DEVANAGARI SCRIPT SEPARATION AND RECOGNITION USING MORPHOLOGICAL OPERATIONS AND OPTIMIZED FEATURE EXTRACTION METHODS

Sushilkumar N. Holambe
Persuing PhD at
Dr. B.A.M.U. Aurangabad

Dr. Ulhas B. Shinde
Principal at
CSMSS chh.shahu college of
Engineering , Aurangabad

Shrikant D. Mali
persuing ME at
TPCT's College of
Engineering,Osmanabd

ABSTRACT

Now days handwritten recognition systems increasingly used for automatic document scanning and analysis purpose. Hence from last two decades this becomes challenging area for researchers. Using semi-automated or automated methods the machine printed documents and scanned documents are recognized which is called as handwritten recognition. Number of methods has been proposed so far for different handwritten language such English, Hindi, Devanagari etc. with their advantages and disadvantages. Devanagari language in India is mainly used for information communication purpose after English, especially in Indian government processing's. The existing methods of Devanagari script separation and recognition based on different segmentation and feature extraction methods having limitations in terms of recognition time, accuracy etc. The recent method presented Devanagari script separation and recognition is done based on morphological operations and zone based features without using classifier. Using particular threshold values classification is done which is limitation for such automated systems. In addition to this approach is only based on zone based features. In this project, novel approach is designed for script separation and recognition based on morphological operations for efficient segmentation of Devanagari script, optimized features extraction approach using zonal features, texture and directional features, and then applying neural network classifier for recognition and accuracy evaluation.

Keywords

Devanagari, Document Scanning, Feature Extraction, Morphological Operations, Classification, Script Recognition, Script Separation

1. INTRODUCTION

In daily life applications like government documents processing, educational documents processing, private industries and banking processing, the documents in form of hard copies are processed in form of soft copies based on electronic media. The use of soft copies delivers the immediate, secure and instant approach for documents processing, sharing and storing. However still there are number of transactions in which hard copies of documents are preferred and widely used. The majorly and vital approach of processing and sharing such documents is fax machines. To ensure the physical documents use for longer period for further analysis, paper is appropriate approach which is easy and secure to handle such communications. But handling large number of papers in day is time consuming, tedious and cost consuming option. Hence the automated approach should be there to capture such hard copies of documents, retrieve information from it and analyze the retrieved information for further processing. This is done by using the image processing terminologies. This automated framework is falls under the domain of document image analysis. Since from last 15-20 years of period document image processing and analysis gained significant attention from research groups [1].

The goal of document image processing and analysis is nothing but character information reading automatically from the document in image form. The reading of characters from document image is done by OCR (Optical Character Recognition) [2]. The processing of reading the scanned physical documents information through machine is called as OCR. The existing OCR is having implicit assumption that type of script that has to be processed is aware before processing. But for automated applications and environment, this type of document processing techniques depending on human intervention in order to choose particular OCR type, and this becomes very inefficient, impractical and undesirable for end users.

2. RELATED WORKS

The goal of this section is to discuss the methods those are previously proposed by various researcher groups by considering recognition accuracy and speed for Devanagari handwritten script recognition. These methods are contributed under three different phases of recognition system like feature extraction, image segmentation and recognition.

In [3], this paper reported horizontal/vertical strokes and end points as the potential features presented for the recognition. This method was having accuracy of 90.50% for handwritten Kannada numerals. But the limitation of this method is that it uses the thinning process which results in the loss of features.

In [4], this paper presented method with three different kinds of features like moment features, density features, descriptive component features etc. This method is presented for Devanagari numerals classification using multi classifier connectionist framework for improving the accuracy of recognition and reliability, they obtained 89.6% accuracy for handwritten Devanagari numerals.

In [5], this paper presented new method of zoning & the directional chain code functions & it has been assumed as a variations vector with size of 100 for handwritten numeral recognition. This method is having better accuracy as compared to previous methods. But the feature extraction process is very complex as well as time consuming.

In [6], this paper has proposed zoning & the directional chain of the code attributes & the known as the verities of the vector of length 100 for handwritten numeral recognition & the have been pointed out of the higher level of recognition accuracy. Whatever, the feature extraction method is being hard & time consuming.

In [7], this paper using Input Fuzzy Modelling for the Recognition of Handwritten Hindi Numerals. This research shown the recognition of the Handwritten Hindi Numerals basis on the modified exponential membership function comfort with fuzzy sets which is derived from the normalized distance features obtained with the use of Box technique. This study has obtained 95% recognition accuracy.

In [8], this paper study the relevance of stroke size and position information for the recognition of the online handwritten Devanagari words by distinct of the three various pre-processing schemes. Experimental results indicate that the word recognition accuracy achieved using a pre-processing scheme which is totally disregards of the main sizes & positions of the strokes.

In [9], this paper presented offline the handwritten Devanagari words recognition: a segmentation based methods novel this segmentation based approach is advanced for recognition of offline handwritten Devanagari words. Stroke based methods & features has been used such as the feature of the vectors A hidden Markov model is used for recognition at pseudo character level. The level of the words has been recognition is to complete on basis on of a string edit distance.

3. PROPOSED APPROACH

There are number methods are targeted to Devanagari scripts due to the fact of using Devanagari language by 600 million people’s daily in their communications. In world, third most used language is Devanagari. This language is used with other major languages such as Marathi, Hindi, Sanskrit, as well as Nepali. The complexity of Devanagari language is more as compared to English due to the various differentiations in writing of different characters which are composed of various order, direction, number, shape, strokes etc. Like English, Devanagari script is also having total 50 numbers of different characters those can be used to construct words. There are very number approaches reported on Devanagari handwritten recognition in literature but having limitations in different views. The main problem of Devangari scripts recognition is the efficient segmentation and script separation.

We proposed novel approach for handwritten Devanagari script recognition using script separation and segmentation methods. First step of handwritten recognition is pre-processing and segmentation. Pre-processing is done for noise removal, resizing and conversion into image which is suitable for segmentation. After that segmentation is perform to extract the real information data from input image in order to do further processing. The second step of handwritten recognition is feature extraction. The selection of feature extraction technique is main factor for delivering highest handwritten recognition accuracy. There are many methods for feature extraction presented in literature. The most commonly used feature extraction methods are Gradient features, structural features, regional features, projection histograms, Zernike moments, zoning etc. Many of methods may require more time for feature extraction, however delivering better accuracy. But in handwritten character recognition, we require both high accuracy and less time.

4. METHODOLOGY

Figure 1 showing the detailed process of Devanagari handwritten recognition based on proposed methodology introduced in this paper. In this section, architecture and algorithm designs are presented for proposed approach. As showing in figure first step is image acquisition in which the handwritten scripts in machine image format is given as input to proposed system. On this input pre-processing is performed, then segmentation, then feature extraction and finally recognition. For each step specialized methods has been designed which are showing in below paragraphs of this section.

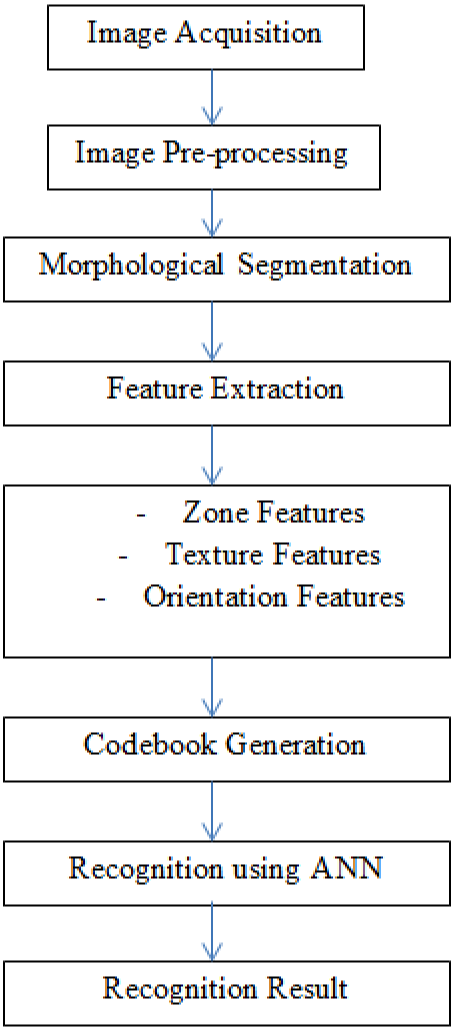

Figure 1: Detailed Processing of Proposed Devanagari Handwritten Script Separation and Recognition

Algorithm 1: Image Acquisition and Pre-Processing

Input: Handwritten Devanagari Image

Output: Pre-processed Image

Step 1: Browse input handwritten image

Step 2: Convert RGB image to grayscale image

Step 3: Resize image

Step 4: Apply Laplacian and mean filtering for image denoising.

Algorithm 2: Morphological Image Segmentation

Input: Pre-processed Image

Output: Segmented Image

Step 1: Apply morphological binary operation on pre-processed image.

Step 2: Apply canny edge detection on binarized image.

Step 3: Apply Dilation Operation

Step 4: Border Removal

Step 5: Apply Morphological Erosion Operation

Step 6: Apply Character Segmentation Vertically and Horizontally

Algorithm 3: Optimized Feature Extraction

Input: Segmented Handwritten Image

Output: Features Codebook

Step 1: Texture Features Extraction using GLCM

Step 2: Extracted texture features are contrast, energy, correlation and homogeneity

Step 3: Zonal features extraction [Algorithm 4]

Step 4: Orientation Features Extraction

Step 5: Apply mean and standard on extracted features

Step 6: Apply fusion operation on all extracted features

Step 7: Generate final codebook of features.

Algorithm 4: Zonal Features Extraction

Input: Segmented handwritten image.

Output: Vector of zonal features

Step 1: Apply skeletonization.

Step 2: Apply zoning by divided image into 9 equal size zones.

Step 3: Extract starters, intersections, and minor starters.

Step 4: Line segments extraction for each zone.

Step 5: Line type detection from line segments such as horizontal, vertical, right diagonal and left diagonal etc.

Step 6: Finding total number of each line type.

Step 7: Finding normalized length of each line type.

Step 8: Store all features in vector.

The recognition is done by using neural network classifier which is responsible to recognize the input handwritten script text by matching with training features. We have proposed Feed Forward Neural Network (FFNN) classifier in place of SVM to improve the accuracy, efficiency and time complexity performances. Below are algorithms for training and recognition:

Algorithm 5: Training using FFNN

Step 1: Input feature matrix M from T reading at layer 1 Fi.

Step 2: Computation of activation value for every neuron [ANi].

Step 3: Searching neuron with maximum ANi value.

Step 4: Extract the step 3 results with its input_id and max_ANi_index.

Step 5: Output Ok is set to 1 for kth neuron who’s having maximum ANi value.

Step 6: Else set output to 0.

Step 7: Feed the input of previous layer to next layer still to the output layer.

Step 8: Repeat above steps for all input layers.

Algorithm 6: Classification Algorithm

Step 1: Read test pattern to be recognized or classified

Step 2: Compute activation value ANi during layer 2

Step 3: Selecting neuron with max ANi.

Step 4: Neuron with max ANi index is extracted and save it as input_id and max_ANi_index for purpose of matching.

Step5: If match is successful, then input_id of max ANi is returned as output.

Step6: Stop

5.RESULTS AND DISCUSSION

The practical work is conducted by using MATLAB simulation tool. MATLAB is powerful and widely used tool for image processing applications. The proposed algorithms are simulated and validated using dataset of 6 different handwritten words written by 30 different persons in different styles.

Total samples collected are 600 handwritten images (100 for each Devanagari word) which are further divided into two parts training (80%) and testing (20%). The test samples are given as input to our designed system in order to recognize them correctly. Below are some samples of handwritten Devanagari script. The performance is measured in terms of two main performance parameters such as recognition accuracy and recognition time using below formulas.

Recognition accuracy = (TP+TN / TP+TN+FN+FP)

Recognition accuracy (%) = (TP+TN / TP+TN+FN+FP) * 100

Recognition time (seconds) = end_time – start_time

Figure 2 and is Figure 3 showing the example Devanagari scripts from dataset.

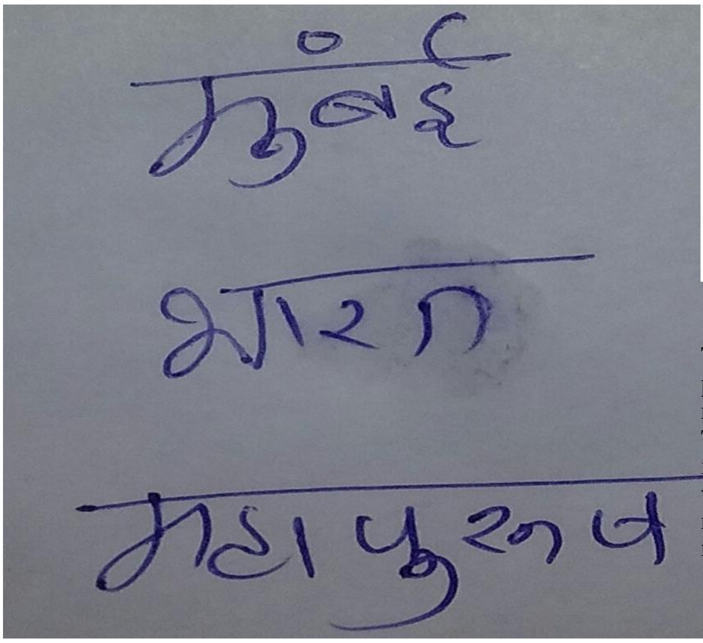
मुंबई
भारत
महापुरुष

Figure 2: Examples of Devanagari Handwritten Scripts

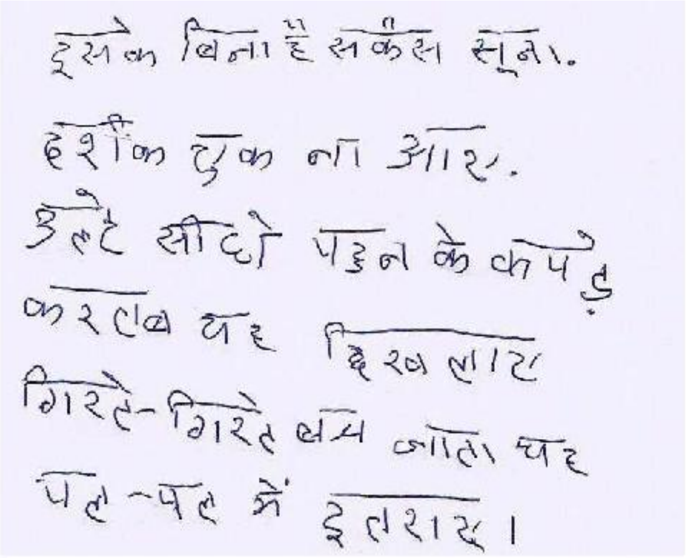

Figure 3: Input Handwritten Paragraph for Recognition

The comparative study between existing method which is recently reported in [1] by author Sukhvir Singh et.al and proposed method which is presented in this paper for Devanagari script recognition is presented in below two graphs. This study is conducted by varying size of training samples in each category.

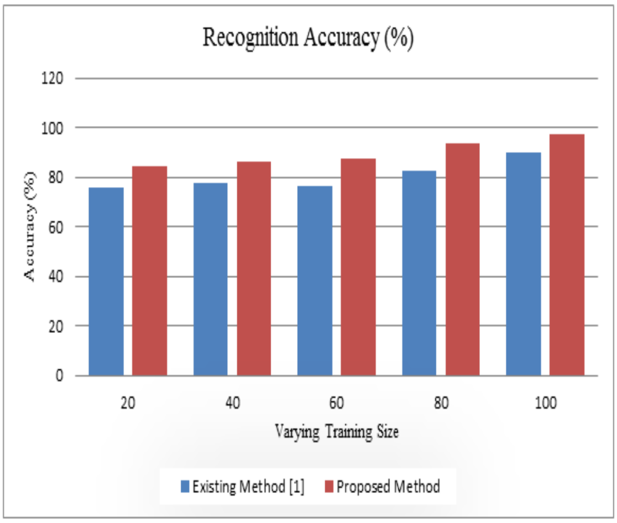

Figure 4: Recognition Accuracy Comparative Analysis

The graphical results showing in figure 3 and 4 showing the proposed approach for Devanagari handwritten script recognition outperforming the existing method reported in [1]. The accuracy is increasing as the number of training samples increases due to the possibility of recognizing the accurate words increases. Similarly as the training size increases, the recognition time is also getting higher. For proposed method recognition is very less as compared to existing method.

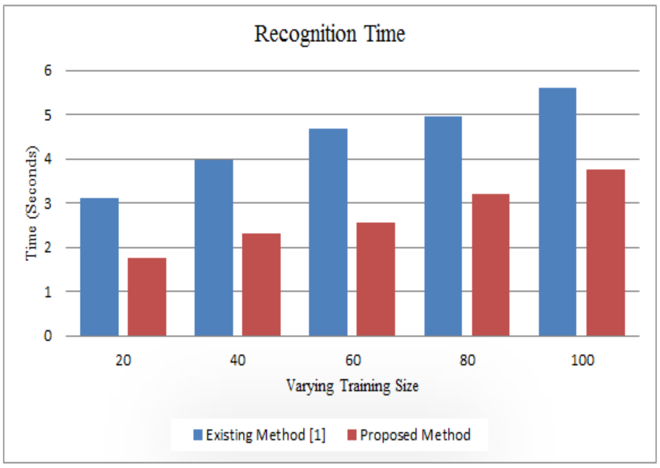

Figure 5: Comparative Study of Recognition Time

4. CONCLUSION AND FUTURE WORK

Now days handwritten script recognition automatic framework or tools are widely used in real time applications for faster and efficient documents processing. Devanagari is largely used communication language in India; hence there must be an efficient tool for automatic Devanagari script separation and recognition. In this paper, new approach is designed for Devanagari script recognition using morphological operators, optimized feature extraction methods and classifier. The simulation results are showing the proposed method outperforming the existing method in terms of recognition accuracy and time. The recognition accuracy of proposed approach is improved by 50 % approximately as well as recognition time minimized by approximately 40 % as compared to existing methods. For future work, we suggest to elaborate and evaluate the performance of this approach using large number Devanagari words and samples.

5. REFERENCES

[1] Sukhvir Singh, Anil Kumar, Dinesh Kr. Shaw and D. Ghosh, “Script Separation in Machine Printed Bilingual (Devnagari and Gurumukhi) Documents Using Morphological Approach”, IEEE 2014.

[2] D. Ghosh, T. Dube, and A.P. Shivaprasad, “Script recognition – a review,” IEEE Trans. Pattern Analysis & Machine Intelligence, vol. 32, no. 12, pp. 2142–2161, Dec. 2010.

[3] Dinesh Acharya U, N V Subba Reddy and Krishnamurthy, “Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster,” IISN-2007, pp-125 -129.

[4] Reena Bajaj, Lipika Dey, and S. Chaudhury, “Devanagari numeral recognition by combining decision of multiple connectionist classifiers”, Sadhana, Vol.27, part. 1, pp.-59-72, 2002

[5] N. Sharma, U. Pal, F. Kimura, "Recognition of Handwritten Kannada Numerals", *9th International Conference on Information Technology (ICIT'06)*, ICIT, pp. 133-136.

[6] N. Sharma, U. Pal, F. Kimura, "Recognition of Handwritten Kannada Numerals", 9thInternational Conference on Information Technology (ICIT'06), ICIT, pp. 133-136.

[7] M. Hanmandlu, J. Grover, V. K..Madasu, S. Vasikarla“ Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals ” International Conference on Information Technology (ITNG'07) 0-7695-2776-0/07 ,2007 IEEE.

[8] Devanagari Word Recognition: An Empirical Study.

[9] Bikash Shaw et al “offline hand written Devanagari word recognition: a segmentation based approach. “978-1-4244-2175 6/08/$25.00 ©2008 IEEE.

[10] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya: Character Recognition Using Matlab‟s Neural Network Toolbox, International Journal of u- and e- Service, Science and Technology Vol. 6, No. 1, February, 2013

.