work has been done in handwritten character recognition and lot of work is to be done. This paper presents a detailed analysis of research work related to devanagari character recognition. This paper presents comprehensive survey of handwritten character recognition system using different features of character and classifiers used.

Keywords
Handwritten character recognition, off-line character recognition, Feature extraction, Segmentation, OCR, classifier.

1. INTRODUCTION
Devanagari is an oldest Indian script[15] that is used to write other languages such as Sanskrit, Hindi, Marathi and several others languages. Hindi is used as an official language more than 1.2 billion people worldwide.

Character recognition [9][14][16] is considered as one of the important technology in today's world. It is used in various fields such as artificial intelligence, computer vision, and pattern. Handwritten character recognition [8][17] is divided in to two parts i.e. offline handwritten character recognition and online handwritten character recognition.

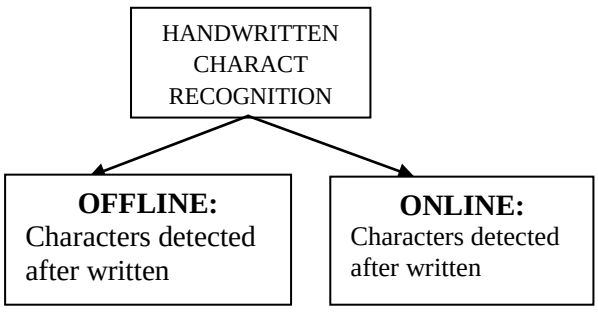

Figure 1: Handwritten character recognition

i) Offline handwritten character recognition [22]:-In this type of character recognition, the typed/handwritten character are scanned and then converted in to digital form. Offline character recognition is more challenging and difficult task as it does not have the advantage of recognizing direction of movements which writing the text.

ii) Online handwritten character recognition:-In this type of online handwritten character recognition, writing and recognition are done simultaneously. In this case, user will write character on any sensory area where sensor will pick up the pen motion and then it recognizes character on the basis of direction of the motion. Online character recognition is much easier than offline character recognition because there is timing information is available. This method is generally available on touchpad, touch screen cell phones etc.

2. LITERATURE SURVEY
R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal, [1] presents All feature-extraction techniques as well as training, classification and matching techniques useful for the recognition are discussed. An attempt is made to address the most important results reported so far and it is also tried to highlight the beneficial directions of the research till date. From the survey, it is noted that the errors in recognizing printed Devanagari characters are mainly due to incorrect character segmentation of touching or broken characters. Because of upper and lower modifiers of Devanagari text, many portions of two consecutive lines may also overlap and proper segmentation of such overlapped portions are needed to get higher accuracy.

U. Pal, B. B. Chaudhuri, [2] presents a review of OCR work done on 12 Indian language scripts. 12 Indian scripts are English, Devnagari, Bangla, Gurumukhi (Panjabi), Tamil, Telugu, Gujarati, Kannada, Kashmiri, Malayalam, Oriya and Urdu.

R. G. Casey and E. Lecolinet, [3] presents a review of Methods and Strategies used in Character Segmentation. In this paper, authors have proposed an organization of these methods under three basic strategies, with hybrid approaches also identified.

Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, [4] presents the comparative study of the well-known methods used in various stages of a pattern recognition system and identify research topics. This paper explains different applications which are at the forefront of this exciting and challenging field. New and emerging applications, such as data mining, web searching, retrieval of multimedia data, face recognition, and cursive handwriting recognition, require robust and efficient pattern recognition techniques. The design of a pattern recognition system requires the following issues: definition of pattern , sensing environment, pattern representation, feature extraction and feature selection, cluster analysis, classifier design and learning, selection of training and test samples, and performance evaluation.

Neha Sahu, Nitin Kali Raman , [5] describes the development and implementation of one such system consisting combination of various stages. Artificial Neural Network technique is used to designed to preprocess, segment and recognize Devanagari characters. The system was designed, implemented, trained and found to exhibit an accuracy of 75.6 % on noisy characters.

Sonika Dogra, Chandra Prakash, [6] describes a recognition system which can be used for the recognition of offline handwritten Hindi characters. For this proposed system Support Vector Machine is used as classifier and Diagonal feature extraction approach is used to extract features. From the results it is clear that combination of SVM classifier and diagonal feature extraction approach is better method for the recognition of handwritten characters.

3. DEVNAGRI HANDWRITTEN CHARACTER RECOGNITION

In India, more than 300 million people who speaks Hindi and uses Devanagri script[10][21] for writing. DOCR need more attention as it is natural language in the world. Nepali , Sanskrit and Marathi are also written in devnagri script. Devnagri is composed of two Sanskrit word "Deva" and "Nagri". Deva means god and Nagri means city. Devnagri script has 11 vowels ('svar') shown in table 1 and 33 consonants ('vyanjan') shown in table 2. There are no capital letters. Devnagri is written from left to right. The concept of upper/lower is missing from Devnagri script.

Table 1 vowels in Hindi

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ

Table 2 Consonants in Hindi

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

In Hindi language various zone portioning is described in Devanagari script.There are three different zones in Devanagari text ie upper zone, middle zone, lower zone. The upper zone and middle zone are always partition by header line which is named as shirorekha. The upper zone encompasses upper modifiers and lower zone encompass lower modifier (shown in figure 2). In Hindi word, upper modifier and lower modifier are not always necessary.

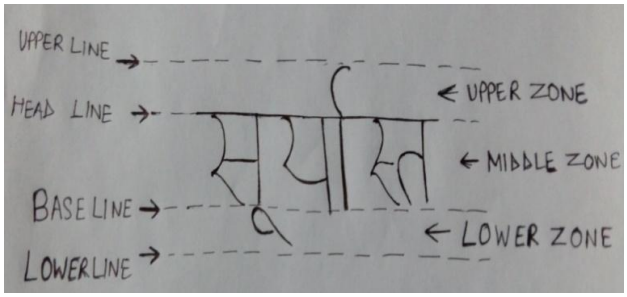

Figure 2. Various zones

4. OCR SYSTEM

OCR (Optical Character Recognition)[7][8][12][13] is consisting of following phases are as given below-:

4.1 Digitization

Digitization is the process of converting paper document in to electronic form. For this handwritten documents are scanned thus an image is processed. This image is the fed in to the next preprocessing stage.

4.2 Pre-processing

Pre-processing[19] is the initial stage of character recognition. The main stages of pre-processing[11] are shown in the figure.

4.2.1 RGB to Gray scale conversion: The scanned image is stored as JPEG images, BMP, TIFF etc which is in RGB format. Now this image must be converted in to a gray scale image. A gray scale image represents an image as matrix where every element has a value corresponding to how bright/dark the pixel at the corresponding position should be colored.

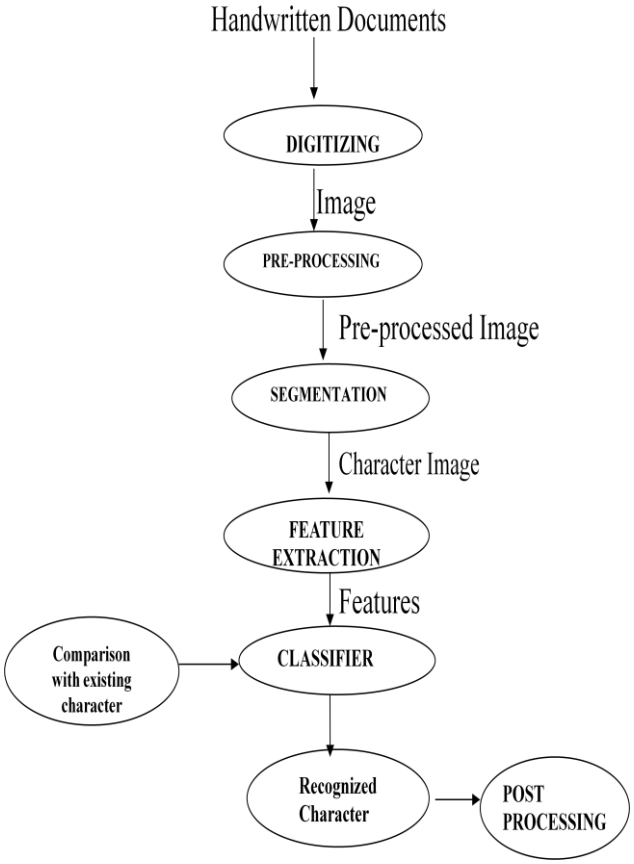

Figure 3: Handwritten character recognition system

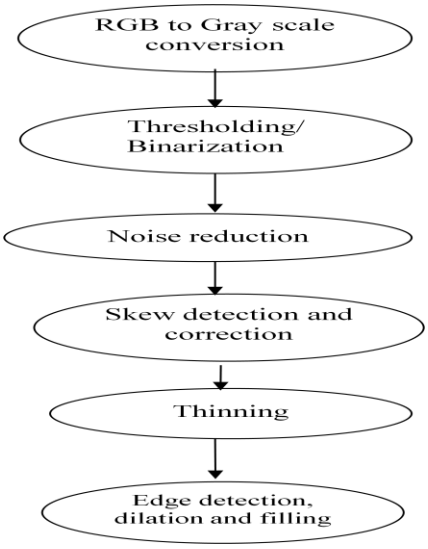

Figure 4: Preprocessing stages

4.2.2 Thresholding/Binarization: Binarization is a process which converts a gray scale image in to a binary image using thresholding technique. Thresholding reduces the storage requirement and increases the rate of processing by converting the gray scale image in to binary image by taking a threshold value.

4.2.3 Noise reduction: Noise reduction is introduced by the optical scanning device which causes disconnected line segment, bumps, and gaps in lines. The distortion including local variations, rounding of corners, dilation and erosion is also a problem. It is necessary to eliminate the imperfection. Noise Reduction techniques can be categorized as (a) Filtering (b) Morphological operation (c) Noise modeling.

(a) Filtering: This process is usually introduced by bad sampling rate of the digitization stage which aims to remove noise and decrease specious points. The basic goal is to elaborate a predefined mask with the image to allocate a value to a pixel as a function of the gray values of its neighboring pixels. Filters can be designed for smoothing, sharpening, thresholding, removing slightly textured or colored background, and contrast adjustment purposes.

(b) Morphological operation: The basic goal behind the morphological operations is to filter the document image replacing the convolution operation by the logical operations. Various morphological operations can be designed to connect the broken strokes, decay the joined strokes, smooth the contours, reduce the wild points, thin the characters, and extract the boundaries. Hence morphological operations successfully remove noise from the document images, due to poor quality of ink and document, as well as erratic hand movement.

(c) Noise modeling: Noise can be removed by various calibration techniques if a model for it were available. Modeling of the noise is not feasible in most of the application. There is very diminutive work on modeling the noise introduced by optical distortion, such as speckle, skew, and blur.

4.2.4 Skew detection and correction: Skew detection of the scanned image specifies the deviation of the text lines from horizontal or vertical axis. This is caused if the paper is not fed straight in to the scanner. Skewed lines are made horizontal by calculating skew angle and making proper correction in the raw image.

4.2.5 Thinning: Thinning extracts the shape information of the character. Thing is a morphological operation which is used to remove selected foreground pixels from the binary images and thin the images to single pixels width level so that their contours are brought out more vividly.

4.2.6 Edge detection, dilation and filling: Detection of images in the binarized image is done using sobel technique. After locating the edge the image is dilated and the holes present in the image are filled. These are the operation performed in the last two stages to produce the pre-processed image suitable for segmentation.

4.3 Segmentation

Objective of segmentation is to partition an image into regions.Segmentation is important stage in Character recognition system. The process of segmentation follows the following pattern:

1. Identify the text line in the pages.
2. Identify the word in individual line.
3. Finally identify individual character in each word.

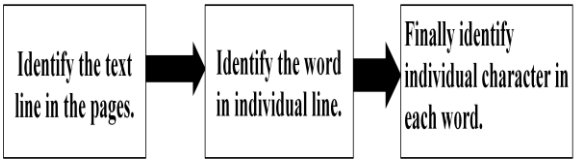

Figure 5: Segmentation

Segmentation can be external and internal. External segmentation is the separation of various writing units such as paragraph, sentences or words. In internal segmentation, an image of series of characters is decomposed in to sub-images of individual character.

There are three basic techniques for segmentation and various hybrid approaches used for segmentation. The basic techniques are:

1) The classical approach, in which segments are identified based on "character-like" properties. This process of cutting up the image into meaningful components is given a special name, "dissection," in discussions below.

2) Recognition-based segmentation, in which the system searches the image for components that match classes in its alphabet.

3) Holistic methods, in which the system seeks to recognize words as a whole, thus, avoiding the need to segment into characters.

In technique 1, the criterion for good segmentation is the agreement of general properties of the segments obtained with those expected for valid characters. Examples of such properties are height, width, separation from neighboring components, disposition along a baseline, etc. In techniques 2, the criterion is recognition confidence, perhaps including syntactic or semantic correctness of the overall result. Holistic techniques (method 3) in essence revert to the classical approach with words as the alphabet to be read.

4.4 Feature extraction

Next step is extracting features from the segmented characters and to distinguish it from other character by comparing it with already stored patterns of all the characters in the library. There are various features used for character recognition. Key features may include height, width, density, loops, lines, stems and other character traits. Chain code histogram feature is used for recognition that is extracted by chain coding the contour points of the scaled character bitmapped image. View based features is also used that is extracted from scaled, thinned one pixel wide skeleton of character image. Shadow features are extracted from scaled character image.

4.5 Classification

Classification is the decision making phase of a handwritten character recognition system. The classification phase is basically relies on the quality of the features extracted in the previous stage for deciding the i/p character belongs to which class.

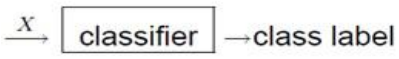

Figure 6. Classifier

X is feature vector. Classifier takes features as input and gives the corresponding character class as output. Classification depends upon the quality of features extracted. Irrelevant features may be removed for better classification accuracy.

There are many classification methods for handwriting recognition is as follows-:

1. Template matching
2. Statistical technique
3. Neural network(NNs) [18]
4. Structural techniques
5. Fuzzy-logic technique
6. Evolutionary computing techniques

4.6 Post processing

Post processing is the final stage of the recognition system. It prints the corresponding recognized character in the structured text form.

5. CONCLUSION AND FUTURE WORK

A lot of research is to be done to handle the Challenges in devanagari Character Recognition [20]. There are big challenges in handwritten character recognition due to different style of writer. Recent research is not directly concern to the characters, but also words and phrases, and even the complete documents. For the character recognition, HMM, neural networks and their combinations are used as the powerful tools. Character recognition, segmentation and classification can be used in an integrated manner for the high reliability and accuracy. This paper covers methodology used for handwritten character recognition using different features and different classifiers. Literature survey tells about the past research work done in devanagrai handwritten character recognition. This paper also describes the different stages used in handwritten devnagari character recognition.

Survey of devanagri handwritten character recognition can be extended to different classifier like neural network, evolutionary techniques on different features.

6. 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] U. Pal, B. B. Chaudhuri, "Indian Script Character recognition: A survey", Elsevier, Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

[3] R. G. Casey and E. Lecolinet, “A survey of Methods and Strategies in Character Segmentation “, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18, pp.690-706, 1996.

[4] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, “Statistical Pattern Recognition: A Review” IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 1, JANUARY 2000.

[5] Neha Sahu, Nitin Kali Raman “An Efficient Handwritten Devnagari Character Recognition System Using Neural Network” IEEE, 2013

[6] Sonika Dogra, Chandra Prakash, “PEHCHAAN: Hindi Handwritten Character Recognition System Based On SVM”, IJCSE, ISSN : 0975-3397 Vol. 4 No. 05 May 2012

[7] George Negi, “Twenty years of Document analysis in PAMI”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp- 38-62, January 2000.

[8] R. Plamondon and S. N. Srihari, “On-line and off-line handwritten recognition: a comprehensive survey”, IEEE Transactions on PAMI, Vol. 22(1), pp. 63–84, 2000.

[9] R.M.K. Sinha and Veena Bansal, "On Automating Trainer For Construction of Prototypes for Devanagri Text Recognition", Technical Report TRCS-95-232, I.I.T. Kanpur, India.

[10] R.M.K. Sinha and V. Bansal, "On Devanagri Document Processing", Int. Conf. on Systems, Man and Cybernetics, Vancouver, Canada, 1995.

[11] Veena Bansal and R.M.K. Sinha, “Integrating Knowledge Sources in Devanagari Text Recognition System”, Technical Report, TRCS-97-248, I.I.T. Kanpur, India, 1997.

[12] R.M.K.Sinha., “Rule Based Contextual Post-processing for Devanagri Text Recognition”, Pattern Recognition, Vol. 20, No. 5, pp. 475-485, 1987.

[13] R.M.K.Sinha, “On Partitioning a Dictionary for Visual Text Recognition”, Pattern Recognition, Volume 23, Issue 5, 1990, Pages 497–500.

[14] R.M.K. Sinha and Veena Bansal, "On Automating Trainer For Construction of Prototypes for Devnagari Text Recognition", Technical Report TRCS-95-232, I.I.T. Kanpur, India.

[15] R. M. K. Sinha, “A Journey from Indian Scripts Processing to Indian Language Processing”, IEEE Annals of the History of Computing, pp8-31, Jan–Mar 2009.

[16] I. K. Sethi and B. Chatterjee, "Machine Recognition of constrained Hand-printed Devanagri", Pattern Recognition, Vol. 9, pp. 69-75, 1977.

[17] R.M.K. Sinha, H. Mahabala," Machine recognition of Devanagri script", IEEE Trans.Systems Man Cybern. 9 (1979) 435–441.

[18] Brijesh k. Verma, "Handwritten hindi character recognition using multilayer perceptron and radial basis function in neural networks," IEEE International conference on Neural Networks,vol. 4,pp. 2111-2115, Nov.1995.

[19] Vyas, M. Verma, K.A "Comprehensive survey of handwritten character segmentation" IEEE, 2014

[20] Malik, L. ; G.H. Raisoni "A Graph Based Approach for Handwritten Devanagri Word Recogntion" IEEE,2012.

[21] Pal, U. ; Sharma, N. ; Wakabayashi, T. ; Kimura, F."Handwritten Numeral Recognition of Six Popular Indian Scripts" IEEE, 2007

[22] Pal, U., Sharma, N., Wakabayashi, T., Kimura, F. "Off-Line Handwritten Character Recognition of Devnagari Script" IEEE, 2007.