ISSN: 2321-7782 (Online)
Volume 3, Issue 2, February 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Recognition of Devanagari Printed Text Using Neural Network and Genetic Algorithm

Prof. Mukund R. Joshi[1]
Faculty of Computer sci & Info-Tech
Sant Gadage Baba Amravati University
Amravati - India

Vrushali V. Sabale[2]
Dept. of Computer sci & Info-Tech
Sant Gadage Baba Amravati Univercity
Amravati - India

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 Devnagri, Bengali, Tamil, China etc. But no more 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: character understanding, character retrieval, segmentation, feature extraction.

I. INTRODUCTION

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 no more research is for printed material. So in our project we propose the An Optical Character Recognition Character Recognition for Marathi Newsprint Scripts. Devanagari is the script used for writing many official languages in India, such as Hindi, Marathi, Sindhi, Nepali, Sanskrit, and Konkani, where Hindi is the national language of the country. Hindi is also the third most popular language in the world

Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. No matter which class the problem belongs, in general there are four major stages in the CR problem: scanning, segmentation, classification and feature extraction. The problem of recognizing handwriting, recorded with a digitizer, as a time sequence of pen coordinates is known as on-line character recognition .But it cannot be applied to documents printed or written on papers. Off-line character recognition is known as Optical Character Recognition (OCR).

II. LITERATURE REVIEW

The work on automatic recognition of printed Devanagari script started in early 1970s. The efforts then were initiated by Sinha [9], [10] at Indian Institute of Technology, Kanpur. A syntactic pattern analysis system for Devanagari script recognition is presented in Sinha's Ph.D. thesis [9]. Another OCR system development of printed Devanagari is by Palit and Chaudhuri [11]

as well as Pal and Chaudhuri [12]. 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.

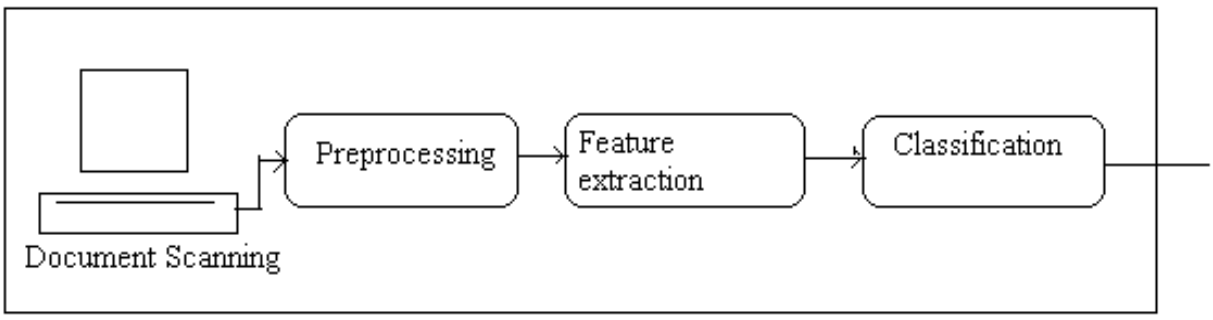
Bindu Philip and R. D. Sudhaker Samuel [2] reported on A Novel Bilingual OCR for Printed Malayalam-English Text based on Gabor Features and Dominant Singular Values. In this paper he developed an OCR that identifies the Malayalam script from English using Gabor filters. A Bilingual Character Recognition System is proposed for the characterization and classification of printed Malayalam-English characters. Indian scripts in general are rich in patterns and variations. Gabor features are extracted after the word level segmentation to identify the script and recognition is based on characterization using Dominant singular values. A recognition rate of 96.5% was achieved for the two–stage classification approach.

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. Proposed Methodology and Discussion

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

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.



Components of an OCR System

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. Therefore for devnagari printed text recognition, we will try to attempts following two main technologies:-

Neural network:- In machine learning, artificial neural networks (ANNs) are a family of statistical learning algorithms inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature.

For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

A neural network 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) 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. Conclusion and Future Research

From the survey, it is noted that 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. Many authors suggest that the postprocessing of classifier outputs by integrating a dictionary with the OCR system can significantly reduce the misclassifications in printed as well as handwritten word recognitions. 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.

Acknowledgement

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. Sanghamitra Mohanty, Himadri Nandini Dasbebartta. An Efficient Bilingual Optical Character Recognition (English-Oriya) System for Printed Documents.
2. 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 IEEE.
3. Bindu Philip and R. D. Sudhaker Samuel. “An Efficient OCR for Printed Malayalam Text using Novel Segmentation Algorithm and SVM Classifiers” International Journal of Recent Trends in Engineering, Issue. 1, Vol. 1, May 2009.

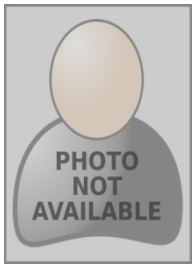
4. Sanghamitra Mohanty, Himadri Nandini Dasbebartta. "An Efficient Bilingual Optical Character Recognition (English-Oriya) System for Printed Documents".

5. Nafiz Arica, Student Member, IEEE and Fatos T. Yarman-Vural. "An Overview Of Character Recognition Focused On Off-line Handwriting" Senior Member, IEEE.

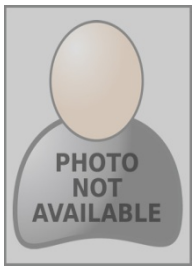
6. R. M. K. Sinha, "A Syntactic pattern analysis system and its application to Devnagari script recognition," Ph.D. Thesis, Dept. Elect. Eng., Indian Institute of Technology, Kanpur, India, 1973.

7. R. M. K. Sinha and H. Mahabala, "Machine recognition of Devnagari script," IEEE Trans. Syst. Man Cybern., vol. 9, no. 8, pp. 435–441, Aug. 1979.

AUTHOR(S) PROFILE



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 Associate Professor at H.V.P.M's college of Engineering and Technology, Amravati.



Miss Vrushali V. Sabale received the B.E. degree in Computer Science Engineering from P.R.Pote college of engineering and Technology in 2013. She is currently persuing Master's Degree in Computer Science and Information Technology from H.V.P.M's College of Engineering And Technology, Amravati.

