International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

RECOGNITION OF DEVANAGARI SCRIPT: A SURVEY

Asma A. Shaikh1, Rahul Dagade2, Marathwada Mitra Mandal's, College of Engineering Pune, India.
1asmamokashi@mmcoe.edu.in, 2 rahul.dagade@gmail.com

Abstract

Hindi is the National language of India. Devanagari Script is the most popular script in India which is used in the Hindi language. Three Hundred million people use Hindi Language in their day today activities. Hindi is third most popular language in the world. Devanagari script is also used for Sanskrit, Hindi, Marathi, Nepali and Konkani languages. The handwriting recognition area has been researched extensively till date, whereas recognition of Devanagari script is progressing area of research. So that it is necessary to learn the existing research for the recognition of Devanagari words.

Introduction

Handwriting recognition of words (HWR) is a system for converting the written text into actual words, which have an important role in many human computer interface uses, including mail sorting, office automation, cheque verification, as well as human–computer interface [1]. Recognition involves three stages, namely preprocessing, feature extraction and classification. First, words from input scripts are segmented and normalized in the preprocessing stage. Then, a set of features are extracted from each of the segmented words in feature extraction stage. Finally, these features are used for the classification.

In the offline recognition, only digital image is given as input, hence it is complicated to identify than on line handwritten word recognition , because in an online system it is possible to map out the process of writing, hence able to find the strong point and chronological order of each fragment when it is written can be recorded for recognition. Offline HWR includes the automatic conversion of text written in a scanned image into letter codes which are utilizable within computer and text processing application [2].

Types of Recognition of handwriting text:

A. An approach based on Segmentation:

An image is divided into lines, the lines are segmented into words and further segmented into characters or letters and these characters are then used for recognition. A word model is created from the concatenation of the character model. This approach is also identified as analytical approach.

B. Segmentation free approach

It is also famous as global approach, the word images are taken for recognition. A global approach makes the recognition simpler by avoiding the difficulty in character segmentation, but requires larger vocabulary when compared with an analytical approach. This model is also known as a holistic approach. Word models are built from word images without segmenting words [1].

1.1.Devanagari Script:

Most of the Indian languages, including Devanagari originated from the ancient Brahmi script through various transformations [3]. In Devanagari script, there are thirty six Consonants “Vyanjan” and 13 vowels “Swar” as shown in the Fig 1. When those thirty six consonants are attached with the vowels, then complex compositions of its constituent symbols are generated [4]. A vowel is followed by a consonant may generate a modified shape, which, depending on the vowels are placed before, after, top and bottom of the consonants are called modifiers or “Matras” because they are used to modify the consonants meaning as shown in Fig 2. There are total twelve modifiers. Devanagari script is a unicase script. There is no concept of upper and lower case letter as Latin script.

Words are written in the Devanagari script as they are pronounced so that script is called as phonetic script. Devanagari script is also called as syllabic script because the text is combinations of consonants and vowels that together from syllables [4].

Apart from this, composite & complex characters are formed by combining more than one basic character. Consonants can have a half form when they are joined with other consonants.

A key feature of the Devanagari script, upper horizontal line on the top of all characters is known as header line or “Shirorekha” [5]. This header line is used to divide the word into three parts, top, that is above the header line, contains top modifiers, bottom part contain lower modifier and core part contains the consonants as shown in Fig 1.

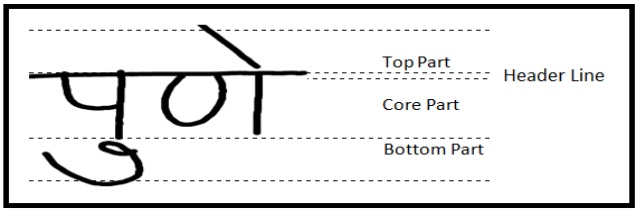

Fig. 1. Header line with three parts of Devanagari words.

Vowels [स्वर]	अ	आ	इ	ई	उ	Modifiers	ा	ि	ी	ु
	[1]	[2]	[3]	[4]	[5]		[1]	[2]	[3]	[4]
	ऊ	ए	ऐ	ओ	औ		ू	े	ै	ो
	[6]	[7]	[8]	[9]	[10]		[5]	[6]	[7]	[8]
	अं	अः	ऋ				ौ	ं	ृ	
	[11]	[12]	[13]				[9]	[10]	[11]	
Consonants [व्यनजन]	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
	ट	ठ	ड	ढ	ण	त	थ	द	ध	न
	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
	प	फ	ब	भ	म	य	र	ल	व	श
	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]	[29]	[30]
	ष	स	ह	ळ	क्ष	ज्ञ				
	[31]	[32]	[33]	[34]	[35]	[36]				

Fig. 2. Devanagari Alphabets: Vowels, Consonants & Modifiers

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	ऋ	अं	अः
क	का	कि	की	कु	कू	के	कै	को	कौ	कृ	कं	कः

Fig 3. Modified Alphabets when Consonants are attached with vowels.

1.2 Recognition System Consist of following stages:

1.2.1 Pre-processing:

The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately. The main objectives of pre-processing are:

1 Noise reduction
2 Binarization
3 Stroke width normalization
4 Skew correction
5 Slant removal

Binarization

Document image binarization (thresholding) refers to the conversion of a gray-scale image into a binary image. Two categories of thresholding:

- Global, picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.
- Adaptive (local), uses different values for each pixel according to the local area information

Noise Reduction - Normalization

Noise reduction improves the quality of the document. Two main approaches:

Filtering (masks) Morphological Operations (erosion, dilation, etc)

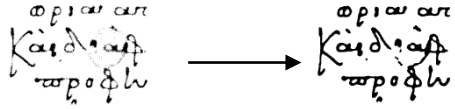

Fig 4. Image after Erosion & Dilation

Normalization provides a tremendous reduction in data size, thinning extracts the shape information of the characters.

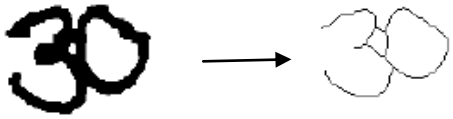

Fig 5. Image after thinning

Skew Correction

Skew Correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include correlation, projection profiles, Hough transform.

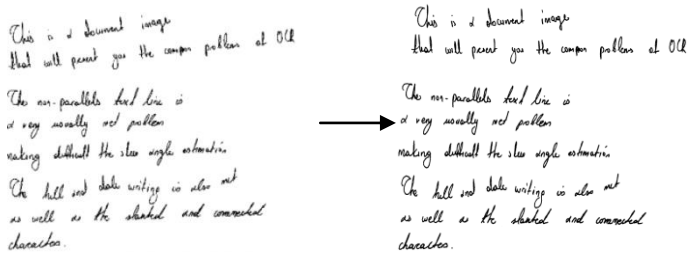

Fig 6. Image after skew correction

Slant Removal

The slant of handwritten texts varies from user to user. Slant removal methods are used to normalize the all characters to a standard form.

Fig 7. Image after Slant Correction

1.2.2 Feature Extraction

In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements.

Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods are based on 3 types of features:

[1] Statistical

[2] Structural

[3] Global transformations and moments

Statistical Features

Representation of a character image by statistical distribution of points takes care of style variations to some extent.
The major statistical features used for character representation are:

1. Zoning
2. Projections and profiles
3. Crossings and distances

Zoning:

The character image is divided into NxM zones. From each zone features are extracted to form the feature vector. The goal of zoning is to obtain the local characteristics instead of global characteristics

The character image is divided into NxM zones. From each zone features are extracted to form the feature vector. The goal of zoning is to obtain the local characteristics instead of global characteristics

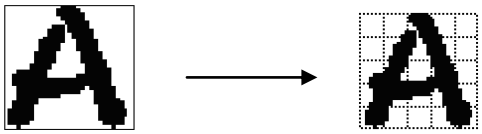

Fig 8. Character Zoning

Zoning – Density Features

The number of foreground pixels, or the normalized number of foreground pixels, in each cell is considered a feature.

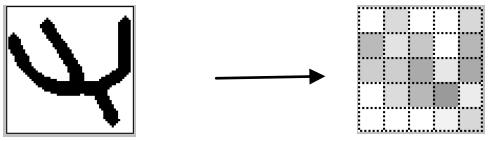

Fig 9. Character Zoning- Density Features

Darker squares indicate higher density of zone pixels.

Zoning – Direction Features

Based on the contour of the character image

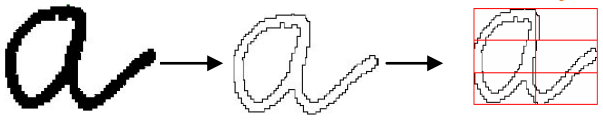

Fig 10. Character Zoning- Direction Features

Structural Features

Structural features are based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc.

Fig 11. Structural Features

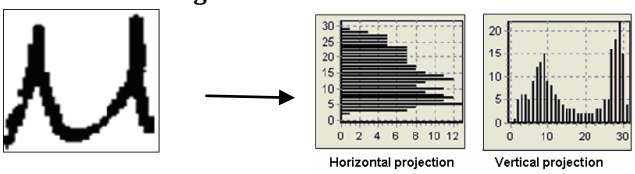

Fig. 12. Horizontal and Vertical projection histograms

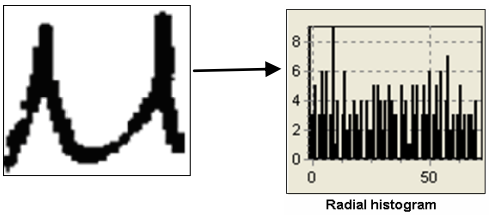

Fig 13. Radial histogram

Radial out-in and radial in-out profiles

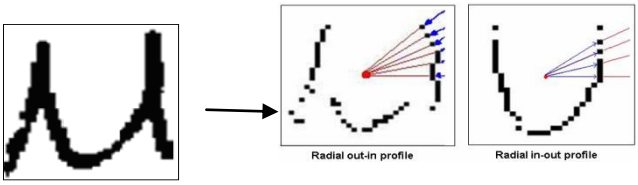

Fig 14. Radial histogram

1.2.3 Classification:

Classification is used for identification of Text.

1. k- Nearest Neighbor:
   The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.
2. Neural Network:

Neural network recognizers learn from an initial image training set. The trained network then makes the character identifications. Each neural network uniquely learns the properties that differentiate training images. It then looks for similar properties in the target image to be identified. Neural networks are quick to set up; however, they can be inaccurate if they learn properties that are not important in the target data.

3. Support Vector Machines:

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

4. Hidden Markov Model:

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states.In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

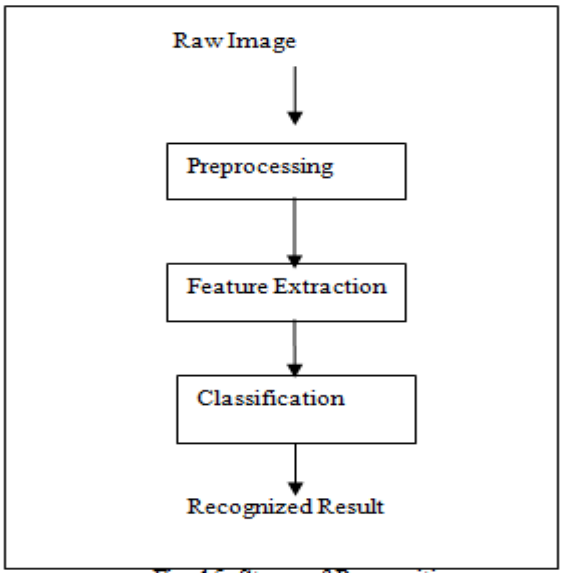

Fig. 15. Steps of Recognition

Literature survey:

Ved Agnihotri [2] proposed a new technique of Chromosomes function generation and fitness function for classification by extracting diagonal features from zones of an image. Handwritten Devanagari script recognition system using neural network is presented in this paper.

Diagonal based feature extraction is used for extracting features of the handwritten Devanagari script. After that these feature of each character image is converted into chromosome bit string of length 378. In the recognition phase and classification Genetic Algorithm is used

Jayadevan R. et.al [4] did a survey of the comparative study of recognition of printed as well as handwritten word recognition by different classification techniques like Artificial Neural Network, Hidden Markov Model, Support Vector Machine, MQDF

B.Shaw et.al [5] proposed a segmentation based approach for offline word recognition, vertical and horizontal stroke based features were extracted for recognition at a pseudo character level by HMM. A word level recognition was done on the basis of a string distance.

Naveen S. et.al [6] proposed a method of detecting Devanagari text of a printed document by mapping word directly to Unicode sequence. Here they consider Unicode as the main recognition unit and use a sequence transcription module to map the words features to corresponding Unicode. A variant of RNN known as BLSTM (Bidirectional long Short term memory) was used for the task.

The problem arise in Devnagari script recognition using Zernike moments, fuzzy rule and quadratic classifier provide less accuracy and less efficiency. For the solution of the above problem and for improve efficiency, Vedgupt Saraf [7] was used genetic algorithm for an excellent means of combining various styles of writing a character and generating a new style.

Bharat et.al [8] proposed method based on HMM for lexicon driven and lexicon free word recognition for online handwritten for feature extraction they used NPen++ feature for curliness, linearity and slop. The two different techniques for recognition of word written in Devanagari Text based on Hidden Markov Models (HMM): lexicon driven and lexicon free. The lexicon-driven technique models each word in the lexicon as a sequence of symbol HMMs according to a standard symbol writing order derived from the phonetic representation. The lexicon-free technique uses a novel Bag-of-Symbols representation of the handwritten word that is

independent of symbol order and allows rapid pruning of the lexicon.

U. Pal et.al [12] did a comparative study of four sets of different feature extracting methods and 12 different classifiers for handwritten character recognition. Projection distance, linear differentiates function, subspace method, modified quadratic discriminate function, support vector machine, Euclidian distance, image learning, nearest neighbor, modifies projection distance, compound projection distance and compound revised quadratic discriminate function were used as different classifiers.

Arora et.al [11] proposed a technique for combining three Multi_Layer Perceptron (MLP) best classifier for recognition of handwritten Devanagari characters using the intersection, shadow feature and chain code histogram features. This paper deals with a new method for recognition of offline Handwritten noncompound Devnagari Characters in two stages. One using neural networks and the other one using minimum edit distance. Each of these techniques is applied on different sets of characters for recognition. In the first stage, two sets of features are computed and two classifiers are applied to get higher recognition accuracy. Two MLP's are used separately to recognize the characters. For one of the MLP's the characters are represented with their shadow features and for the other chain code histogram feature is used. The decision of both MLP's is combined using weighted majority scheme.

Anoop Namboodiri [15] presented a method for online recognition of handwritten text by a K nearest neighbor and support vector machine classifier and sequential floating search method for feature extraction. It classified words and lines in an online handwritten document into one of the six major scripts: Arabic, Cyrillic, Devanagari, Han, Hebrew, or Roman.

Gunjan Singh et.al [19] proposed an offline handwritten Hindi character recognition system using neural network. Neural networks are good at recognizing handwritten characters as these networks are insensitive to the missing data. The paper proposed the approach to recognize Hindi characters in four stages— 1) Scanning, 2) Preprocessing, 3) Feature Extraction and, 4) Recognition. Feature extraction includes extracting some useful information out of the thinned image in the form of a feature vector. The feature vector comprises of pixels values of normalized character image. A Backpropagation neural network is used for classification.

K. Y. Rajput et.al [16] proposed to recognize the characters of handwritten text using neural network by replacing the recognized characters by standard fonts.

Ashutosh Aggarwal et.al. [20] proposed method to solve the problem of OCR System ie. handwritten, machine-print, grayscale, and binary and low-resolution character recognition, so that all sample images of Devanagari characters used are normalized to 90*90 pixel sizes. Then extracted features by using Gradient Feature Extraction are converted to Gradient Feature Vector. Then for classification Support Vector Machine, supervised Machine Learning technique is used

Veena Bansal et.al [17] presented technique of a tree classifier for recognition of Hindi handwritten character by using vertical feature bar, horizontal zeroes, crossing moments.

Mitrakshi B. Patil et.al [22] proposed method for recognition of offline handwritten Devanagari characters using segmentation and Artificial neural networks. The whole process of recognition includes two phases- segmentation of Characters into line, word and characters and then recognition through feed-forward neural network.

Swapnil A. Vaidya et.al [21] proposed method in which add all the sample character image matrices and divide the resultant matrix by total number of matrices added, called as Avg_matrix. Then subtract it from each sample character image matrix, which results in unique features because of their positional properties of pixels present in that image. They used singular value decomposition technique to get projection vector matrix then used generalized regression neural network for resulting feature vectors and obtain classification performance in the character recognition task.

Prachi Mukherji et.al[23] proposed method in which thinned character is segmented into segments (strokes), using basic structural features like endpoint, cross point, junction points and adaptive thinning algorithm. The segments of characters are coded using our Average Compressed Direction Code (ACDC) algorithm. The knowledge of script grammar is applied to identify the character using shapes of strokes, mean row and column co-ordinates, relative strength, straightness and circularity. Their location in the image frame is based on fuzzy classification. Characters are pre-classified using a tree classifier. Subsequently unordered stroke classification based on mean stroke features is used for final classification and recognition of characters.

Table1: Details of Offline Handwritten Recognition System

Sr. No.	Reference Paper	Types	Feature Extraction	Classifier	Dataset	Recognition Accuracy
1	Ved Agnihotri [2]	Offline Handwritten Devanagari	Diagonal based feature extraction	Genetic Algorithm	1000	85.78%
2	B.Shaw et.al [5]	Offline Handwritten Devanagari Word	Vertical and horizontal stroke based features	HMM	22500	81.63%
3	Arora et.al [11]	Offline Handwritten noncompound Devanagari Characters	Shadow feature and chain code histogram features	Three Multi_Layer Perceptron (MLP) best classifier	7154	90.74%
4	Gunjan Singh et.al [19]	Offline handwritten Hindi character recognition system	One-dimensional 49x1 vector form	Neural network	1000	93%
5	Swapnil A. Vaidya et.al [21]	Offline handwritten character recognition	Positional properties of pixels	Neural network	4500	82.89 %
6	Prachi Mukherji et.al[23]	Offline Devanagari Handwritten	Structural features like endpoint, cross point, junction points	Tree classifier	250	86.4%.

Table2: Details of Online Handwritten Recognition System

Sr. No.	Reference Paper	Types	Feature Extraction	Classifier	Dataset	Recognition Accuracy
1	Bharat et.al [8]	Online handwritten	NPen++ feature for curliness, linearity and slop	Hidden Markov Models (HMM): lexicon Driven and lexicon free.	9407	74.83%
2	Anoop Namboodiri [15]	Online recognition of handwritten text	Sequential floating search method	K nearest neighbor and support vector machine classifier	2,155	95.5 %
3	Veena Bansal et.al [17]	Online Hindi handwritten character	Vertical feature bar, horizontal zeroes, crossing moments	Tree classifier	12,000	90%

Table3: Details of printed text Recognition System

Sr. No.	Reference Paper	Types	Feature Extraction	Classifier	Dataset	Recognition Accuracy
1	Ashutosh Aggarwal et.al. [20]	Handwritten, machine-print, grayscale, and binary and low-resolution character recognition	Gradient Feature Extraction	Support Vector Machine, supervised Machine Learning technique	7200	94%
2	Naveen S. et.al [6]	Devanagari text of a printed document	Unicode sequence	BLSTM (Bidirectional long Short term memory)	1.5M	

Conclusion:

To recognized words written in Devanagari Script is useful for different applications in our Daily life. All the above research helps the new researchers to get the benefits from their ideas to generate new ideas for the recognition. Recognition of Devanagari Script is divided into three major parts – Preprocessing, Feature Extraction & Classification.

Reference:

[1] Jawad AlKhateeb, Jinchang Ren. Jianmin Jiang, Husni Al. Muhtaseb, "Offline handwritten Arabic cursive text recognition using Hidden Markov Models and Re-ranking" in Pattern Recognition vol.32, pp.1081-1088, 2011.

[2] Ved Prakash Agnihotri, "Offline Handwritten Devanagari Script Recognition" in MEC, pp. 37-42, 2012.

[3] U. Pal and B. B. Chaudhuri, "Indian script character recognition: A survey," Pattern Recognition., vol. 37, pp. 1887–1899, 2004.

[4] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil and Umapada Pal, "Offline Recognition of Devanagari Script: A Survey" in IEEE Transaction on Systems, Man and Cybernetics –Part C. Applications and Review, vol.41, pp-782-796, 2011.

[5] Bikash Shaw, Swapan Kr. Parui and Malayappan Shridhar, "Offline Handwritten Devanagari Word Recognition: A Segmentation Based Approach", 19th International Conference on Pattern Recognition (ICPR'08), December, 8-11, 2008, Tampa, Florida, USA.

[6] Naveen Shankaran, Aman Neelappa and C.V. Jawahar, "Devanagari Text Recognition: A Transcription based Formulation" in ICDAR, pp. 678-68, 2013.

[7] Vedgupt Saraf, "Offline Handwritten Character Recognition of Devanagari script uses Genetic Algorithm for Improve efficiency" in ICCSE, pp.161-164, 2013.

[8] A. Bharat and Sriganesh Madhavnath, "HMM – Based Lexicon Driven and Lexicon-Free word Recognition for Online Handwritten Indic Scripts" in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol-34, pp.670-682, 2012.

[9] Sandhya Arora and Debotosh Bhattacharjee, "Multiple classifier combination for Offline Handwritten Devanagari Character Recognition".

[10] Umapada Pal, T. Wakabayashi, F. Kimura, "Comparative study of Devanagari Handwritten Character Recognition using Different Features and Classifiers" in 10th ICDAR, IEEE, pp.1111-1115, 2009.

[11] Rajiv K., Deepak Bagal, T.S.Kamal, "Skew Angle detection of a cursive handwritten Devanagari Script character image", J.Indian Inst Sci, vol.82,pp 161-175, 2002.

[12] M.S. Khorsheed, "Off-Line Arabic Character Recognition – A Review", Pattern Analysis & Apps., vol.5, pp.31-45,2002.

[13] Shaw B. and S. K. Parui, "A Two Stage Recognition Scheme for Offline Handwritten Devanagari Words", Machine Interpretation of Patterns-Image Analysis and Data Mining, pp. 145-165, 2010.

[14] Desai A. and Dr. Latesh Malik, "A Modified Approach to thinning of Devanagari characters" in IEEE 2011, pp.420-423

[15] Anoop Namboodiri, "Online Handwritten Script Recognition", Pattern Analysis & Machine Intelligence, IEEE Vol.26 No.1, 2004

[16] K.Y. Rajput and Sangeeta Mishra, "Recognition and Editing of Devanagari Handwriting Using Neural Network", IEEE Colloquium and International Conference, Mumbai, Vol. 1, pp. 66-70

[17] R.M.K. Sinha and Veena Bansal, "On Devanagari Document Processing", IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada

[18] Jayadevan R ,Umapada Pal and Fumitaka Kimura , "Recognition of Words from Legal Amounts of Indian Bank Cheques", 12th International Conference on Frontiers in Handwriting Recognition,2011.

[19] Gunjan Singh, Sushma Lehri, "Recognition of Handwritten Hindi Characters using Backpropagation Neural Network",(IJCSIT) International Journal of

Computer Science and Information Technologies, Vol. 3 (4) , 2012,4892-4895

[20] Ashutosh Aggarwal, 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.

[21] Swapnil A. Vaidya, Balaji R. Bombade, "A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction", IJCSMC, Vol. 2, Issue. 6, June 2013, pg.179 – 186.

[22] Mitrakshi B. Patil ,Vaibhav Narawade , "Recognition of Handwritten Devnagari Characters through Segmentation and Artificial neural networks", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 1 Issue 6, August – 2012.

[23] Prachi Mukherji, Priti P. Rege, "Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition ", Journal of Pattern Recognition Research 4 (2009),pp 52-68.

Biographies

ASMA A. SHAIKH received the AMIE degree in Computer Science & Engineering from the Institutions of Engineers (India) Kolkatta, the M.C.A. degree the SMU University and the Pursuing M.E. degree in Computer Engineering from the University of Pune, respectively. Currently, she is a Lecture of MCA at University of Pune. Her teaching and research areas include Image Processing, Pattern Recognition and Cloud Computing .

RAHUL DAGADE is working as Assistant Professor in Department of Computer Engineering in Marathwada Mitra Mandal's College of Engineering, Pune. His research interest are image processing, video processing, computer vision, opensource contribution to OpenCV.

Acknowledgments

We are thankful to IJATER Journal for the support to develop this document. We are thankful to Prof. Dr. Jayadevan R., Department of Computer Engineering, for his valuable Guidance. With due respect, We thanks H.O.D Prof. Ram Joshi, Department of Computer Engineering, for his motivating support ,keen interest which kept our spirits alive all through. We are thankful to our ME Coordinator Prof. Harmeet Khanuja for her kind support.