RESEARCH ARTICLE OPEN ACCESS

An Optical Character Recognition for Handwritten Devanagari Script

Trupti R. Zalke[1], Vasant N. Bhonge[2]

[1] M.E. Student, Department of Electronics and Telecomm, Engineering, SSGM College of Engineering, SHEGAON 444203 India
[2] Associate Professor, Department of Electronics and Telecomm. Engineering, SSGM College of Engineering, SHEGAON 444203 *India*

ABSTRACT
Optical Character Recognition is process of recognition of character from scanned document and lots of OCR now available in the market. But most of these systems work for Roman, Chinese, Japanese and Arabic characters . There are no sufficient number of work on Indian language script like Devanagari so this paper present a review on optical character recognition on handwritten Devanagari script.

***Keywords* –** Feature Extraction, Handwritten Document, OCR, Preprocessing, Segmentation

I. INTRODUCTION

Handwriting recognition of words is a system for converting the written text into actual words, which have an important role in many human computer interface uses. Handwritten character recognition is an important and challenging filed of Optical Character Recognition (OCR) handwritten character recognition is a difficult problem due to the great variations of writing styles, different size of the characters. Multiple types of handwriting styles and so on so it is major area of research.

1. Optical Character Recognition

OCR is process of conversion of printed or handwritten scanned document into machine encoded text. Computer system equipped with such system can improve the speed of input operation and decrease some possible human error. The process of OCR involves several steps including segmentation, feature extraction, and classification. OCR finds its applications in a wide area. Some of the important areas are as automatic number plate recognition, sound recording and reproduction.

2. Indian Language Characteristics

India is a multi-lingual and multi-script country comprising of Eleven different scripts like Devnagari, Gurumukhi, Tamil, Bangala, Oriya and so on .In all Devanagari is third most widely used script.Devanagari is used for several major languages such as Hindi, Sanskrit, Marathi and Nepali. It is Useful for writing and documentation purpose of most of the Indian languages.

3. Characteristics of Devanagari Script

In India more than 300 million people around the world use Devanagari script. This script forms the foundation of Indian languages. So Devanagari script plays a very major role in the development of literature and documentation purposes. Most of Indian script including Devanagari Originated from ancient Brahmi script. This script has complex composition of its constituent symbols. It has about 11 vowels and 33 consonants along with 14 modifiers. Devanagari script is written from left to right and it does not have any upper or lower case letters. It is Phonetic and syllabic script. Phonetic means words are written exactly as they pronounced and Syllabic means text is written using consonant and vowels that together form syllables A horizontal line on the top of all characters known as 'Header line or Shirorekha'

II. LITERATURE REVIEW

Ashutosh Aggarwal et.al. proposed system on Handwritten Devanagari Character Recognition Using Gradient Features. In this features are extracted by using Gradient Feature Extraction then converted into Gradient Feature Vector. For classification purpose Support Vector Machine, supervised Machine Learning technique is used. accuracy achieved in this is 94%.

Naveen Shankaran et.al. presented a Devanagari Text Recognition: A Transcription Based Formulation .they Proposed a method of detecting Devanagari text of a printed document by mapping word directly to unicode sequence. And for classification BLSTM (Bidirectional long Short term memory)classifier is used.

Vijaya Rahul Pawar presented a Performance Evalution of Multistage Offline Marathi Script

Recognition System, they proposed work on an artificial neural network based classifier and statistical and structural method based feature extraction approach is used for the recognition of the script. For classification purpose Self organizing map (SOM) is used.accuracy achieved in this project is 93%.

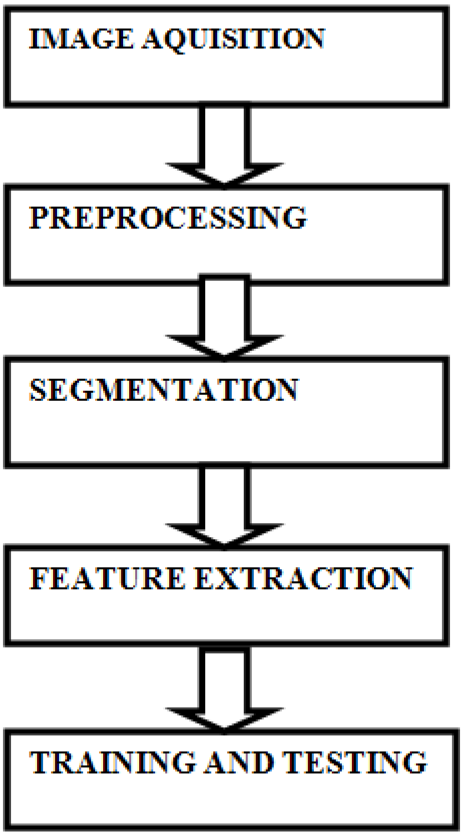
Jayadevan R. et.al [4] did a survey of the comparative study of recognition of printed as well as handwritten word recognition by different classification techniques like Artificial Neural Network, Hidden Markov Model, Support Vector Machine, MQDF Bharat et.al [8] proposed method based on HMM for lexicon 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] presented a comparative study of four sets of different feature extracting methods and 12 various 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.

Veena Bansal et.al [17] presented A Complete OCR for Printed Hindi Text in Devanagari script, proposed technique of a tree classifier for recognition of Hindi handwritten character by using vertical feature bar, horizontal zeroes, crossing moments. Tree classifiers are used for classification.
Overall accuracy obtained at the character level is 93%.

III. PROPOSED MODEL

We perform the following steps in proposed model:



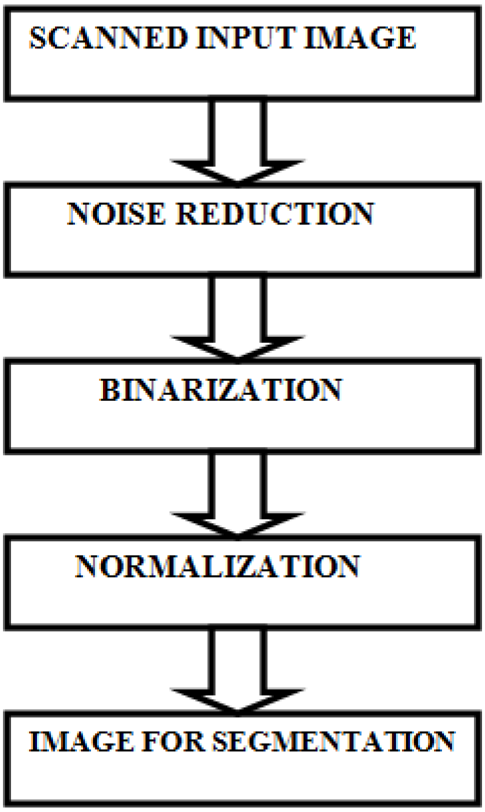
1.Image Acquisition

Handwritten Image is captured from optical scanner and is converted into digital images. The scanner is used is 300 dpi scanner.

2. Pre-processing

Pre-processing aim to produce data that are easy for OCR system.Pre-processing phase is applied to remove unwanted parts from the image by applying one or more method.
Method Involves:-

1) RGB to Gray
2) Threshold
3) Complement the Image
4) Morphological Operations like opening and closing
5) Linearization
6) Noise removal using filters



3. Segmentation

Segmentation of handwritten word is very important task, it is important to improve the accuracy of handwritten word since recognition system is heavily depending upon segmentation phase. Segmentation means to subdivide .technique involves Line, Word and Character segmentation.

3.1 Line segmentation

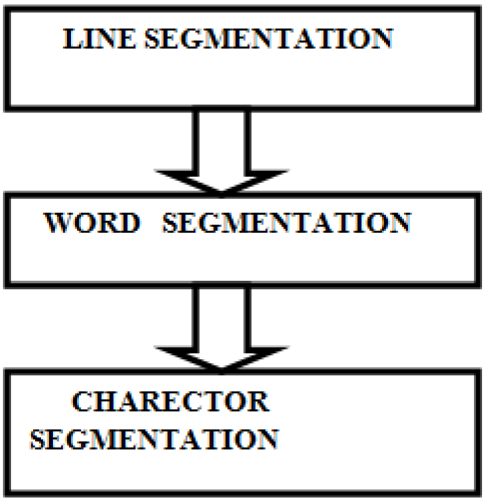
Separation of text lines from text blocks is called line segmentation.

3.2 Word segmentation

Separation of words from each text line is called word segmentation.

3.2 Character segmentation

Separating words into constituent characters is called character segmentation.



4. Feature extraction

Feature extraction is one of the most important steps in developing a classification system. It is represented as a feature vector. Major goal of feature extraction to extract a set of feature which maximizes recognition. Feature extraction and selection can be defined as extracting the most representative information from the raw data, which minimizes the within class pattern variability while enhancing the between class pattern variability. For this purpose, a set of features are extracted for each class that helps distinguish it from other classes, while remaining invariant to characteristic differences within the class. Some features that have been carried for recognition are geometric feature, topological, directional, mathematical & structural features.

5. Classification

The classification is nothing but matching of database characters with the input image characters. For classification purpose various classifier used like Support Vector Machine, K-Nearest Neighbours, Bayesian Classification, and Decision Tree Classification.

IV. CONCLUSION

In this paper, we have presented various step and method like pre processing, segmentation, feature extraction, classification and matching techniques required for optical character recognition of handwritten Devanagari scripts. As In India huge volumes of historical documents and books of handwritten Devanagari script remain to be digitized for better access, sharing, indexing, etc. This project will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics and other recognition system.

REFERENCES

[1] Vijaya Rahul Pawar Arun Gaikwad, Ph.D "*Performance Evalution of Multistage Offline Marathi Script Recognition System*" *International Journal of Computer Applications (0975 – 8887) Volume 88 – No.4, February 2014*

[2] Ratnashil N Khobragade1 Dr. Nitin A. Koli Mahendra S Makesar "*A Survey on Recognition of Devnagari Script*" *International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)*

[3] 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*

[4] 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*

[5] Ashwin S Ramteke, Milind E Rane " *A Survey on Offline Recognition of Handwritten Devanagari Script " International Journal of Scientific & Engineering Research Volume 3, Issue 5, May-2012 1 ISSN 2229-5518*

[6] Kiran R.Dahake S.R,Suralkar S.P.*Ramteke " Optical Character Recognition for Marathi Text Newsprint*" *International Journal of Computer Applications Volume 62– No.16, January 2013*

[7] Veena Bansal and R M K Sinha, "*A Complete OCR for printed Hindi Text in Devanagari Script", IEEE 800 - 804 2001*

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

[9] Asma A. Shaikh, Rahul Dagade," RECO*GNITION OF DEVANAGARI SCRIPT: A SURVEY* " *International Journal of Advanced Technology & Engineering Research (IJATER)*

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