

BANGLA CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK

Anik Saha¹, Kaushik Deb¹, Mohammad Ibrahim Khan¹, and Kang-Hyun Jo²

¹ Faculty of CSE, Chittagong University of Science and Engineering (CUET), Chittagong, Bangladesh

² Faculty of EE and Information systems, University of Ulsan, Ulsan, South Korea

Corresponding Author: Dr. Kaushik Deb, E-mail: debkaushik99@cuet.ac.bd

Abstract- In this paper, we proposed a method for recognizing printed Bangla Character that uses artificial neural network. The proposed Bangla Character recognition method consists of three basic steps. Initially, convert the input character image to binary image and then normalized the binary character image. Secondly, the normalized binary images send into the proposed neural network which consists of 144 input neurons and 100 output neurons. No hidden layer is used here. In that stage random weight between input layer and output layer neuron is taken. Then we apply our proposed training algorithm to train the network. Finally, randomly choose a character image and send it to trained neural network. By using proposed testing algorithm the corresponding character is detected and represented it into editable text format. Conversion of a character in such away saves a lot of time because if we want to manually convert a script into editable text it takes a long time but if we scan the document and by applying the proposed algorithm to recognize the character it required less times. Unsupervised learning is used here. In case of unsupervised learning output neuron is not fixed rather it depends on how many character it has to be recognized. Vowels, Consonants, Compound characters, Digits can be recognized by this technique. This paper recognizes font size (26-78) of 99% accuracy of the font type that we have trained.

Keywords: Bengali Character, Neural network, Character Recognition, Unsupervised learning

1. INTRODUCTION

Character is the basic building block of any language that is used to build different structure of a language. Characters are the alphabets and the structures are the words, strings and sentences etc. in [1]. Optical character recognition (OCR) is the process of converting scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used to convert books and documents into electronic files, to computerize a record-keeping system in an office, or to publish the text on a website. Normally the text in images cannot be edited. During OCR, the software analyzes the image and converts the picture of characters to editable text based on the pattern of the pixels-image. After OCR it can be used in wide variety of word processing, page layout and spreadsheet applications. All major Indian scripts including Bangla are mixtures of syllabic and alphabetic scripts. Bangla script runs from left to right and it has no equivalent to capital letters of Latin scripts in [2]. There are 11 vowels and 39 consonant in bangle alphabet. The consonant are grouped together and behave like a single character which is known as compound character. There are also some modified vowels which are used to make a string or word. A neural network is a mathematical or computational model which is similar like biological neural network where a set of input is trained and then the trained network is used

to recognized object. The network can be trained in two ways: supervised learning and unsupervised learning. In supervised learning output neuron is fixed but in case unsupervised learning output neuron varies.

2. RELATED WORKS

This section provides a descriptive summary of some methods that have been implemented and tested for character recognition. As far as detection of the character recognition is concerned, researchers have found many methods of recognizing character. An author in [3] implements a method to recognize multi-font Bangla character using digital Curvelet Transform. Moreover, authors don't explain anything about the recognition of compound character. In [4] the character shapes are recognized by a combination of template and feature matching approach. A structural-feature-based tree classifier has been used in [5]. To recognize the basic characters if in case of compound characters, feature based tree classifier is initially used to separate them into small groups and next, an efficient template matching approach is employed to recognize individual character. Supervised learning has been used as Multi-layered feed-forward back propagation neural network in [7] & Conventional back propagation algorithm in [6] to recognize bangla character.

3. METHODOLOGY

For recognizing bangla character first convert the scanned image into binary image and normalize the binary image so that we can get 12×12 matrixes. Then sent it to the input layer of the neural network. Unsupervised learning is used here. In case of unsupervised learning output neuron is not fixed, rather it varies. The number of character that has to be recognized is number of output layer neuron. No hidden layer is used here. In the testing process, select an image of a character, normalize the character into 12×12 matrixes, sent it to the trained neural network, determine which bangla character it is and then print the corresponding Unicode of the character. Two layer neural network used here. No hidden layer is used in this paper. Input layer consists of 144 neurons. This paper is used to recognize 11 vowels, 39 consonants, 10 digits and 40 compound characters.

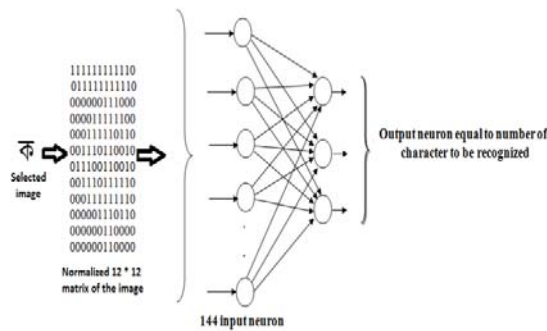


Fig 1: The Proposed neural network

So output layer consists of 100 neurons. Each input neuron is connected to every output layer neuron and a random weight is given to each input to output layer connection.

The process of image analysis to detect character symbol by examining pixel is the core part of input set preparation in both the training and testing phase. This paper assumes that there is nothing else character in the image and font color of image is black. So we put a 1 whether we get a black pixel and put 0 whether we get any other color pixel. As the character can be placed anywhere of the image so our first job is to find the boundary of the character. The following algorithm is used to do so:

1) Algorithm for detecting boundary of character

1. Calculate the sum of each row, the first row which sum is greater than zero is selected as top position.
2. After selecting top position, we will continue the summing of row but then find which row has the sum equal to zero. The row whose value is equal to zero will be selected as bottom position.

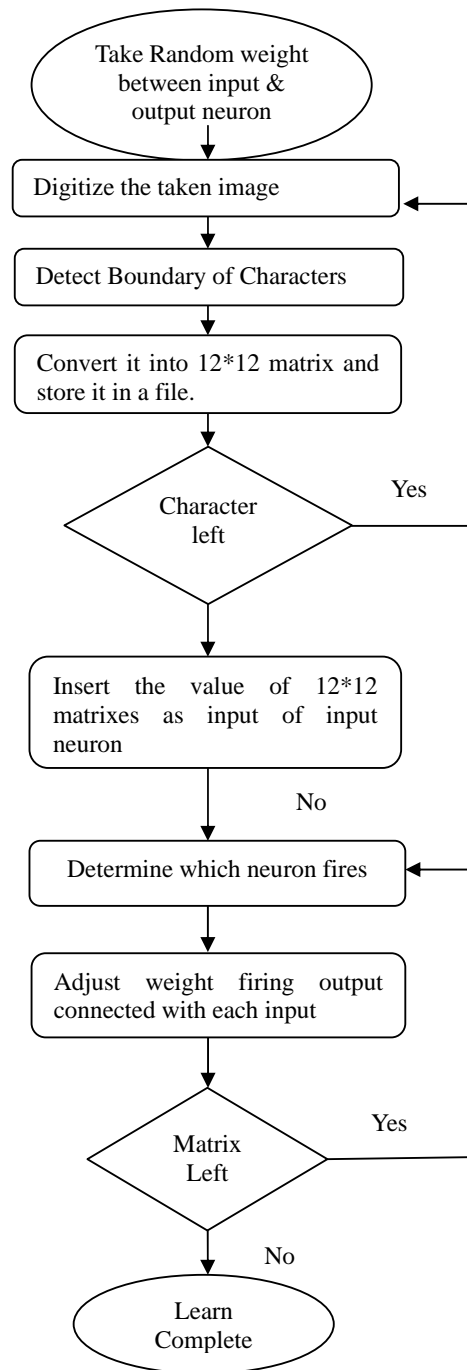


Fig. 2: The proposed character training scheme.

3. If no row value is equal to zero then last row is the bottom position.

4. Continue the process column wise and find left position and right position. Now take the index value from top-left to bottom right position, which is the boundary value.

4. Result and discussion:

We have trained the network using unsupervised learning which is quite different than supervised learning, in supervised learning each time we train the network the combination of output neuron is fixed, but in case of unsupervised learning, we do not get fixed output neuron for each training. For example if we train the network then for a character if we get at testing phase output neuron 7 fires, Then if again we train the same network then may be the firing neuron differs, in that case we may get output neuron 10 fires for the same character. For that we keep a record of firing neuron. Suppose for a testing process output neuron 12 fires for 1st character, output neuron 7 fires for 2nd character then, after selecting a character image and sending it into neural network, if we get output neuron 12 fires then we can assume that we have got 1st character, if we get neuron 7 for the character image then we mark it as 2nd character. Using

this technique it is being able to recognize font size from 26-78 with 99% accuracy. The paper is done in matlab environment. It requires 0.218482 seconds to convert a character into 12*12 matrixes. For training process 0.083225 seconds required and 0.059489 seconds required for testing process.

5. Conclusion

We have presented here a technique that recognizes Bangla characters which include vowels, consonants, digits and compound characters. Our proposed character recognition algorithm operates on input image and efficiently recognizes the character and converts it into editable text.

6. References

- [1] Y. LeCun, B. Boer, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard and L.D. Jackel "Handwritten zip code recognition with multilayer networks" *International Conference on Pattern recognition*, 1990, pp. 35-44
- [2] S. K. Parui, K. Guin, U. Bhattacharya and B. B. Chaudhuri " *Online Handwritten Bangla Character Recognition Using HMM*"
- [3] Angshul Majumdar " *Bangla Basic Character Recognition Using Digital Curvelet Transform*" *Journal of Pattern Recognition Research*, 2007 pp.17-26
- [4] U. Pal, B. B. Chaudhuri, "OCR in Bangla: an Indo-Bangladeshi Language", *Pattern Recognition*, Vol. 2 - Conference B: Computer Vision & Image Processing, Proceedings of the 12th IAPR International, 1994.
- [5] B. B. Chaudhuri, U. Pal, "An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari(Hindi)", IEEE Computer Society, 1997.
- [6] Jalal Uddin Mahmud, Mohammed Feroz Raihan and Chowdhury Mofizur Rahman, "A Complete OCR System for Continuous Bangla Characters", IEEE TENCON-2003: Proceedings of the Conference on Convergent Technologies for the Asia Pacific, 2003.
- [7] Ahmed Asif Chowdhury, Ejaj Ahmed, Shameem Ahmed, Shohrab Hossain and Chowdhury Mofizur Rahman "Optical Character Recognition of Bangla Characters using neural network: A better approach". 2nd International Conference on Electrical Engineering (ICEE 2002), Khulna, Bangladesh.