NOVATEUR PUBLICATIONS
International Journal of Research Publications in Engineering and Technology [IJRPET]
ISSN: 2454-7875
VOLUME 3, ISSUE 4, Apr.-2017

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

SHRIKANT D. MALI
PG Student, Department of Computer Science & Engineering, TPCT's College of Engineering, Osmanabad- 413501, Maharashtra, India

DR. ANILKUMAR N. HOLAMBE
HOD & P.G. Coordinator, Department of Computer Science & Engineering, TPCT's College of Engineering, Osmanabad-413501, Maharashtra, India

ABSTRACT:

For different applications handwritten recognition frameworks 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 this approach is only based on zone based features. In this paper, 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 analyse 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. If the document itself having different types of languages, then document analysis as well as recognition becomes more challenging and complex, as OCR needs to select any one type of language before processing that document. To mitigate such research challenges, language type detection and recognition methods has been presented recently [3]. There are two types of handwritten recognition such as offline and online under the OCR domain. There are number applications of ORC like vehicle number plate recognition, criminal investigations, bank documents recognition, digits recognition, postal address block detection and recognition etc. Which factors are considered for evaluation of efficient handwritten recognition systems? Accuracy and speed are two main performance metrics are the answers to this question. These two performance metrics are important in evaluating the handwritten recognition system efficiency for variety of OCR applications. The speed and accuracy of handwritten character recognition efficiency is mainly based on use of feature extraction methods and classification methods [2] [4].

In this paper, novel approach proposed for the Devanagari script separation and recognition based efficient methods for segmentation, feature extraction and recognition. This approach is 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 goal of this paper is to introduce new algorithm and methods for accurate and less time consuming approach for Devanagari script separation and recognition. The proposed framework is categorized into four main functions such as pre-processing, image segmentation, feature extraction and finally recognition. For pre-processing Laplacian and Mean filtering methods used over the grayscale and resized handwritten images. For efficient and accurate segmentation, morphological operations such as dilation, binarizations etc. are used [5]. The segmented output is used for extraction of features like zone features, orientation features, texture features. There are three types of features extracted from segmented text. These features are combined using fusion operation in order to generate final codebook of feature vector [6] [7]. At the recognition step, codebook of features is input to classifier which performs the task of recognition with outcome of recognized words. The reminder of paper is having sections such as section II presenting the study over the existing methods of Devanagari script recognition using different algorithms and techniques. Section III presenting the detailed flowchart and algorithm design for each step of proposed approach. Section IV presenting the results in terms of accuracy and recognition time. At last conclusion and suggestions are discussed in section V of this paper.

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.

In [10], this paper presented method for English character recognition. The English character from input scanned document is recognized by using neural networks. In this paper author presented extensive study over different feature based methods of classifications for recognition of offline English character. In optical character recognition feature extraction is most important phase. Now days a recent method for the automatic handwritten recognition performing well either in achieving speed efficiency or achieving accuracy efficiency, but not both. But these methods are replacing existing methods of OCR like for English script. Therefore in [10], author is introduced simple, accurate method for recognizing optical characters by using the neural networks.

In [11], this paper introduced the new handwritten character recognition system using hybrid feature extraction method and support vector machine (SVM). This system is composed of three phases such as pre-processing, extraction of features and training as well as classification using SVM. Author used hybrid feature extraction approach which resulted into more accuracy. These features are able to extract local as well as global variations in handwritten character font styles. The extracted feature vector was a combination of correlation function based features and some statistical/structural features. But the limitation of this method is that it takes more time for recognition.

In [12], this paper proposed multi-layered neural network classification method for recognizing the handwritten characters. From the results showing in paper, the simulation of this work is done by using MATLAB GUI. Author designed this GUI in such way that end user can either test or train network on the basis of at time one character. The size of feature vector used for this method is 6 and this satisfying to achieve accurate identification of characters with multi layered neural network.

In [13], this paper introduced the recent method for Devanagari script character recognition using genetic algorithm, this method showing accuracy of 97 % in an average.

3. PROPOSED APPROACH:

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.

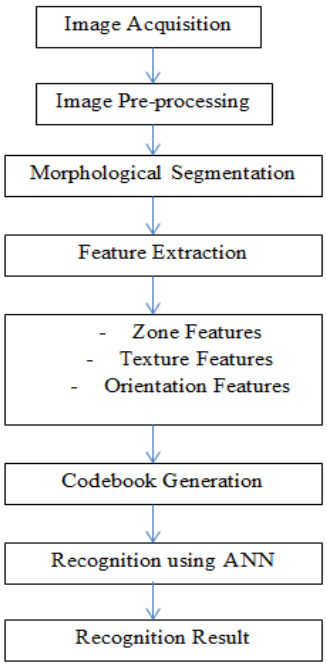

Figure 1: Detailed Processing of Proposed Devanagari Handwritten Script Separation and Recognition

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.

4. 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 is showing the example Devanagari scripts from dataset.

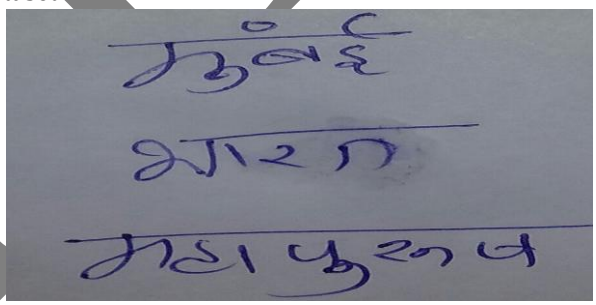

Figure 2: Examples of Devanagari Handwritten Scripts

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.

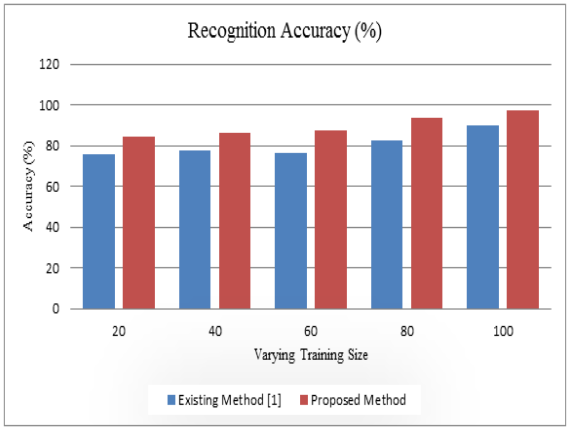

Figure 3: Recognition Accuracy Comparative Analysis

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.

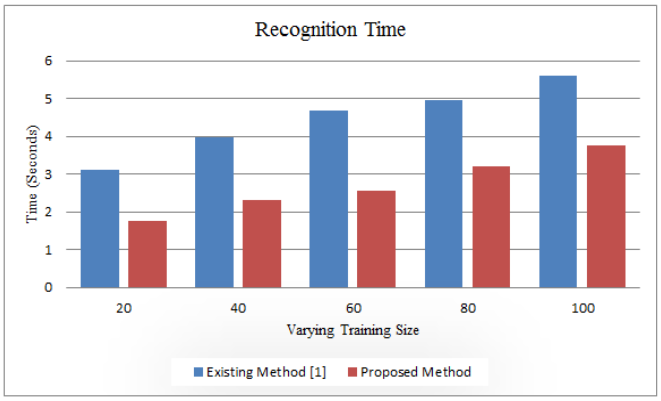

Figure 4: Comparative Study of Recognition Time

5. 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.

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
11) Muhammad Naeem Ayyaz, Imran Javed and Waqar Mahmood 1 2: *Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction,* Pak. J. Engg. & Appl. Sci. Vol. 10, Jan., 2012 (p. 57-67).
12) Sunit Bandaru *"Handwritten character recognition using neural network".*
13) Vedgupt Saraf, D.S. Rao, *"Devanagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency*", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, April 2013.