Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue for ICPCDECT 2016,Volume 3 Issue 1

OPTICAL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK

Extracting structured data from unstructured data using OCR and ANN

Anamika Bhaduri[1], Deeksha Gulati[2], Sanvar Inamdar[3] and Mayuri Kachare [4]

[1-4]Computer Science, AISSM's Institute of Information Technology,

Abstract:- *The increasing use of computers for documentations have lead to a large amount of data in the form of various unstructured documents which are not arranged in a uniform, understandable and integrated way. The processing required for extracting information is still only in its preliminary stage and the hardly predictable document structure make it very hard to extract information automatically. This project focuses on extracting structured data from unstructured data using OCR(Optical Character Recognition) and Neural Network. OCR is a technology which is required to deal with common facts as well as complex designed fonts .It focuses on recognizing characters of a document, that is it does script identification from a variety of unstructured printed or handwritten documents. Neural Network uses trained data, that is, the system will already be trained for recognizing characters of the input unstructured document using synaptic weights. The techniques of feedforwarding and backpropagation will be used in Neural networks which will match the patterns and add new patterns on recognition. The Multilayer Perceptron(MLP) which will match the input to the output using previously stored data will be the model for neural network. The system will be implemented and simulated using Java with Neural Network as the backend for the optical character recognition process. Such an OCR system with Neural Network at it's back focuses towards increases accuracy by eliminating human errors that would occur if the work had be done manually. It also focuses on extracting data irrespective of the noise and the image processing defects.*

Keywords-Optical Character Recognition (OCR), Artificial Neural Network, BPN, Image Processing,Multilayer Perceptron, Synaptic weights.

I.INTRODUCTION

Today's scenario is the amount of unstructured data increasing at an exponential rate. It is difficult to access, extract and edit that large amount of data for human manually. Because of human's workload error get increase while extraction. Also there is a need to store the information in a proper, understandable and usable format. This data has to be stored so that it can be used for further references and information retrieval. This paper focuses on extracting useful and meaning information from an image using Artificial Neural Network and storing it in an editable, easily accessible format. OCR is works on retrieve and extract the data from image and ANN is used for training and comparing the samples of English language.

II.PROPOSED SYSTEM

The proposed system works in the following way:

1. The neural network is trained using synaptic weights, the digitized characters are stored in a .dat file and the pixel connectivity is stored in a xml file.

2, The input image goes through the pre processing steps of resizing ,digitization, segmentation.

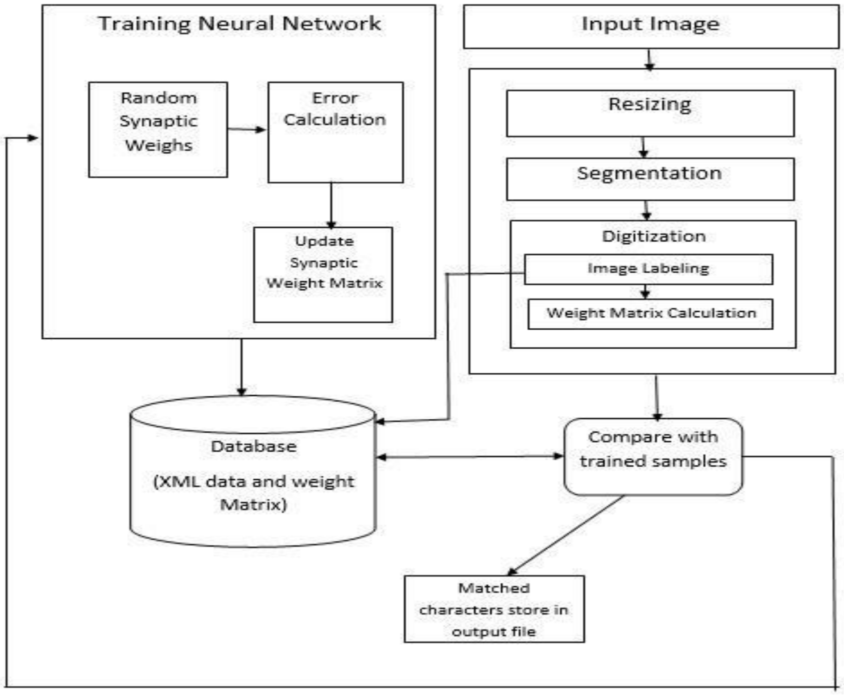
3. The data of the input image is compared with the already trained data, and if the value of Φ is greater than the set threshold, then it indicates a match. and the neural network weights are updated using back propagation and the output , that is, the structured data is then stored in a file.

The proposed system consists of the following modules:

1 Image Processing

2 Neural Network Training.

3 Recognizing and storing in document.


Training Neural Network
Random Synaptic Weighs
Error Calculation
Update Synaptic Weight Matrix
Input Image
Resizing
Segmentation
Digitization
Image Labeling
Weight Matrix Calculation
Database
(XML data and weight Matrix)
Compare with trained samples
Matched characters store in output file

2.1 Image Processing

2.1.1 Image Scanning

In image scanning process, an scanned image will be browsed and will be used for recognition. The image will be in a .jpeg format.

2.1.2 Preprocessing

Image preprocessing comprises of converting the image into a binary format, segmenting it and the digitizing it. The image will also be resized to a standard format. This also involves resizing of the alphabets in a 15*15 pixel size by cropping the top, bottom, left and right boundaries.

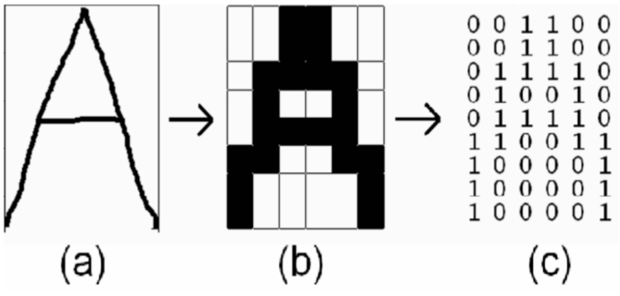
2.1.3 Segmentation and Location

Segmentation is the process that determines the constituents of an image. It is necessary to locate the regions of the document where the text area can be distinguished from the logos and symbols. Only the text area will be used for recognition, whereas the symbols, diagrams and logos will be ignored.

Also, the image which is present in the form of sequence of characters is segmented, that is, separated into parts, which are meaningful and easier to analyze. This is done with the help of pixel and region classification. Labeling process, which will assign a number to each character, will also be used for segmentation of pre-processed input image into isolated characters. This labeling provides information about number of characters in the image.

2.1.4 Digitization

The process of digitization is an important step in the recognition process using neural network. When a document is used for recognition it is expected to have set of printed or handwritten characters pertaining to one or more scripts or fonts.



```
0 0 1 1 0 0
0 0 1 1 0 0
0 1 1 1 1 0
0 1 0 0 1 0
0 1 1 1 1 0
1 1 0 0 1 1
1 0 0 0 0 1
1 0 0 0 0 1
1 0 0 0 0 1
```

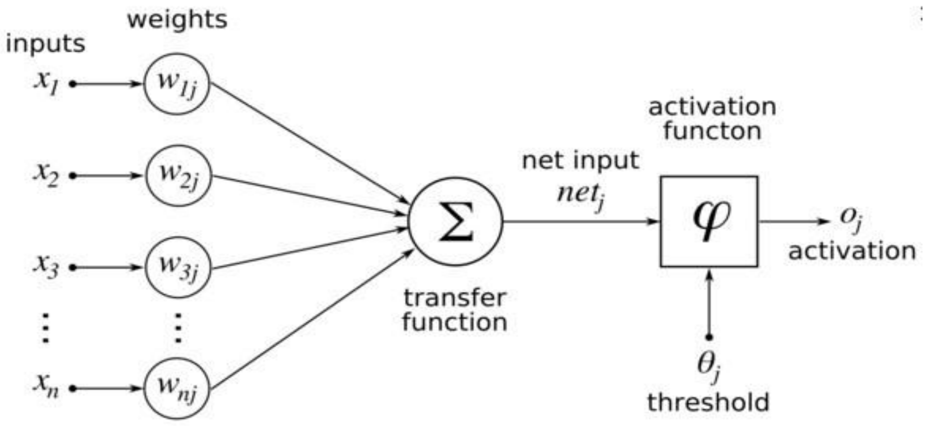
In this process an image matrix is created for each recognized alphabet where 0 is assigned to a white pixel and 1 is assigned to every black pixel. This image matrix which is stored in a .dat file is then used an input for the recognition system In the above figure the digitization of alphabet A has into 6X8 = 48 digital cells has been depicted. Digitization of an image or character into a binary matrix of predefined dimensions makes the input image invariant of its actual dimensions.

2.2 Neural Network Training

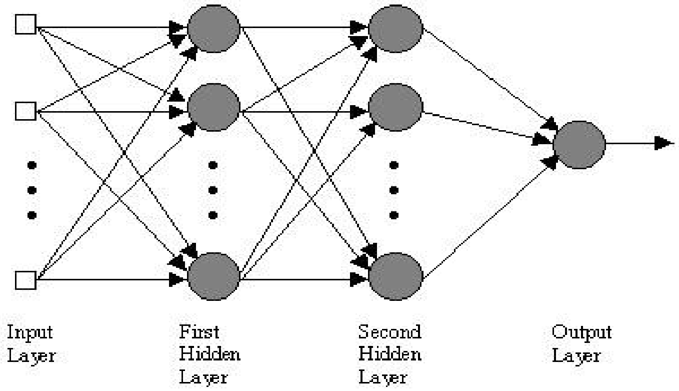
A neural network is a powerful tool which is designed to model the way in which human brain performs a particular task. It has the ability to capture and represent complex input/output relationships. It is a system which performs intelligent tasks similar to those performed by the human brain. In the proposed system

1.A neural network acquires knowledge through supervised learning.

2.A neural network's knowledge is stored within inter-neuron connection strengths known as synaptic weights.



The Multilayer Perceptron Model will be used for simulation of neural network. This model is based on supervised learning as it obtains an output based on the trained set and also learns from the new output obtained. This type of network focuses on creating a model that correctly maps the input to the output using previously stored or trained data so that the model can then be used to produce the output when the desired output is unknown. A graphical representation of an MLP is shown below.



The MLP is a feed forward neural network which uses back propagation for training the neural network. After each data is processed, if a match is obtained, then the perceptron learns by calculating the amount of error in the output compared to the expected result. In this way each time a similar pattern is found the MLP learns using the back propagation technique.

2.2.1 BPN Algorithm steps

1. Apply random synaptic weights as input to the network and find out the output.
2. For each input image, which satisfies the threshold value, work out the error for neuron.
3. The error is What you want – What you actually get
4. Update the synaptic weights.
5. Calculate the Error for hidden layer in neurons.
6. Unlike the output layer we can't calculate errors directly So BACK -PROPOGATATE them in from output layer. Hence algorithm is called the Backpropagation Algorithm.

2.3 Recognizing and storing in document.

2.3.1 Recognition by Adjusting the Weight Matrix

Each character has a label assigned to it and every time a new variant or pattern of the character is found, it is taught to the system under the same label name. Thus the network has different variations of the same pattern under the same label name.

During the training process, the input matrix G is defined as follows:

If I(i, j)=1 then G(i , j)=1

Else

If I(i , j)=0 then G(i , j)=-1

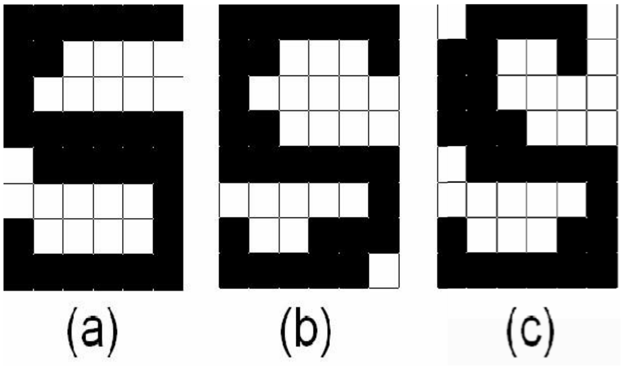
For the k^{th} character to be taught to the network the weight matrix is denoted by Wk .Weight matrix Wk is updated using the following algorithm:

Algorithm 1:

Weight Matrix Wk updation

```
for all i=1 to m;
 {
          for all j=1 to n;
          {
                    W(i , j)= W(i , j) - G(i,j)
          }
}
```

Where m and n are the dimension of matrix Wk.



Printed character vary because then patterns slightly differ from each other.

Then the recognition of patterns is done based on certain statistics.

ϕ=s'/w'

w=Ideal weight

s'= Candidate score

ϕ=Recognition quotient

2.3.2 Image Labeling

Image labeling is a two pass algorithm. It is iterated through 2-Dimensional binary data. Then it is identified by using 8-connectivity and 4-connectivity label of pixels. 8-connectivity uses North-East, North-West and West of current pixel. 4-connectivity uses North and West of current pixel.

2.3.3 Algorithm for First Pass:

Step 1: Iterate each pixel of data by column and row

Step 2: Get 8 neighboring pixels of current pixel.

Step 3: Match the color of 8 neighboring pixels with current pixel.

Step 4: Matching color is added to connected list .

Step 5: If not matching uniquely label the pixel and continue.

Step 6: Otherwise find the connected neighbor with the smallest label and assign to current element.

Step 7: Store the equivalence between the neighboring labels which are not equal.

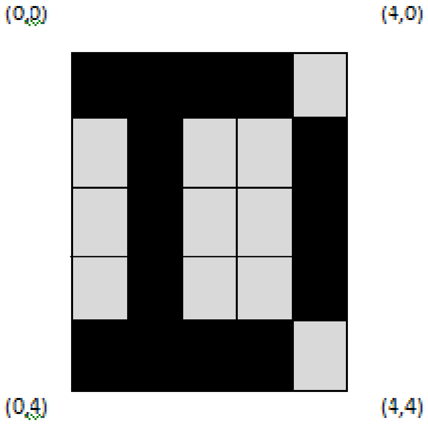
2.3.4 Algorithm for Second Pass:

Step 1: Iterate through each element of data by column and row.

Step 2: Re-label element with lowest equivalent label

Step 3.Finding boundary and generating x,y co-ordinate pixel array:

Left co-ordinate is given by starting x co-ordinate and Right co-ordinate is given by ending x co-ordinate value.The top index is given by lowest y co-ordinate and the bottom index is given by highest y co-ordinate.Width is given by right-left co-ordinate. Height is given by bottom up o-ordinate.



The connected array for the above character is given as

A = { (0,0) (1,0) (2,0) (3,0) (1,1) (4,1) (1,2) (4,2) (1,3) (4,3) (0,4) (1,4) (2,4) (3,4) }

The character is recognized using the digitized input pattern and the above connected array.

Step 4:Matching connected pixels with learned set [.xml]. The XML data is matched with the connected component bit array. According to the x,y co-ordinates, each pixel is matched.The fully matched pixel co-ordinates is the matched character from XML.

<u>Xml Code:</u>

```
<characterinfo>
<ParamValue>a<
<PixelInfo>

                         (0,0) (1,0) (2,0) (3,0)

                         (1,1) (4,1) (1,2) (4,2)
```

(1,3) (4,3) (0,4) (1,4)

(2,4) (3,4)

</ParamValue>
</characterinfo>

2.3.5 Forming words

The left most index of the character in the bit map is represented by LeftXindex. The right most x c-ordinate of the character is represented by RightXindex. If the difference between the current character and previous character is greater than or equal to 3 pixels then a new word is formed.

2.3.6 Storing

Recognized character will be store in file in the editable and accessible format.

III CONCLUSION

The ever increasing data needs to be present in an editable, usable, machine understandable format. Optical Character Recognition aims at reducing human errors by automating the processes of extracting useful information from various unstructured documents. The input image is processed using segmentation, digitization and image labeling. Also, synaptic weights are used to store training data for the neural network. Artificial neural network is commonly used for training the system. Artificial Neural Network for OCR uses Multilayer Perceptron model to compare the input image with the trained set to obtain highly accurate characters. Current scenario neural network is used for recognition. Two most important parameters of software are time and accuracy. Proposed system currently concentrates on accuracy of recognition of printed text data by using BPN and Xml data.Only printed data recognition is not sufficient in today's scenario so the next step of this proposed system is to concentrate on handwritten text recognition with high accuracy.

IV REFERENCES

[1]. Mrs. B.Vani, Ms. M. Shyni Beaulah, Mrs. R. Deepalakshmi,"High accuracy Optical Character Recognition algorithms using learning array of ANN", 2014 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[2]. Simone Marinai, Marco Gori, Fellow, IEEE, and Giovanni Soda, Member, IEEE Computer Society, "Artificial Neural Networks for Document Analysis and Recognition", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 1, JANUARY 2005

[3]. Sameeksha Barve,"Optical Character Recognition using Artificial Neural Network" , ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 4, June 2012

[4]. Sameeksha Barve,"Optical Character Recognition using Artificial Neural Network" , ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 2.

[5]. Nan Li, Jinying Chen, Huaigu Cao, Bing Zhang Raytheon, Prem Natarajan, "Applications of Recurrent Neural Network Language Model in Offline Handwriting Recognition and Word Spotting", 2014 14th International Conference on Frontiers in Handwriting Recognition

[6]. Burcu Kır, Cemil Öz, Ali Gülbağ, "The Application of Optical Character Recognition for Mobile Device via Artificial Neural Networks with Negative Correlation Learning Algorithm", 978-1-4799-3343-3/13/$31.00 ©2013 IEEE

[7]. George Nagy, Thomas A. Nartker, Stephen V. Rice, "Optical Character Recognition: An illustrated guide to the frontier", Optical Character Recognition: An illustrated guide to the frontier Procs. Document Recognition and Retrieval VII.

[8]. Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat, " Optical Character Recognition Implementation Using Pattern Matching", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2088-2090.

[9]. Sarojini B.K. Sireesha.K, "A 3-Level Mapped Segmentation based Handwriting Recognition System", Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012