An Optical Character Recognition System for Indian Scripts

Trupti R.Zalke, Prof.V.N.Bhonge

***Abstract*— Character recognition is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. In India, more than 300 million people use Devanagari script for documentation. There has been a significant improvement in the research related to the recognition of printed as well as handwritten Devanagari text in the past few years. As in India 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 In this project, we present work done on OCR of Indian language devanagari script.**

Handwriting is the most effective way by which civilized people speaks. Devanagari is the basic Script widely used all over India. In the proposed work Optical isolated Marathi Characters are taken as an input image from the scanner. An input image is pre processed and segmented. Features are extracted.Feature vector is applied to classifier and then required post processing is done.

***Index Terms*— Classification, Feature Extraction, OCR, Pre processing ,Segmentation,**

I. INTRODUCTION

Machine simulation of human reading has become a topic of serious research since the introduction of digital computers. The main reason for such an effort was not only the challenges in simulating human reading but also the possibility of efficient applications in which the data present on paper documents has to be transferred into machine-readable format. Automatic recognition of printed and handwritten information present on documents like cheques, envelopes, forms, and other manuscripts has a variety of practical and commercial applications. Handwritten 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.

II. LITERATURE SURVEY

Vijaya Rahul Pawar[1] 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.

Trupti R. Zalke, Electronics and communication department,SGBAU/SSGMCE Shegaon, Shegaon Dist: Buldhana, India,
Prof.V.N.Bhonge, Associate Professor Electronics and communication department,SGBAU/SSGMCE Shegaon, Shegaon Dist: Buldhana, India

For classification purpose Self organizing map (SOM) is used.accuracy achieved in this project is 93%.

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.

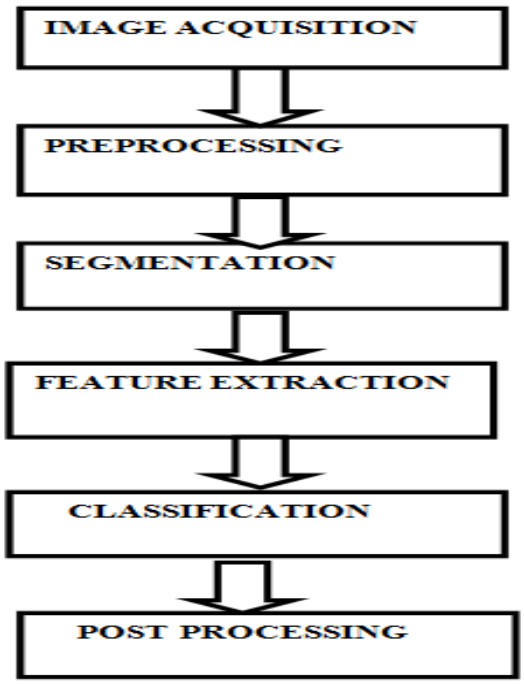
Jayadevan R. et.al [4] did a survey of the comparative study of recognition of printed as well as handwritten word recognition by different classification techniques like Artificial Neural Network, Hidden Markov Model, Support Vector Machine.

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.

Vedgupt Saraf [7] was used genetic algorithm for an excellent means of combining various styles of writing a character and generating a new style.

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.

III. PROPOSED MODEL



1. Image Acquisition:

Handwritten Image is captured from optical scanner and are converted into digital images.

The scanner is used is 300 dpi scanner. The image should have a specific format such as .jpeg, .bmp, .png etc.

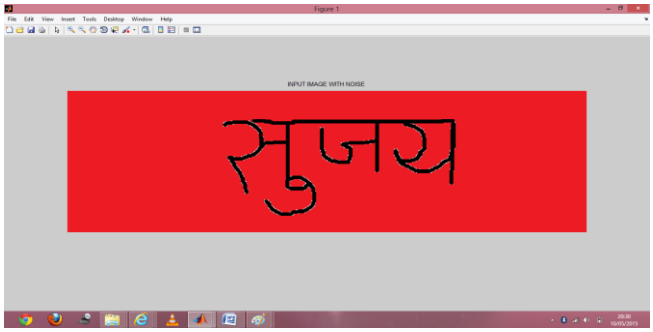


Fig.1 Scanned Input Image

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) Binarization
6) Noise removal using filters

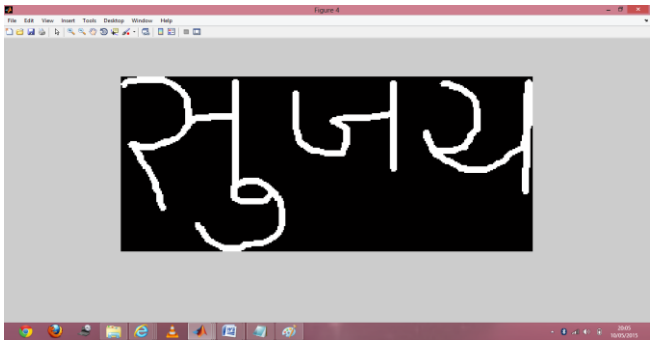


Fig.2 Pre Processing Output

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 depend upon segmentation phase. segmentation means to subdivide .technique involves Line, Word and Character segmentation .

Line segmentation:

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

Word segmentation:

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

Character segmentation:

Separation of character from each word is called word segmentation

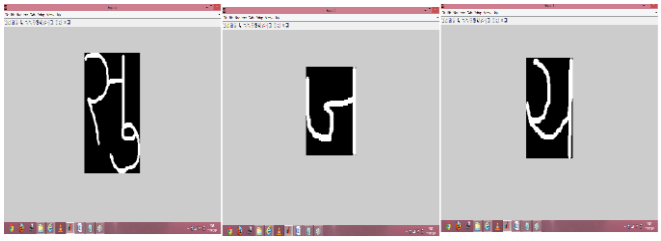


Fig. 3 Segmentation Output

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.

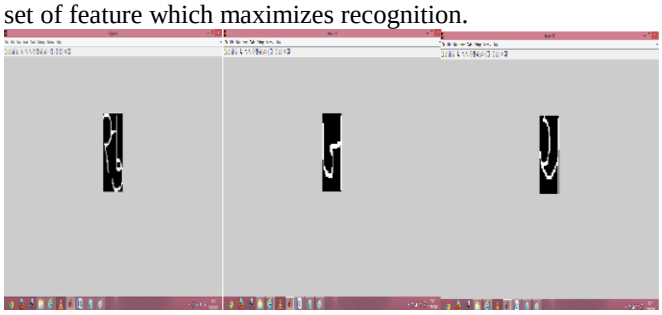


Fig. 4 Cropping and Resizing Output

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.

Here used Template Matching for classification purpose.

Template matching:

This is one of the simplest approaches to pattern recognition.In this approach a prototype of the pattern that is to be recognized is available. The given pattern that is to be recognized is compared with the stored patterns. The size and style of the patterns is ignored while matching.

Flow:

- Image from detected string is selected.
- Size of template is rescaled.
- Matching matrix is computed.
- Find Highest match .
- Index of best match is stored.

6. Post Processing

Post-processing stage is the last stage of the proposed recognition system. Post-processing step involves grouping of symbols.

The process of performing the association of symbols into strings is referred to as group.In post processing other considerations like cost of errors are considered for the final decision.

It prints the corresponding recognized characters in the structured text form. It generates the text file.

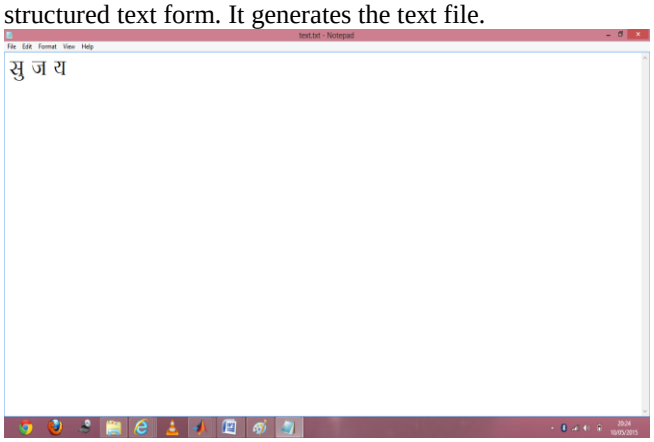


Fig.5 Output of Character Recognition

IV. CONCLUSION

Only a few work is done on script identification. Generally researchers assume that a given document is written in a specific script. In countries like India, where many languages and scripts exist, the identification of script has to be done prior to the recognition in various application.

This presented various step and method like pre processing, segmentation, feature extraction, classification ,post processing and matching techniques required for optical character recognition of handwritten devanagari scripts. As in India huge volumes of historical documents and books (handwritten or printed in 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.

V. FUTURE SCOPE

1) OCR for poor quality documents:

Most of the work reported on Indian languages are on good-quality documents. Elaborate study on poor-quality documents are not undertaken by the scientists in the development of Indian script OCR.

Experiments should be made to observe the effect of poor quality paper as well as noise of various types, and take corrective measures.

2) OCR for the visually handicapped:

One of the primary motivations of early development of OCR system was a reading aid for the visually handicapped.

In India too there is a great need for reading aid for the blind. One possible way of achieving this goal is to convert the OCR output into speech format.

REFERENCES

[1] Vijaya Rahul Pawar, ArunGaikwad, 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] Ved Prakash Agnihotri, ―Offline Handwritten Devanagari Script Recognition‖ in MEC, pp. 37-42, 2012. 3)Bindu Philip and R. D. Sudhaker Samuel. "A Novel Bilingual OCR for Printed Malayalam-English Text based on Gabor Features and Dominant Singular Values" 2009 IEEE4)

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal Offline Recognition of Devanagari Script: A SurveyIEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 6, NOVEMBER 2011

[10] Sandhya Arora and Debotosh Bhattacharjee, ―Multiple classifier combination for Offline Handwritten Devanagari Character Recognition‖.

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