International Engineering Journal For Research & Development

Vol.4
Issue 6

E-ISSN NO:-2349-0721

Impact factor : 6.03

OPTICAL RECOGNITION OF DIGITAL CHARACTERS USING MACHINE LEARNING

Aparna Patil
PRMITR , Badnera ,Dist Amaravati.
aparna_vp@rediffmail.com

S.W. Mohod
PRMITR , Badnera ,Dist maravati.
sharadmohod@rediffmail.com

M. D. Ingole
PRMITR , Badnera ,Dist Amaravati.
manikjadhao@yahoo.co.in

Abstract-

Optical Character Recognition (OCR) plays an important role in document image processing. Recognition of characters in a smart way is gaining importance in the modern days, as huge piles of data is generated, and it needs to be processed and manipulated. In world more than 300 millions peoples uses Devanagari script such as Marathi ,Hindi, Sanskrit has uses Devanagari as base script As compared to English language Devanagari character recognition is complicated. This paper aims at presenting an OCR utility which recognizes text characters, using a machine learning model.

***Keywords*— Artificial intelligence, classification algorithm, machine learning, Optical character recognition, machine learning**

INTRODUCTION

Recognition of optically processed character is known as optical character recognition(OCR). In India 500 million people uses Devanagari script. Many languages like Hindi, Marathi, Guajarati , Sanskrit uses Devanagari as its base script. Devanagari script consist of 34 consonants (vyanjana)and 18 vowels(swar)[2]. Character recognition of Devanagari script is very complicated. All possible pronunciation can not be written perfectly as Devanagari script is partly phonetic. Word written in Devanagari can be pronounced in one way only. OCR is a technique can process variety of documents PDF or digital image into ASCII or other machine recognizable form. Due to lot of variation in fonts, size of written character recognition become difficult. OCR is classified in two types . online OCR in which it process the character as it is written directly avoiding initial stage of identifying the character. In offline OCR handwritten or printed document is processed by scanning into binary or grayscale image to the recognition algorithm.

There are six major stages in character recognition system.

1. Scanning
2. Preprocessing
3. Segmentation
4. Feature extraction
5. Classification
6. Post processing

In OCR input is an image it can be obtained by scanning by digital scanner or loading it from internal storage. Preprocessing involves getting dataset i.e. the image on which we work then import the required library .Noise removal is done in preprocessing stage after segmentation neural network is used to train the dataset which can be used for classification purpose.

BLOCK DIAGRAM FOR OCR

Any OCR system goes through following step i.e. Data acquisition, preprocessing, segmentation, feature extraction ,classification and post processing The block diagram for OCR system as given below.

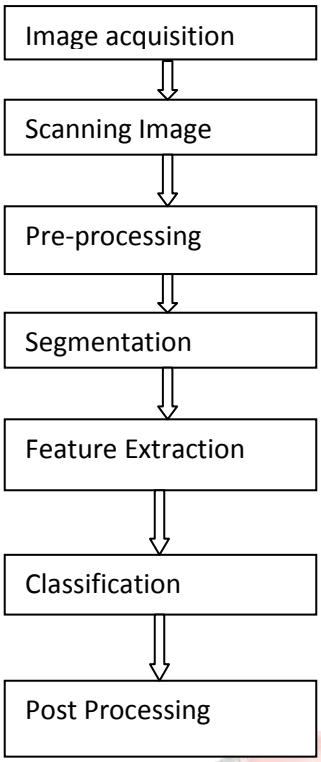

Fig1: Block diagram of OCR

Following are the stages of character recognition system.

a. **Scanning**: In this stage hardcopy of printed document is scanned and converted the image into jpg format.

b. **Preprocessing** : In this stage of OCR system Dataset is divided into two category . Test dataset is on which system works and train dataset is to train machine. Following are the steps of preprocessing.

   Binarization: The scanned image is gray scale image in this step of preprocessing gray scale image is converted into binary image .Two intensity values are obtained as black and white.

   Normalization: For getting uniform size of character normalization is applied.

   Noise Removal: Scanned image contains noise like distortion, gap in line in it The presence of noise in image degrades quality of image which affect on accuracy of recognition of character.

   Thinning :This process involves selected pixel from image.

c. **Segmentation**: In Character Recognition techniques, the Segmentation is the most important process. Segmentation is done to make the separation between the individual characters of an image. The Devanagri words can be separated by removing Shirorekha. Each Separated character generates a sub-image. By vertical projection of document lines are segmented and for word and character segmentation horizontal projection is used.

d. **Feature extraction**: In feature exaction phase the segmented image is tends to scaling or zoning i.e. features of each image examined in detail and recorded by machine for prediction or recognition purpose. Diagonal feature extraction method is efficient feature extraction method.

e. **Classification**: There are various methods of classification to obtain better result number of methods are combined. Template matching, Statistical technique, syntactic approach these are different methods of classification.

f. **Post processing:** Post processing is last stage of proposed OCR system recognized character are printed or displayed in this stage of OCR.

III FEATURES OF MARATHI SCRIPT

Devanagari script has about 14 vowels and 33consonants.Some of the vowels and the consonants are shown in figure l (a) and figure l (d) respectively. In English as well as in Marathi, the vowels are used in two ways:

1. They are used to produce their own sounds. The vowels shown in figure l (a) are used for this purpose in Devanagari.

2. They are used to modify the sound of a consonant. In order to modify the sound of a consonant, we attach an appropriate modifier form symbols shown in figure l (b) in an appropriate manner to the consonant. Each modifier has been attached to the first consonant of the script क (see figure l(c)). **A** visual inspection of figure l(c) reveals that some of the modifier symbols are placed next to the consonant (core modifiers),some above (top modifiers) and some are placed below(lower modifiers) the consonant. Some of the modifiers contain a core modifier and a top modifier, the core modifier is placed before or next to the consonant; the top modifier is placed above the core modifier. Devanagari script has a pure form for most of the consonants.

A consonant in pure form always touches the next character, yielding conjuncts, touching characters, or fused characters. Figure l (e) shows some of the conjuncts formed by writing form consonants followed by consonant ' य. We can use almost any consonantin place of ' य and write over 100 conjuncts.

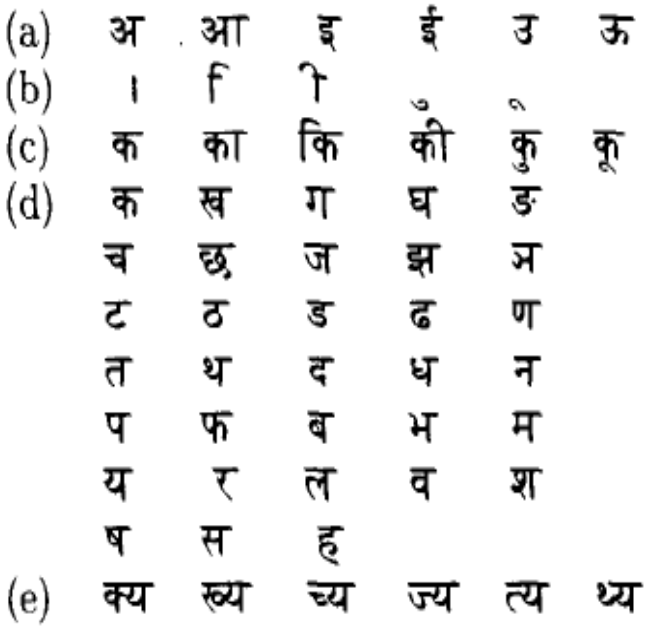

Fig4.3: Characters and Symbols of DevanagariScript; (a) Some of the vowels; (b) Modifier Symbolsfor the above vowels; (c) The modifier symbolsattached to the consonant to indicate क their placing;(d)Consonants; (e) Pure Form of Consonants; (f)Some sample Conjuncts

ARTIFICIAL NEURAL NETWORK

Neural network is also known **as Artificial Neural Network (ANN)**, is an artificial intelligent system which is based on biological neural network. Neural networks able to be trained to perform a particular function by adjusting the values of the connections (weight) between these elements. [8]

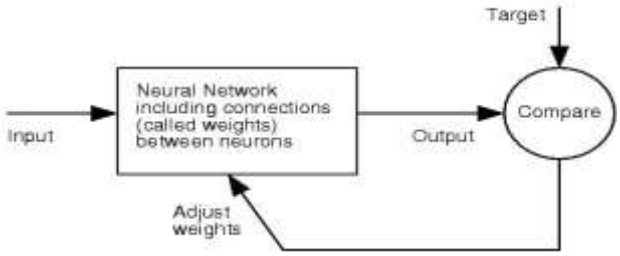

Fig3.1: Neural Network Block Diagram

Neural network is adjusted and trained in order the particular input leads to a specific target output .Example at Figure 3.1, the network is adjusted, based on a comparison of the output and the target until the network output is matched the target .Nowadays, neural network can be trained to solve many difficult problems faced by human being and computer

The feedforward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. In computing, feed-forward normally refers to a multi-layer perceptron network in which the outputs from all neurons go to following but not preceding layers, so there are no feedback loops. Fig 3. below shows a representation of a simple feed-forward Neural Network with four inputs, one hidden layer and four outputs.

Neural networks learn by changing their weights

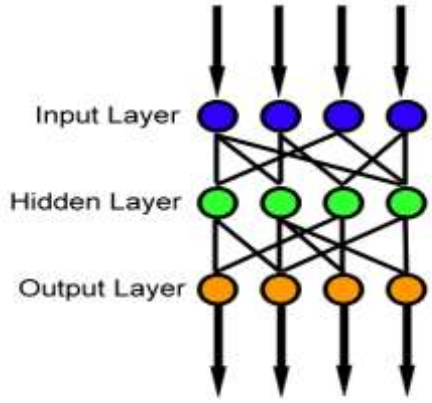

Fig 3.2. A simple feed forward Neural Net

V. RELATED WORK

Proposed OCR System:

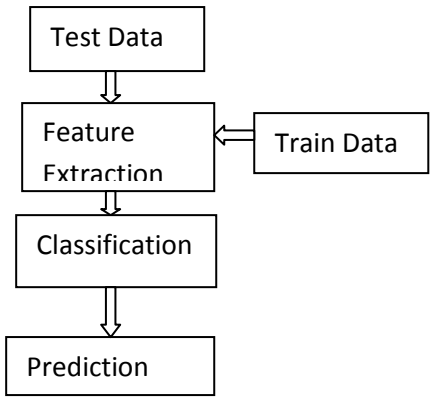

Fig 3.2. Block diagram of proposed OCR

Input to the system is scanned image It is in special format JPEG or BMT This input is taken through scanner or by digital camera ,by any other suitable digital device Preprocessing of scanned input done before segmentation .In preprocessing input scanned image converted into gray scale image and bounding is performed on image before segmentation.

Segmentation: For efficient machine learning model segmentation is necessary In character recognition system input image is sequence of character the segmentation technique divide this sequence of character into image sub image of individual character . These individual image of character is resized individually for further processing.

Feature extraction: In this phase input image subject to features scaling . the image features are examined in detailed and stored it by machine for future prediction. Various variant features are extracted. Here diagonal feature extraction method is used.

It examines the image diagonal and extract the features present over there. It identifies the static and manuscripted character in segmentation resized image is further classified. Thus converted into zone. The feature of each zone extracted In OCR feature extraction method records features of each zone.

Support Vector Machine: Support Vector Machines is a supervised machine learning algorithm which requires labeled data to create and train a model that could align well to the actual function. It classifies two classes bythe means of adjusting a hyper plane between those two classes. We then use the created model to label the test data. Testing the model on test data helps us find the accuracy of the model so created. Suppose we have n-dimensional input vector, the aim of SVM is to find n-1 hyper planes that can identify the training classes. SVM try to find that hyper plane whose margin is the highest between the classes. This is demonstrated in

VI . Result and Conclusion

We tested our Devanagari document reading system on various printed documents and gathered various results. A performance of approximately 98% character level is obtained. A sample text page and the

The preference is given to the mappings that are known OCR confusions. At present ,we store all forms of a word in the dictionary and do not worry about inflection. The methodology described here makes use of the structural properties of the script, namely - presence of top modifiers, lower modifiers and core modifiers that is unique to Indian scripts. We have automated the training process for generating the prototypes. The training process uses the same segmentation process as the OCR.

Table 1 Results obtained from OCR

	Connected Components		Upper Modifiers		Lower Modifiers		Punctuation Marks		Total	
Correct	825	78.34%	123	50.0%	38	77.55%	8	29.6%	994	72.3%
Incorrect	228	21.66%	123	50.0%	11	22.45%	19	70.3%	381	27.7%
Total	1053		246		49		27		1375	

VI. References

1. S. Antani and L. Agnihotri, *Gujrati Character Recognition ,in* Proceedings of the international conference on Document Analysis and Recognition (ICDAR-99), Bangalore, India, pp. 418-421, 1999.
2. Veena Bansal and R.M.K. Sinha, *On how to describe shapes* of *Devanagari characters and use them for recognition*, in Proceedings - Fifth International Conference on Document Analysis and Recognition IEEE Publication, held at Bangalore from Sep 21-23, 1999, pp. 653-656.
3. Veena Bansal and R.M.K. Sinha, *Partitioning and Searching Dictionary for Correction* of *Optically-Read Devanagari Character Strings*, in Proceedings - Fifth International Conference on Document Analysis and Recognition, IEEE Publication, held at Bangalore from Sep21-23, 1999,pp. 410-413.
4. Veena Bansal and R.M.K. Sinha, *Segmentation of Touching Characters in Devanagari*in Computer Vision Graphics and Image Processing, Recent Advances, Eds. S. Chaudhury and S.K. NayarViva Books Private Limited, 1999, pp. 371-401.
5. R.G. Casey and E. Lecolinet, "A survey of Methods andStrategies in Character Segmentation", IEEE Transactionson Pattern Analysis and Machine Intelligence,

6. *G.* C. Cash and M. Hatamian, "Optical character recognition by the method of moments", ComputerVision, Graphics and Image F'rocessing, vol. 39, pp.291-310, 1987.
7. B. B. Chaudhuri and U. Pal, "A Complete PrintedBangla OCR System", Pattern Recognition, vol. 31,no. 5, pp. 531-549, 1997.
8. G. S. Lehal and Chandan Singh, "A Gurmukhiscript recognition system", Proceedings 15th International Conference on Pattern Recognition, Barcelona,
Spain, Vol 2, pp 557-560, 2000.
9. *S.* S. Marwah, S. K. Mullick and R. M. K. Sinha, "Recognition of Devanagari characters using a hierarchical binary decision tree classifier", IEEE International Conference on Systems, Man and Cybernetics,October 1994.
10. R.M.K.Sinha, "Rule based contexual postprocessing for Devanagari text recognition", Pattern Recognition, 20(5), pp. 475-485, 1987.
11. I. K. Sethi, "Machine recognition of constrained hand printed Devanagari" , Pattern Recognition, vol.18, pp. 690-706, 1996. 9, pp. 69-75, 1977.
12. I. S. Oh, J. S. Lee, C. Y. Suen, "Analysis of class separation and Combination of Class-Dependent Features for Handwriting Recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.21, no.10, pp.1089-1094, 1999.
13. S.V. Rajashekararadhya, P. Vanaja Ranjan, "A Novel Zone Based Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Indian Scripts", Digital Technology,
Journal, Vol. 2, pp. 41-51, 2009.
14. D. Trier, A. K. Jain, T. Text, "Feature Extraction Method for Character Recognition - A Survey", Pattern recognition,vol.29, no.4, pp.641-662, 1996.
15. G. G. Rajput, S. M. Mali," Fourier Descriptor based Isolated Marathi Handwritten Numeral Recognition", Int'l Journal ofComputer Applications, Volume 3 – No.4, June 2010.
16. R. M. K. Sinha, "A journey from Indian scripts processing to Indian language processing,"
IEEE Ann. Hist. Comput., vol. 31, no. 1, pp. 8–31, Jan./Mar. 2009.
17.S. Acharya, A. K. Pant and P. K. Gyawali, "Deep learning based large scale handwritten
Devanagari character recognition," *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Kathmandu, 2015, pp. 1-6.

doi: 10.1109/SKIMA.2015.7400041

18. D. Berchmans and S. S. Kumar, "Optical character recognition: An overview and an
insight," *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kanyakumari, 2014, pp. 1361-1365.

doi: 10.1109/ICCICCT.2014.6993174