Optical Character Recognition of Devanagari Script Using Machine Learning- A Survey

Mr.Sunil Kumar Dasari
Research scholar, Department of Electronics and Communication Engineering
School of Engineering, Presidency University, Bengaluru, Karnataka, India

Dr.Shilpa Mehta
Professor and Head-Department of Electronics and Communication Engineering
School of Engineering, Presidency University, Bengaluru, Karnataka, India

Ms.Diana Steffi D.D
Research Scholar, Department of Electronics and Communication Engineering
School of Engineering, Presidency University, Bengaluru, Karnataka, India

Abstract- Optical character recognition(OCR) became the powerful tool when it comes to most of the digital world applications, there are wide variety of languages and script styles throughout the world when it comes to countries like America,Russia,Europe the script is almost identical or similar and the research has been started few decades ago and they developed efficient algorithms to any printed document or content on the image or the content on the handwritten document to editable text on a digital device. The beauty of India is it has one basic language that is Devanagari lipi (script), many languages like Hindi, Telugu,Malayalam,Tamil,Kannada from different parts of the county consisting of similar kind of characters which are from devanagari character set , due to this similarity of the character styles the existing algorithms which are efficient for foreign languages are identifying the wrong letter from the printed document or hand written document, to overcome this problem in this problems in this paper mainly we are concentrating to do survey of the work that has been done over few decades in Devanagari fonts identification using Neural Networks.

Keywords – Devanagari Character Recognition, Image Processing, Image Acquisition, Segmentation, feature extraction, Identification and Neural Networks.

I. INTRODUCTION

The concept Optical character recognition (OCR) is the process of identifying the correct character from any printed document or handwritten document or from any image, the OCR is a field which comes under Digital Image processing and its applications, the main motto on concentrating on devanagari script is because of the population who speaks and writes as well as the complexity of the language. Like other foreign languages developing an algorithm for Devanagari OCR is not so easy. Devanagari script has total 52 set of characters in that 34 are consonants and the remaining are 18 vowels, the basic nature of devanagari language is it has unique style of pronunciation not like other languages, the characters in this script varies in size and font styles which is the major task for a computer to identify the font, The process involved in character recognition is capturing the image then preprocess the image and convert the image in to binary streams ,from the binary stream we are going to extract the features of the character that can be identified .

There are six major stages in the Character Recognition those are

- Image Acquisition
- Pre-Processing
- Image Segmentation
- Feature Extraction
- Image classification

- Post processing

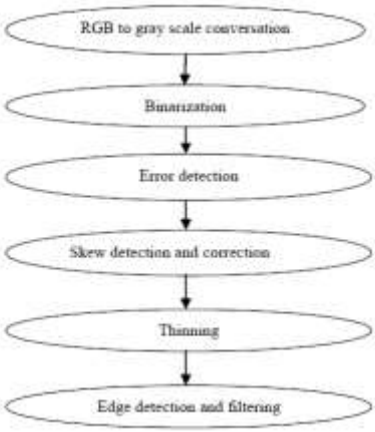


Image acquisition is the process of capturing the image of document by any digital device such as digital camera or document scanner, the captured image may have noise[] that has to be removed in the second stage that is pre-processing, the pre-processed image will be segmented and those symbols will be used in train the Artificial Neural Network to extract the features ,from the features the algorithm able to classify and identify the character.

II. CONVENTIONAL PROCESS OF OCR

The character recognition of any language has the following steps which are shown the block diagram i.e capturing,pre-processing,segmenatation,feature extraction ,classification and post processing.

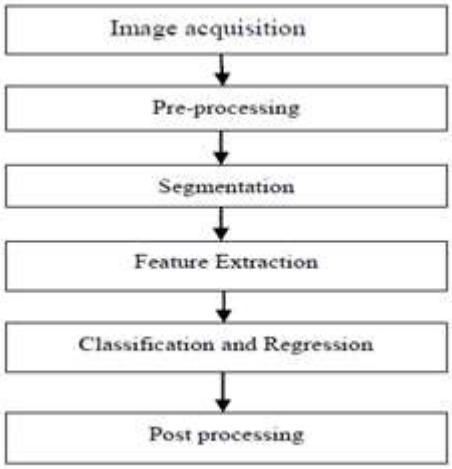


Fig.1: Block Diagram of OCR Steps.

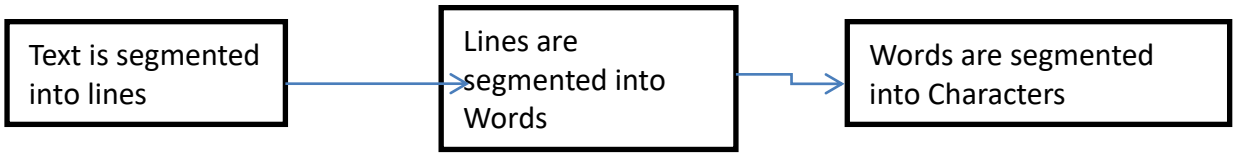
The flow of character recognition is as follows:-

1) Image Acquisition:- Take a picture using digital camera of the document or scan the document and save it in a computer with proper image extention.
2) Pre-processing:- In this process the image undergoes the following operations as shown in the block diagram,the input to the Pre-Processing stage is the stored image in the computer.

Fig.2 Pre-processing Block Diagram

In this stage the stored image should be converted to a gray scale image if it is color image, because our interest is only on the pattern of the character but not the color, and the gray scale image is best suitable to process further in terms of memory requirement and levels of information it consists of. After conversion the signal has to be converted to Binary image which will have only two intensity levels 1 and 0.The Characters in Devanagari lipi has different sizes once it is converted to binary image that has to be brought in to uniform matrix like 16x16 or 128x128 using Image Normalization techniques. Once the image segments brought in to a uniform matrix the redundant bits has to be removed to reduce the unwanted content in the image so that the image smoothness can be improved as well as the chances of increasing in the options to identify correct font.

3) Image Segmentation: - Image segmenation is nothing but dividing the whole image into small sub-images based on the uniqueness in the content,this is the pulse of the whole OCR to improve hit rate of the algorithm.the flow of segmentation in Devanagari lipi OCR is as follows



4) Feature extraction:- Feature extraction is the main part of the Character identification process, this is the process where each character will be represented as a feature vector ,the unique feature of this step,the focus of this stage is to extract a set of features of the segmented image to improve character recognition rate, these features are extracted from global content which was available after segmentation stage therefore the burden on the next stage will be reduced as the input data will be filtered and fed.in thi stage statistical (zoning, projection) ,structural and global transformations(Discrete Cosine Transform) and moment features will be extracted.

5) Image Classification: Once the features are extracted in feature vectors the will be given to image classifiers such as K-nearest Neighbourhood (K-NN),Bayes Clasifier ,neural networks,Hidden Markov Model (HMM) and so on,these clasiffiers are the decision makers of the algorithms.

The best conventional way of classifying the feature vector by different classifiers based on the quality of feature vector derived from the raw input in feature extraction stages are by matching the features of the output with the database which is already stored in memory so that if any the output matches with any vector in database the decision will be taken as template matching, in another way of classification based on the magnitude of the feature vector in multi-dimensional vector space to estimate statistical classifications, on the other hand it is also easy to identify or classify the feature vector by the patterns of the standard scripts, font styles, font size,etc.., gives a chance for decision maker based on syntax of that particular script either by scanning the letters or words or sentences.

6) Post-Processing: In this stage based on the decision from classification stage the recognized fonts will be printed in editable form on digital screen.

III. INTRODUCTION TO DEVANAGARI (LIPI) SCRIPT

Devanagari evolved from the Brahmi script. The word Devanagari has been mystery to scholars, there is a hypothesis that it might be combination of two Sanskrit words 'Deva' (God, king or Brahmans) and 'Nagari'(city).Literally it combines to form 'City of Gods', 'Script of Gods'.

Devanagari, a development of Brahmi system of phonetics, is the only script which has specific signs (grapheme) for the phonetically arranged sounds of the human speech (phonemes), and it is flexible enough to write foreign sounds by attaching marks to the nearer grapheme. The Roman, Greek, Hebrew, and Arabic alphabets have certain traditional names for indicating sound pictures but there is no guarantee that one sign will have only one phonetic value.

The Devanagari script is an important and widely used script of India. It is mainly used to write Hindi, Marathi, Nepali and Sanskrit languages. It also serves as an auxiliary script for other languages such as Punjabi, Sindhi and Kashmiri.

The basic characters of Devanagari script consist of 36 consonants (Vyanjan) and 13 Vowels (Swar). Devanagari script has specific composition rules for joining consonants, vowels and modifiers [6].

Table 1: Vowels and Corresponding Modifiers.

Vowels:	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
Modifiers:		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ

Table 2: Consonants

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट
ठ	ड	ढ	ण	त	थ	द	ध	न	प	फ
ब	भ	म	य	र	ल	व	श	ष	स	ह

Table 3: Half Form of Consonants with Vertical Bar.

क	ख	ग	घ		च		ज	झ	ञ	
			ण	त	थ		ध	न	प	फ
ब	भ	म	र		ल	व	श	ष	स	

Table 4: Examples of Combination of Half-Consonant and Consonant.

क क्क	क ल क्ल	घ न घ्न	च म च्म	श च श्च	त न त्न	प त प्त	प ल प्ल
ब व ब्व	भ न भ्न	म ल म्ल	ल ल ल्ल	श न श्न	श व श्व	श ल श्ल	स न स्न

Table 5: Examples of Special Combination of Half- Consonant and Consonant.

क ष क्ष	ज ञ ज्ञ	ट ट ट्ट	ट ठ ट्ठ	त र त्र	द द द्द
द ध द्ध	द व द्व	द व र द्व	श र श्र	द भ द्भ	द य द्य

Table 6: Special Symbols

क़	ख़	ग़	ज़	फ़	ड़	ढ़	ॅ	॰	:	।	ऽ	॑

IV. ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are one of the main tools used in machine learning. As the "neural" part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize. While neural networks (also called "perceptions") have been around since the 1940s, it is only in the last several

decades where they have become a major part of artificial intelligence. This is due to the arrival of a technique called "back propagation," which allows networks to adjust their hidden layers of neurons in situations where the outcome doesn't match what the creator is hoping for — like a network designed to recognize dogs, which misidentifies a cat, for example. Another important advance has been the arrival of deep learning neural networks, in which different layers of a multilayer network extract different features until it can recognize what it is looking for.

General structure of Neural Network

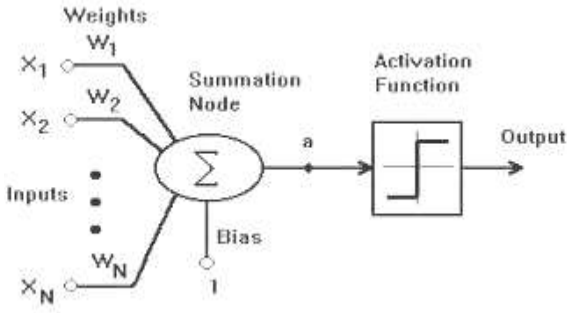


Fig3: Structure of Neural Network

The most common model used in neural network modeling is the multilayer Perceptron (MLP). A multilayer perceptron[11] (MLP) is a deep, artificial neural network. It is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP. MLPs with one hidden layer are capable of approximating any continuous function. The block diagram represents the technical flow of multilayer perceptron [12]

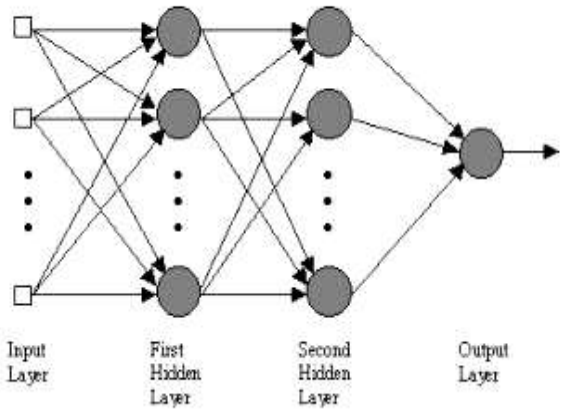


Fig 3:-Block diagram multiplayer Perceptron (MLP).

The Multi-layer perceptron takes the input from the input layer which experience a sequence of multiplication operations based on the predefined weights the quantities gets added and given as input for first hidden layer, here the inputs of first hidden layer undergo the same process based on the weights and processes by some gradient function always nonlinear ,once the processing completed the data will be fed to 2^{nd} hidden layer – where similar process will happen and given at the end to output layer where the final value occurs, therefore the process from input layer to output layer through Hidden layes based on the weights includes additions, multiplications[13].

V. RELATED WORK

In this paper we reviewed the performance of existing algorithms and the efficiency of the algorithms in different environments and the method of approaches to character recognition.

Bansal, Veena & Sinha[1] designed an algorithm to recognize Hindi characters with an efficiency of 93%,they have used robust filters and two distance based classifiers to classify the segmented image in to known cases,and used two level partitioning scheme and search algorithm.

M. Avadesh and N. Goyal[2] proposed an image segmentation algorithm for calculating pixel intensities to identify letters in the image and they trained a 5-layer artificial neural network (ANN) on the data. The ANN contains 2500, 2000,1000 and 800 neurons in the first, second, third and fourth layer respectively. The ANN finally terminates with 602 neuron softmax layer and achived 93.32% accuracy.

Nayak, M. and Nayak, A.K[3] proposed a 4 layer neural network to identify odiya characters of 211 symbols and used binary feature extraction with a modified back propagation neural network to achive better accuracy than the exixting algorithms. S. D. Chame and A. Kumar[4] propsed a thresholding abd Blob analysys for segmentation to detect overlapped regions in the text regions using freeman chain code and SVM for classification and achived an accuracy of 93% for devanagari numerals.

Yadav D, Sánchez-Cuadrado S, Morato J [5] proposed three feature extraction techniques histogram of projection based on mean distance, histogram of projection based on pixel value, and vertical zero crossing, have been used to improve the rate of recognition and with a two level hidden layer back propagation neural network achieved an accuracy of 90% for printed hindi text on document.

Holambe, A. N., et al [6], proposed Euclidian Distance-Based K-NN Classification and used Gaussian filter and a Robert filter are applied to the character image to obtain a gradient image in feature extraction stage on the data handwritten characters from different peoples of different age group (i.e. 03 to 75), i.e. of 7000 people of different age groups and achieved an accuracy of 95% . the author [7], proposed Discrete cosine transform to obtain the data sets for classification and recognition and he also proposed a multilayer perceptron model neural network for classification.

Bhopi, Ms Smita Ashokrao, and Mr Manu Pratap Singh [8],in their work in the field of optical character recognition on devanagari script proposed and implemented four different neural network models using three different feature extraction techniques,the author tested with hybrid evolutionary algorithm with feed forward neural networks to improve the accuracy.

In [9], the author proposed removal of shirorekha in the preprocessing stage itself for the Hindi characters to reduce the noise levels and feature vectors generated by using K-means clustering and for classification the author proposed linear kernel based technique. The author in his work proposed Fuzzy technique [10], for the characters and numerals written in Hindi script as well as in devanagari script and developed an algorithm with an accuracy of 90.6 for Hindi scrip and 92.6 for devanagari script.

Neetu Bhatia [13] in his research work proposed numerous techniques for off line character recognition mainly focused on the complexities in recognizing the characters those are in images with more noise or distortion, he also observed that the conversion from gray color image to binary image and to feature vectors it's a complex process than that of printed text detection in online.

The researcher in [14] suggested that the features of the image extracted using gradient representation then used support vector machines(SVM) and feed forward artificial neural network for classification to take a decision on the output from feature vectors, the author also used the dataset of 8000 samples of Marathi script and normalized at 20x20 matrix size and achieved highest accuracy .

Hinduja, R. Dheebhika and T. P. Jacob[15] in their survey on identification of characters using deep neural network and compared the performance of the OCR algorithm in accuracy and the time to identify the character post processing, the authors experienced very good results when the numbers are trained separately and as well as characters separately before decision making, the authors used feed forward neural network for classification stage and they have worked on the images captured with digital camera for their study.

N. Sankaran and C. V. Jawahar[16] ,In their work the authors proposed a OCR algorithm for Devanagari script to improve accuracy and efficiency by using Recurrent Neural Network i.e Bidirectional LongShort Term Memory(BLSTM) by avoiding few complexities in segmentation stage to minimize the number of wrong identification of the character.

VI. CONCLUSION

In this paper we listed the most popular techniques or algorithms that are already used in optical character recognition field under the branch of Image processing and pattern recognition, after the study we can

understand that as on today there is no single algorithm or OCR model to recognize the characters of devanagari language with 100% accuracy, the existing algorithms with classifiers which are developed based on convolutional neural networks have the highest accuracy compared with other classifiers, and there are plenty of feature extraction methods using standard transformations, still there is lot more scope in extracting the good quality features from the binary image segments, and also its observed that noise filtering at the preprocessing section improves the quality of algorithm. We can put lot more research in to it this field to improve the accuracy of the algorithm for devanagiri script.

VII. REFERENCES

[1]. Bansal, Veena & Sinha, R.. (2001). A Complete OCR for Printed Hindi Text in Devanagari Script.. 800-804. 10.1109/ICDAR.2001.953898.

[2]. M. Avadesh and N. Goyal, "Optical Character Recognition for Sanskrit Using Convolution Neural Networks," 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, 2018, pp. 447-452, doi: 10.1109/DAS.2018.50.

[3]. Nayak, M. and Nayak, A.K. (2017) 'Odia character recognition using backpropagation network with binary features', Int. J. Computational Vision and Robotics, Vol. 7, No. 5, pp.588–604.

[4]. S. D. Chame and A. Kumar, "Overlapped Character Recognition: An Innovative Approach," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, 2016, pp. 464-469, doi: 10.1109/IACC.2016.92.

[5]. Yadav D, Sánchez-Cuadrado S, Morato J. Optical character recognition for Hindi language using a neural-network approach. JIPS. 2013 Mar 1;9(1):117-40.

[6]. Holambe, A. N., et al. "Brief review of research on Devanagari script." *International Journal of Computational Intelligence Techniques* 1.2 (2010): 06-09.

[7]. "OPTICAL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK Extracting structured data from unstructured data using OCR and ANN ", by Anamika Bhaduri1 , Deeksha Gulati2 , Sanvar Inamdar3 , Mayuri Kachare 4. In International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 02, Issue 04; April - 2016 [ISSN: 2455-1457].

[8]. Bhopi, Ms Smita Ashokrao, and Mr Manu Pratap Singh. "Performance analysis of Handwritten Devnagari Character Recognition using Feed Forward, Radial Basis, Elman Back Propagation, and Pattern Recognition Neural Network Model Using Different Feature Extraction Methods." *May* 18: 152-158.

[9]. Gaur A., and Yadav S.: Handwritten Hindi character recognition using K-means clustering and SVM. Fourth International Symposium on Emerging Trends and Technologies in Libraries and Information Services, pp. 65–70, (2015).

[10]. Hanmandalu M., and Murthy O. V. R.: Fuzzy model based recognition for handwritten Hindi numerals. International conference on Recognition, pp. 490-496, (2005).

[11]. Handwritten Character Recognition using Neural Network Chirag I Patel, Ripal Patel, Palak Patel in International Journal of Scientific & Engineering Research Volume 2, Issue 3, March-2011 1 ISSN 2229-5518.

[12]. Application of Neural Networks in Character Recognition , V. Kalaichelvi Assistant Professor Dept of Electronics & Instrumentation Engg BITS PILANI, DUBAI CAMPUS Ahammed Shamir Ali Student BITS PILANI, DUBAI CAMPUS in International Journal of Computer Applications (0975 – 8887) Volume 52– No.12, August 2012.

[13]. " A Review of Research on Devnagari Character Recognition", by Vikas J Dongre Vijay H Mankar ,Department of Electronics & Telecommunication, Government Polytechnic, Nagpur, India. Published in International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010.

[14]. "Evolutionary Computing Techniques in Off-Line Handwritten Character Recognition: A Review ", by Gauri Katiyar Shabana Mehfuz. In UACEE International Journal of Computer Science and its Applications - Volume 1 : Issue 1 [ISSN 2250 - 3765]

[15]. . Hinduja, R. Dheebhika and T. P. Jacob, "Enhanced Character Recognition using Deep Neural Network-A Survey," *2019 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, 2019, pp. 0438-0440, doi: 10.1109/ICCSP.2019.8698008.

[16]. N. Sankaran and C. V. Jawahar, "Recognition of printed Devanagari text using BLSTM Neural Network," *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Tsukuba, 2012, pp. 322-325.