A HYBRID FEATURE EXTRACTION AND RECOGNITION TECHNIQUE FOR OFFLINE DEVNAGRI HADWRITING

Poonam Sharma
Department of Computer Science
The NorthCap University
Email-Id: poonamsharma@ncuindia.edu

Shivani Sihmar
Department of Computer Science
The NorthCap University
Gurgaon 122001, INDIA
Email-Id: shivanisihmar@gmail.com

Abstract—: Optical character Recognition is a very important aspect of computer applications. In case of printed character the accuracy of recognition has improved, researchers are still working on improving the accuracy of handwritten script recognition. To achieve proper segmentation of characters, one has to consider problems like unequal height of characters, unequal spacing between characters, and the lower, upper modifiers if the segmentation will have error, it will increase at later stages. In this paper, a hybrid of feature extraction techniques has been used and segmentation approaches are used for recognizing the handwritten Hindi documents. Accuracy rate of 98% at word, 70% at characters and 60% at shirorekha removal is achieved. Recognition rate of 80-85 % is observed (abstract).

Keywords- Devnagari character recognition; Offline character recognition; Feature extraction, Segmentation, classifiers (key words).

I. INTRODUCTION

Handwriting Character Recognition' is an orderly activity through which the characters or any symbol in the given handwritten document is recognized. The user transmits the handwritten document to the computer system with the help of digital scanners, tablets, or other electronic devices. Then the computer system interprets the input into its language and displays the results. For automatically converting the character-based document into a different format which is understandable by computers; humans have been thinking of different algorithms and machines with the capability of reading and converting of these character-based documents into alternate format or media. System which converts digitized images of printed or handwritten manuscripts and/or typewritten documents into character-based files is known as 'Optical Character Recognition' [1].

There are several systems available for recognizing the English language. Hindi is a national language of India which is written using Devnagari script. OCR's for Devanagari script are still evolving. Still lot of work is required to develop a OCR system which have high accuracy rate. Handwritten character recognition for Devnagari script is a challenging task due to the complexity in the Hindi language and different handwriting style [2].

In this paper a recognition system for handwritten Devnagari script is developed by using Matlab's tool box of Neural Network. The aim of this paper is to enhance the accuracy of existing work done in handwritten Devnagari script. A GUI is made using Matlab's tool for the user to communicate with the system.

Outline of the paper is as follows: Section II explains basic characteristics of Devanagari script. Section III explains some existing research on offline character recognition. Section IV explains steps of handwriting recognition. Section V explains feature extraction techniques. Section VI explains recognition and classification techniques. Section VII explains the comparison of various techniques. Section VIII describes the conclusion and the future scope.

II. LITERATURE REVIEW

In 1979 [3], **Sinha et. al** used Structural approach for feature extraction and technique of syntactic Pattern Analysis for recognition. Importance of spatial association between the constituent symbols was showed by Sinha for better understanding of Devanagari script. In 1989 [4], **Jayanti et. al** used Statistical approach for feature extraction and Binary Tree to recognize the Devanagari script. Jayanti used two major features for printed Devanagari script i.e. horizontal lines and vertical lines. The other feature that was used was height to width ratio. For computer programming Binary tree is one of the fastest approaches for decision process making. In the year

1997 [5], **Chaudhary et. al** used statistical approach for feature extraction and tree classifier, and template matching for recognizing Devnagari script. Firstly tree classifier separates the compound characters into smaller groups. After feature extraction, template matching is used to recognize the constituent characters. The disadvantage in template matching is that template matchers would be prevented for even reading the limited number of letters because of the slight deviations in orientation, shape, and size. In 2003 [6], **Jawahar et. al** used Support Vector Machine for recognition and for feature extraction Principle Component Analysis(PCA) was used. In 2004 [7], **Govindaraju et. al** used Neural Networks along with Gradient approach for feature extraction. In multi- classification, 38 characters and 83 conjunct characters were considered. Four categories were made on the basis of their structural properties. In 2006 [8], K-Nearest neighbor was used by **Kompali,** for recognition of characters. GSC was used alongside for feature extraction. For recognition of handwritten characters, **Pal et al.** [9] used a modified quadratic classifier. Directional information was obtained from the arc tangent of the gradient and Gaussian filter formed the basis for the feature extraction.. The basis of features was directional information In 2008 [10], for higher accuracy in recognizing the characters having the same features, combination of two classifiers was used by **Pal et. al.** In 2008 [11], for recognition of handwritten Devanagari words, a segmentation-based approach was proposed by **Shaw et.** For extracting the features Chain code was used. For recognition of handwritten Devanagari script Hidden Markov model was used.. Not much work is reported toward handwritten character string (word) recognition of Devanagari.In 2009, for recognizing the Devnagari script Hidden Markov Model was used by **Natrajan** [12]. In 2010 [13] for classification of non-compound characters in handwritten Devnagari, combination of a minimum edit distance(MED) and two MLPs were used by **Arora et.al.** For this CH features and shadow features are computed. In 2013, [2] **Sahu et. al** used neural networks for recognizing the Devnagari characters. Using Back Propagation network Devnagari character set is trained and on the word set the testing is performed. Accuracy of the system is average. Difficulty in recognizing some characters is highlighted. And the data set is tested again and the characters which cannot be recognized are separated from the data set to be tested. In 2015, [14] features were extracted by using multiple feature extraction techniques and for recognizing the text neural network was used by **Dongre et.al** . From each image 72 structural features i.e. (9 x 8) are used to extract features. The features extracted are called zonal features. In a similar way, before partitioning, 9 global geometric features are extracted.First, confirm that you have the correct template for your paper size. Maintaining the Integrity of the Specifications

III. INTRODUCTION TO DEVNAGARI SCRIPT

'Devanagari' word is a combination of two words "DEVA"+ "NAGARI" which mean the "City of Gods" [15] In INDIA, Devanagari is the principle script. Devanagari script is used to write various official languages in India, like Sanskrit, Hindi, Bangla, Nepali, Sindhi, and Marathi etc. In the world, the third most popular language is Hindi [1]

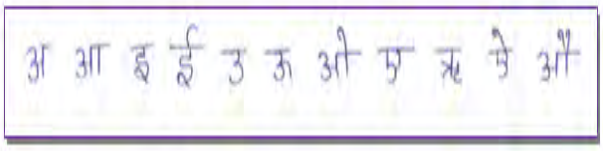

Figure.1 Swar in Hindi language

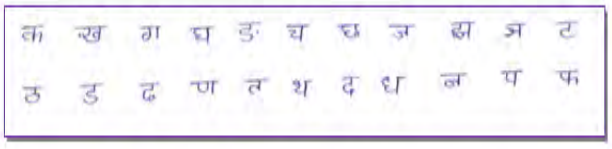

Figure. 2 Vyanjan in Hindi language

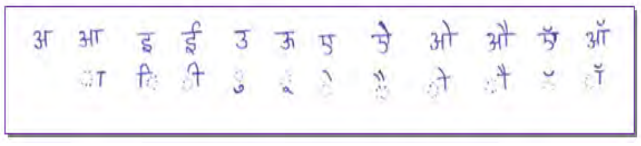

Figure. 3 Modifiers in Hindi language

Hindi language has 33 consonants, 13 vowels, and corresponding modifiers or we can say 'matras' known as the Basic characters. This language has compound characters as well which makes it even more complex which are formed by combining the basic characters [16]. There are some half consonants in this language which when combined with full consonants they form new characters known as 'Compound characters [17, 18]. Compound character's shape is more complex than the basic characters. If a vowel is followed by a consonant than a new

character or compound character is formed. Numerals, Text or characters all are written from left to right. There is no such thing as upper case or lower case. Devanagari script is phonetic and syllabic in nature. It is phonetic in a way that the words are pronounced exactly in the same way as they are written. And it is syllabic in a sense that vowels and consonants together form the syllables [16]. All the characters in this script are connected through a line known as 'Shirorekhha' or the Header line. The presence of compound characters, overlapping characters, conjuncts, makes it difficult to develop Devanagari OCR [19] .

IV. PROPOSED HANDWRITTEN SCRIPT RECOGNITION SYSTEM

A database is created for training and testing. For doing the training, The input image is scanned and provided as an input to the application. As, it consists noise, the image is enhanced for further processing using morphological operations of Matlab. After preprocessing is done, the image is gone under segmentation. Blob analysis is done on the image. With regionprops and boundingbox the image is segmented and a rectangular box is put on the character. After segmentation is done, Gabor features and Freeman chain code features are used to extract the features of the characters. An input and output vector pair is stored for the recognition purpose. Then neural network is used to recognize the script.

TRAINING:

To recognize the characters one part of this application trains the network. During training, input-output vectors are provided to the neural network. The weight array of the network is trained so as to minimize performance measure; i.e. reduce the errors using BPN (back propagation) algorithm.

Five images consisting of consonants, vowels, modifiers, some half consonants and some variations of consonants are used to train the network.

A. Preprocessing:

Preprocessing is a series of operations for enhancing the input image before further processing. The various operations done on the input image includes Binarization by which the gray scale image is converted to binary image using Otsu's thresholding method. By using the sobel operator, the image is enlarged using matlab's imdilate function. By applying the preprocessing, the image becomes suitable for segmentation.

B. Segmentation:

In the proposed application, in the training phase segmentation is done by using Matlab function regionprops which is used to measure the properties of the image and returns a matrix containing the properties of the image and then Bounding box is used to place a rectangle over the segmented characters. The image is resized in 720*360 pixels for making the image uniform for the feature extraction step.

C. Feature Extraction:

This step is very important for the correct recognition of characters. Characters are defined by the presence or absence of pixels which are known as features of characters. These features may include height, width, loops, lines or other features. In this, only the relevant information about the character is extracted to minimize the data and a feature vector is generated with scalar values.

In this proposed algorithm, hybrid of Gabor features and freeman Chain Code features are used.

STEPS:

1. First 50 features are extracted using Gabor features and next 50 features are extracted using freeman chain code method.

2. Then these features are extracted for all the letters for which the network has to be trained.

3. For each character a feature vector is calculated using the below two techniques.

4. These features vectors are manually pasted into a separate matrix. For each character an input vector is stored with its corresponding output vector to train the network.

TESTING:

A. Preprocessing:

Colored images are first converted into gray scale and then further techniques are applied which includes binarization which converts the gray scale image into binary image. The input image may contain gaps in lines, disconnected line segments, bumps etc. the distortion caused by this may include erosion, rounding of corners etc. before further processing it is necessary to remove all these imperfections. For this purpose, morphological techniques of preprocessing are being used. For skew detection and correction Skew angle is calculated and proper corrections are made by aligning the skewed lines horizontally using various transforms and Hu moments [20, 21].

B. Segmentation:

Segmentation is performed on the preprocessed image. First of all the lines are segmented, then words are segmented. Shirorekha is removed and finally characters are segmented. Histogram for the image is plotted. All the high intensity pixels i.e the white pixels are searched and find the rows having white pixels and change them to black pixels. Compute the area and put bounding box and save it to different file. The words will be segmented. After segmenting the words, shirorekha is removed by counting the number of black pixels per row and converting them to white pixels.now the same process that was applied for training is used for character segmentation.

C. Feature Extraction

Same way is used for extracting the features as in the training part.

D. Recognition:

Back propagation neural network is used for the recognition. An input vector and target vector is passed to the neural network and an image to be recognized is provided as an input to the network.

V. FINDINGS

Figure. 4 Input Image

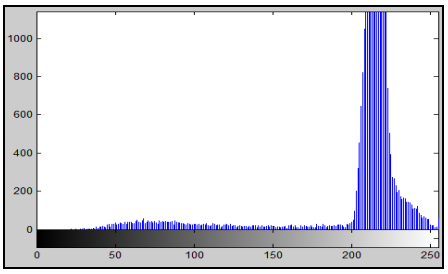

Figure.5 Histogram of Image

Figure.6 Dilated image

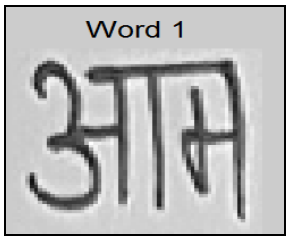

Figure.7 Segmented word as an image

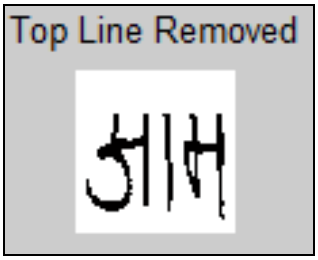

Figure.8 Shirorekha removed

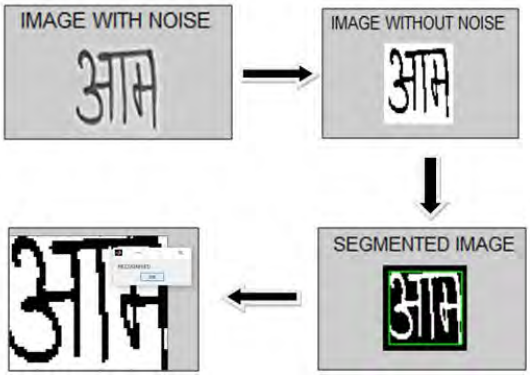

Figure. 9 Recognized word

VI. CONCLUSION AND FUTURE SCOPE

This application deals with the offline character recognition of Devanagari script. First of all, the scanned image is preprocessed for enhancement of quality using morphological operation and other techniques like skew detection and correction, thinning and Binarization and noise removal.

Three algorithms for segmenting the words, characters and removal of shirorekha are proposed. The results of Word segmentation algorithm are 98%. Almost all the words are segmented correctly and each word is stored in a different file as an image. After word segmentation, shirorekha of the individual word is removed. There is an issue while removing modifiers (matras) above the shirorekha. The shirorekha (header line) on words without upper modifiers is removed successfully. In the last, the characters are segmented, with an accuracy rate of 70% but still there is scope of improvement. Accuracy rate of 50-60 % at shirorekha removal is achieved.

Features are extracted using Gabor features and Freeman chain code. For training, each character's are extracted the vector formed is of 1 x 100, and for the 33 consonants it is a matrix of 33 x 100. For each character the vector is stored in the vector matrix through which the network is trained.

In the recognition phase, neural network with back propagation learning has been used. The weights are adjusted accordingly. The network is able to recognize the characters it is trained with. But due to the segmentation issue, the letters are not shown properly. Recognition rate achieved is 80-85%.

Due to issue in shirorekha removal, the character segmentation also suffers a bit. When removing shirorekha, a small portion of the word from the upper side also gets removed. More work can be done on the removal of shirorekha. This is the problem when we consider a paragraph. This application successfully segments and recognizes the individual characters with 70% accuracy.

FUTURE WORK:

Researchers can work on proper removal of shirorekha without affecting the upper modifiers to achieve higher accuracy. A system can be developed for handwriting analysis for identifying a person or a writer. The algorithm is tested against Hindi language. It can further be extended to recognize Gurumukhi, or Sanskrit etc.

VII. REFERENCES

[1] Deepa Berchmans, SS Kumar, "Optical Character Recognition: An overview and an Insight", International conference on control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014.
[2] Ms.Neha Sahu, Mr.Ntin Kali Raman,"An Efficient Handwritten Devnagari Character Recognition Using Neural Network", IEEE, 2013.
[3] R. M. K. Sinha and H. Mahabala, "Machine recognition of Devnagari script," IEEE Trans. Syst. Man Cybern., vol. 9, no. 8, pp. 435–441, Aug. 1979.
[4] K. Jayanthi, A. Suzuki, H. Kanai, Y. Kawasoe, M. Kimura, and K. Kido, "Devanagari character recognition using structure analysis," in Proc. IEEE-TENCON, 1989, pp. 363–366.
[5] B. B Chaudhuri and U. Pal, "AnOCR system to read two Indian language scripts: Bangla and Devanagari," in Proc. 4th Conf. Document Anal. Recogniton., 1997, pp. 1011–1015.
[6] C. V. Jawahar, P. Kumar, and S. S. R. Kiran, "Bilingual OCR for Hindi- Telugu documents and its applications," in Proc. 7th Conf. Document Anal. Recognition, 2003, pp. 1–5.
[7] V. Govindaraju, S. Khedekar, S. Kompalli, F. Farooq, S. Setlur, and R. Vemulapati, "Tools for enabling digital access to multilingual indic documents," in Proc. 1st Int.Workshop Document Image Anal. Libraries, 2004, pp. 122–133.
[8] S. Kompalli, S. Setlur, and V. Govindaraju, "Design and comparison of segmentation driven and recognition driven Devanagari OCR," in Proc. 2nd Int. Conf. Document Image Anal. Libraries, 2006, pp. 1–7.
[9] U. Pal, N. Sharma, T.Wakabayashi, and F. Kimura, "Off-line handwritten character recognition of Devnagari script," in Proc. 9th Conf. Document Anal. Recogniion., 2007, pp. 496–500.
[10] U. Pal, S. Chanda, T. Wakabayashi, and F. Kimura, "Accuracy improvement of Devnagari character recognition combining SVM and MQDF," in Proc. 11th Int. Conf. Frontiers Handwrit. Recognition, 2008, pp. 367–372.
[11] B. Shaw, S. K. Parui, and M. Shridhar, "A segmentation based approach to offline handwritten Devanagari word recognition," in Proc. IEEE Int. Conf. Inf. Technol., 2008, pp. 256–257.

[12] R. M. K. Sinha, "A journey from Indian scripts processing to Indian language processing," IEEE Ann. Hist. Comput., vol. 31, no. 1, pp. 8– 31, Jan./Mar. 2009.
[13] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, and M. Kundu, "Recognition of non-compound handwritten Devnagari characters using a combination ofMLP and minimum edit distance," Int. J. Comput. Sci. Security, vol. 4, no. 1, pp. 1–14, 2010.
[14] Vikas J. Dongre, Vijay H. Mankar,"Devanagari Offline Handwritten Numeral and Character Recognition using Multiple Features and Neural Network Classifier", @ IEEE, 2015.
[15] Gayathri P, Sonal Ayyappan,"Off-line Handwritten Character Recognition using Hidden Markov Model", IEEE, 2014.
[16] R.Jayadevan, Satish R. Kolhe, Pradeep M. Patil and Umapada Pal,"Offline Recognition of Devnagari Script: A Survey",IEEE transactions on systems, man, cybermatics—part c: applications and reviews, vol.41,no.6,November 2011.
[17] Poovizhi P," A Study on Preprocessing Techniques for the Character Recogntion", International Journal of Open Information Technologies, 2014.
[18] Akanksha Gaur, Sunita Yadav," Handwritten Hindi Character Recognition using KMeans Clustering and SVM", 2015, International Symposium on Emerging Trends and Technologies in Libraries and Information Services, IEEE, 2015.
[19] Nisha Sharma, Tushar Patnaik, Bhupendra Kumar," Recognition for Handwritten English Letters: A Review", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.
[20] Mohamed Cheriet, Nawwaf Karma, Cheng-Lin Liu, Ching Y. Suen, (2007), " Character Recognition System: A guide for students and Practitioners, " John Wiley & Sons, Inc., Hoboken, New Jersey, 2007.
[21] U.Pal, M.Mitra and B.B. Chaudhary, (2001) "Multi-Skew Detection of Indian Scripts Documents", CVPRU IEEE, pp 292-296.