Recognition of Marathi Newsprint Text Using Neural Network and Genetic Algorithm

Miss Vrushali V. Sabale
Dept of computer sci.& Information Tech
Sant Gadage Baba Amravati University
Amravati, India
vrushali19sabale@gmail.com

Prof. Mukund R. Joshi
Fac.of computer sci.& Information Tech.
Sant Gadage Baba Amravati Univercity
Amravati,India
mukundjoshi98@yahoo.co.in

Abstract— Now a day there are many new methodologies required for the increasing needs in newly emerging areas, with this methodologies there are many techniques are present for the character recognition of handprint Devanagri, Bengali, Tamil, China etc. But very little research is for printed material. So in our project we propose the recognition of devnagari printed text using neural network and genetic algorithm. 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.. All feature-extraction techniques as well as training, classification and matching techniques useful for the recognition are discussed in various sections of the paper. An attempt is made to address the most important results reported so far and it is also tried to highlight the beneficial directions of the research till date. Moreover, the paper also contains a comprehensive bibliography of many selected papers appeared in reputed journals and conference proceedings as an aid for the researchers working in the field of Devanagari printed text using neural network and genetic algorithm.

Keywords*- Devanagari,Optical character recognition, Feature extraction, Segmentation.*

*****

I. INTRODUCTION

Artificial Neural Network (ANN) is a very efficient method for recognizing characters. Attempts have already been made to recognize English alphabets using similar type of methods. Expert systems can be developed using Artificial Neural Network (ANN) which can recognize hand written English alphabets easily and accurately. Now a day there are many new methodologies required for the increasing needs in newly emerging areas, with this methodologies there are many techniques are present for the character recognition of handprint Devnagri, Bengali, Tamil, China etc. But very little research is for printed material for devnagari text. So in our project we propose the Recognition Of Devanagari Printed Text using neural network and genetic algorithm.

Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. The study investigates the direction of the CR research, which can be classified based upon two major criteria: the data acquisition process (on-line or off-line) and the text type (machine-printed or hand-written). No matter which class the problem belongs, in general there are five major stages in the CR[1] problem: pre-processing, segmentation, representation, training and recognition and post processing.

Hindi is the national language of India and one of the most popular languages in the world which is the form of Devanagari script. Hence in this paper Devanagari texts are chosen for recognition. 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 thechallenges 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 applicationsin banks, post offices, libraries, and publishing houses. Genetic algorithm: In the field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a meta heuristic) is routinely used to generate useful solutions to optimization and search problems.[1]Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems.

II. LITERATURE REVIEW

The work on automatic recognition of printed Devanagari script started in early 1970s. The efforts then were initiated by Sinha at Indian Institute of Technology, Kanpur. A syntactic pattern analysis system for Devanagari script recognition is presented in Sinha's Ph.D. thesis . Another OCR system development of printed Devanagari is by Palit and Chaudhuri as well as Pal and Chaudhuri . A team comprising Prof. B. B. Chaudhuri, U. Pal, M. Mitra, and U. Garain of Indian

Statistical Institute, Kolkata, developed the first commercial level product for printed Devanagari OCR.

Some of the existing techniques used in OCR for Indian scripts work is presented here.

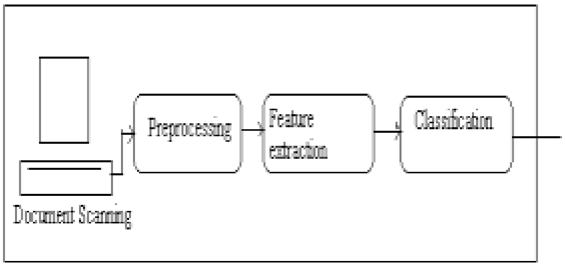
K.h.aparna, sumanth jaganathan [1] reported on "An Optical Character Recognition System For Tamil Newsprint". In this project, they present version of a complete Optical character recognition (OCR) system for tamil newsprint. They implement all the standard elements of OCR process like deskewing, preprocessing, segmentation, character recognition and reconstruction. They used the ability of artificial neural networks to learn arbitrary input/output mappings from sample data for solving the key problems of segmentation and character recognition. After the project they conclude that when the text block having a few touching characters is sent for character recognition, 94% recognition rate is obtained. In general, for other documents the recognition rate varied from 85 to 90 percent depending on the touching characters present in the text part.

Bindu Philip and R. D. Sudhaker Samuel [3] reported on An Efficient OCR for Printed Malayalam Text using Novel Segmentation Algorithm and SVM Classifiers with a recognition rates between 90.22% and 95.31 %.In this method the system segments the scanned document image into text lines, words and further characters and sub-characters. The segmentation algorithm proposed is motivated by the structure of the script. A novel set of features, computationally simple to extract are proposed. The approaches used here are based on the distinctive structural features of machine-printed text lines in these scripts. A lateral cross-sectional analysis is performed along each row of the normalized binary image matrix resulting in distinct features. The final recognition is achieved through classifiers based on the Support Vector Machine (SVM) method.

III. PRAPOSED METHODOLOGY

Praposed Work:

With the advent of computer and information technology, there has been a dramatic increase of research in the field of Devanagari OCR[4] since 1990. Different strategies using combination of multiple features, multiple classifiers, and multiple templates have been considered extensively in the state of the art. Only a few works have been reported in the areas of Devanagari printed recognition. Some research is really required to find ideal combinations of classifiers for the purpose of recognition.



Components of an OCR System

Fig 1. Components of OCR system

The classical paradigm for character recognition has three steps: segmentation, feature extraction, and classification.

A. Preprocessing

The preprocessing stage is a collection of operations that apply successive transformations on an image. It thereby simplifying the processing of the rest of the stages. Different image processing steps like enhancing, cleaning, extracting region of interest comes under same family named as pre-processing. As far as documents containing Devanagari text are concerned, the most important characteristic to be considered for skew estimation .But as we are dealing with the printed character there is no need of preprocessing, as our image quality is already enhanced.

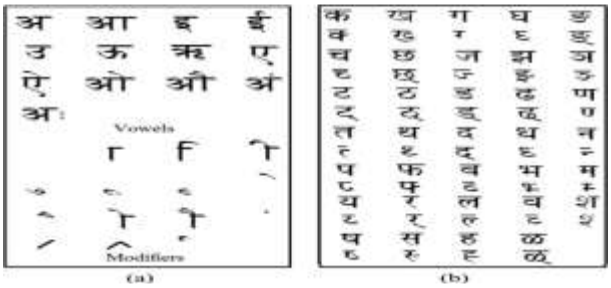
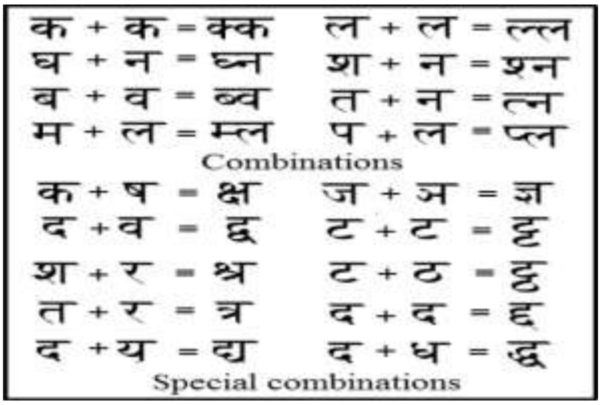

Vowels
Modifiers
(a)
(b)

Fig. 2 (a) Vowels and modifiers of Devanagari script. (b) Consonants and their corresponding half forms (shown below the consonants) in Devanagari script.

B. Segmentation

In segmentation we will try to attempts to segment words into letters or other units with or without use of feature based dissection algorithms. The most simple and straight forward segmentation algorithm is a vertical scan. The algorithm binarizes the image into black and white pixels and simply looks for unbroken columns of white pixels. This work well for machine printed characters or handwritten characters in which a prescribed amount of white space is guaranteed. A more robust technique is to isolate regions of connected black pixels. This method separates the black pixels into sets in which each black pixel is adjacent to another black pixel in the set. This method works very well for digits that are not overlapped, touching or disjoint. In most of the recognition systems, in order to avoid extra complexity and to increase the accuracy of the algorithms, a more compact and characteristic representation is required.



क + क = क्क ल + ल = ल्ल
घ + न = घ्न श + न = श्न
ब + व = ब्व त + न = त्न
म + ल = म्ल प + ल = प्ल
Combinations
क + ष = क्ष ज + ञ = ज्ञ
द + व = द्व ट + ट = ट्ट
श + र = श्र ट + ठ = ट्ठ
त + र = त्र द + द = द्द
द + य = द्य द + ध = द्ध
Special combinations

Fig. 3. Some combinations of consonants with themselves

C. Feature Extraction

As far as the development of an OCR[8] system is concerned, the most important step can be the extraction of precise features from the characters, symbols, and words. There are several ways for feature extraction, but the most important is to extract the features, which can distinct different patterns. Feature extraction is one of the most important steps in developing a classification system. This step describes the various features selected by us for classification of the selected characters.

There are many features are extracted for the recognition of Marathi characters for that we consider features as follows:

a. GLCM (Gray level co-occurrence matrix)
b. Histogram of individual characters
c. Color Dominant

GLCM: Level Concurrence Matrix (GLCM) method is a way of extracting second order statistical texture features. The approach has been used in a number of applications, Gray Level Co-Occurrence Matrix (GLCM) has proved to be a popular statistical method of extracting textural feature from images.

Histogram:

Histogram: The histogram's *x*-axis reflects the range of values in Y. The histogram's *y*-axis shows the number of elements that fall within the groups; therefore, the *y*-axis ranges from 0 to the greatest number of elements deposited in any bin. The *x*-range of the leftmost and rightmost bins extends to include the entire data range in the case when the user-specified range does not cover the data range; this often results in "boxes" at either or both edges of the distribution.

Color Dominant: Use surf and surfc to view mathematical functions over a rectangular region. surf and surfc create colored parametric surfaces specified by X, Y, and Z, with color specified by Z or C.surf(Z) creates a a three-dimensional shaded surface from the *z* components in matrix Z, using x = 1:n and y = 1:m, where [m,n] = size(Z)

D. Classification

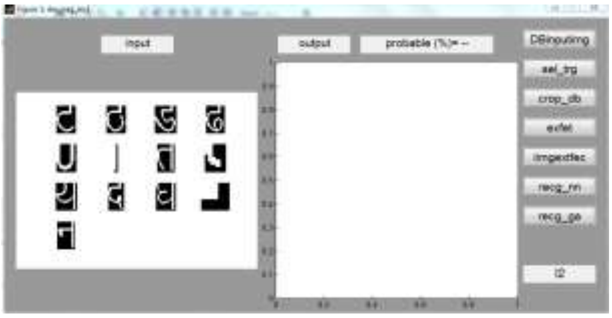
The classification stage is the main decision making stage of an OCR system and uses the features extracted in the previous stage to identify the text segment according to preset rules. The post-processing stage, which is the final stage, improves recognition by refining the decisions taken by the previous stage and recognizes words by using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies,

Neural network:-A neural network[2] is defined as a computing architecture that consists of massively parallel interconnection of adaptive 'neural' processors. Because of its parallel nature, it can perform computations at a higher rate compared to the classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. A neural network contains many nodes. The output from one node is fed to another one in the network and the final decision depends on the complex interaction of all nodes**.**

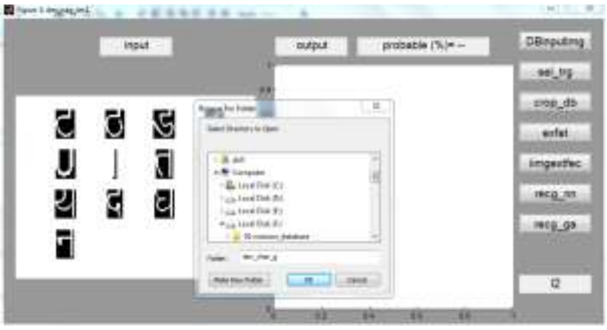
Genetic algorithm: In the field of artificial intelligence, a genetic algorithm (GA)[7] is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a meta heuristic) is routinely used to generate useful solutions to optimization and search problems.[1]Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

IV. .RESULT

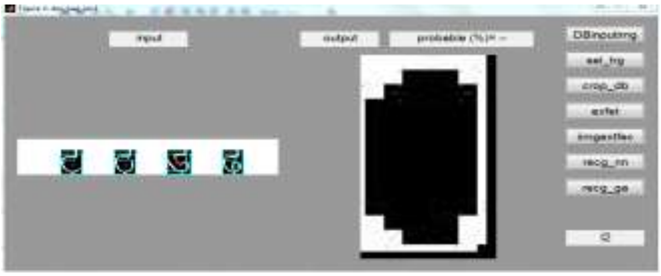
With the advent of computer and information technology, there has been a dramatic increase of research in the field of Devanagari OCR since 1990. Different strategies using combination of multiple features, multiple classifiers, and multiple templates have been considered extensively in the state of the art. There is a great scope of research in these areas for the future researchers in the area of printed Devanagari OCR.

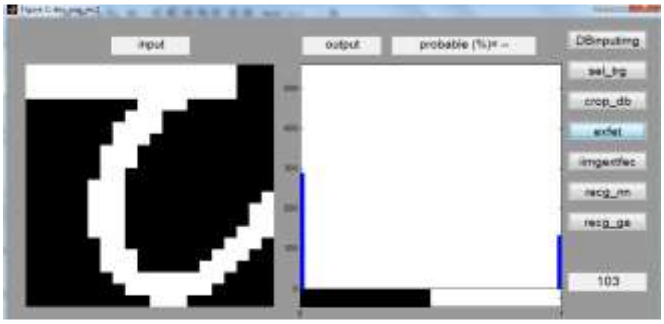


While doing research over Devnagari printed text, it is noted that, you have to extract the scanned images of devnagari text from the database ,the errors in recognizing printed Devanagari characters are mainly due to incorrect character segmentation of touching or broken characters. Because of upper and lower modifiers of Devanagari text, many portions of two consecutive lines may also overlap and proper segmentation of such overlapped portions are needed to get higher accuracy.



After extracting the image you have to select target folder so that you can perform various operation on themlike feature extraction ,classification etc.In India huge volumes of historical documents and books (printed in Devanagari script) remain to be digitized for better access, sharing, indexing, etc. This will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics.





Though neural network architecture is complex but accuracy of ANN can be further increased by increasing the number of samples for training the network India's national language Hindi (written in Devanagari script) is world's third most popular language after Chinese and English.

V . .Future research

Lots of effort has been taken over character recognition by various peoples like B. B. Chaudhuri, U. Pal, M. Mitra, and U. Garain of Indian Statistical Institute, Up till now we used Neural network, OCR technology in future with the addition of that we can do use of genetic algorithm. This is the field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a meta heuristic) is routinely used to generate useful solutions to optimization and search problems.

VI. CONCLUSION

As seen from the result the overall performance of the system showed that though neural network architecture is complex but accuracy of ANN can be further increased by increasing the number of samples for training the network the errors in recognizing printed Devanagari characters are mainly due to incorrect character segmentation of touching or broken characters. In India huge volumes of historical documents and books(printed in Devanagari script) remain to be digitized for better access, sharing, indexing, etc. This will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics.

Authors and Affiliations

Miss Vrushali V. Sabale received the B.E. degree in Computer Science Engineering from P.R.Pote college of engineering and Technology in 2013 and pursuing M.E in H.V.P.M. college of engineering,Amravati India .

Prof. Mukund R. Joshi received the B.E. and M.E. degree in electronics and telecommunication from Prof Ram Meghe Institute of Technology and Research Badnera , He is currently working as Assistant Professor at H.V.P.M's college of Engineering and Technology, Amravati,India.

Acknowledgment

My thanks to the Guide, Prof. M.R.Joshi and Principal Dr.A.B.Marathe, who provided me constructive and positive feedback during the preparation of this paper.

References

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

[2] Raghuraj Singh1:International Journal of Computer Science & CommunicationVol. 1, No. 1, January-June 2010, pp. 91-95 :Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network

[3] B.Indira 1: I.J. Image, Graphics and Signal Processing, 2012, 6, 15-21 Published Online July 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijigsp.2012.06.03 : Classification and Recognition of Printed Hindi Characters Using Artificial Neural Networks International Journal of Advanced Technology & Engineering Research

[4] Sameeksha Barve :(IJATER) ISSN NO: 2250-3536 VOLUME 2, ISSUE 2, MAY 2012 139 OPTICAL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK

[5] Chucai Yi, *Student Member, IEEE*, and Yingli Tian, *Senior Member*, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 7, JULY 2014Scene Text Recognition in Mobile Applications by Character Descriptor and Structure Configuration

[6] **S.L. Mhetre*** Volume 4, Issue 2, February 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering: Recognition of Devanagari Handwritten Numerals using Two Different Approaches

[7] Vedgupt Saraf, D.S. Rao :International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, April 2013 :Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency

[8] Smeet D. Thakur and Prof. Smita S. Sikchi Offline Recognition of Image for content Based Retrieval International Journal of Latest Trends in Engineering and Technology (IJLTET).