Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi/ International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 1,Jan-Feb 2012, pp.659-662

A Review on the Various Techniques used for Optical Character Recognition

Pranob K Charles[1] V.Harish[2] M.Swathi[2] CH. Deepthi[2]

[1](Associate Professor, Dept. of Electronics and Communications, K L University)
[2](Project Students, Dept. of Electronics and Communications, K L University)

ABSTRACT

Handwriting recognition has been one of the most interesting and challenging research areas in field of image processing and pattern recognition in the recent years. This paper describes the techniques for converting textual content from a paper document into machine readable form. The computer actually recognizes the characters in the document through a revolutionizing technique called Optical Character Recognition. Several techniques like OCR using correlation method and OCR using neural networks are reviewed in this paper.

Keywords: OCR, neural networks, correlation.

I. INTRODUCTION

Optical Character Recognition also referred to as OCR is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning the document [1]. Documents are scanned using a scanner and are given to the OCR systems which recognizes the characters in the scanned documents and converts them into ASCII data. OCR has three processing steps, Document scanning process, Recognition process and Verifying process. In the document scanning step, a scanner is used to scan the handwritten or printed documents. The quality of the scanned document depends up on the scanner. So, a scanner with high speed and color quality is desirable. The recognizing process includes several complex algorithms and previously loaded templates and dictionary which are crosschecked with the characters in the document and the corresponding machine editable ASCII characters. The verifying is done either randomly or chronologically by human Intervention.

Optical Character Recognition is classified into two types, Offline recognition and Online recognition. In offline recognition the source is either an image or a scanned form of the document whereas in Online recognition the successive points are represented as afunction of time and the order of strokes are also available[4][5]. Here in this paper only offline recognition is dealt.

A brief description of the history of OCR is as follows. In 1929 Gustav Tauschek obtained a patent on OCR in Germany, followed by Handel who obtained a US patent on OCR in USA in 1933 . In 1935 Tauschek was also granted a US patent on his method. Tauschek's machine was a mechanical device that used templates and a photo detector. RCA engineers in 1949 worked on the first primitive computer-type OCR to help blind people for the US Veterans Administration, but instead of converting the printed characters to machine language, their device converted it to machine language and then spoke the letters. It proved far too expensive and was not pursued after testing [2][3].

In 1978 Kurzweil Computer Products began selling a commercial version of the optical character recognition computer program. LexisNexis was one of the first customers, and bought the program to upload paper legal and news documents onto its nascent online databases. In about 1965 Reader's Digest and RCA collaborated to build an OCR Document reader designed to digitize the serial numbers on Reader's Digest coupons returned from advertisements. Two years later, Kurzweil sold his company to Xerox, which had an interest in further commercializing paper-to-computer text conversion. Kurzweil Computer Products became a subsidiary of Xerox known as Scansoft, now Nuance Communications.

II. CORRELATION METHOD FOR SINGLE CHARACTER RECOGNITION

A.Preprocessing: The image is taken and is converted to gray scale image. The gray scale image is then converted to binary image. This process is called Digitization of image. Practically any scanner is not perfect, the scanned image may have some noise. This noise may be due to some unnecessary

details present in the image. So, all the objects having pixel values less than 30 are removed. The denoised image thus obtained is saved for further processing. Now, all the templates of the alphabets that are pre-designed are loaded into the system.

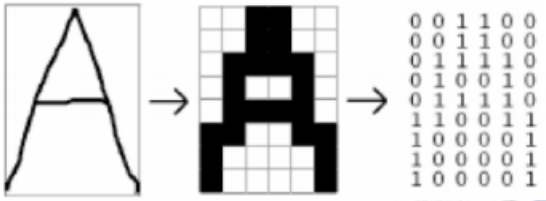

Fig 1. Digitized image

B.Segmentation: In segmentation , the position of the object i.e., the character in the image is found out and the size of the image is cropped to that of the template size.

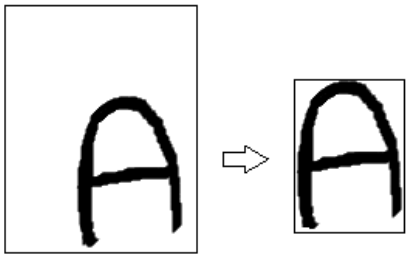

Fig 2. Segmented image

C.Recognition: The image from the segmented stage is correlated with all the templates which are preloaded into the system. Once the correlation is completed, the template with the maximum correlated value is declared as the character present in the image.

D. Remarks: In correlation method there are many unnecessary comparisons and the efficiency of recognition is same for a particular pattern and the given set of templates. However extra templates can be added to the system for providing a wide range of compatibility but doing so will increase the computational intensity of the system. Another important drawback of this method is it requires lot of memory and execution time.

III. CORRELATION METHOD FOR CONTINUOUS CHARACTER RECOGNITION

A.Preprocessing: A noisy image is read from the scanner and is converted to a binary image. The noise is removed from the image by removing all details of the image less than 30 pixels. Then the image is segmented by splitting the image into lines, each line representing a row of words in the image each with a separate label for identification. Now each line consists of different number of words each with many number of letters. Each letter should be separated and resized to the size of the preloaded templates. The recognition process is similar to 'correlation method for single character recognition'.

B.Creating Templates: Images from A to Z and numbers from 1 to 9 are taken into different variables and are preprocessed. All these variables are stored in the form of a cell in which each sub matrix represents a letter. The same process is done for printed upper case, printed lower case, printed numbers, hand written upper case, hand written lower case and hand written numbers. All the model inputs are saved under the same variable name like 'templates.dat' to the hard disk.

C. Dividing into lines: The image is first clipped and an array containing the coordinates of non-zero elements is found. Then the empty row i.e., the row with all elements as zero is found ,this row is taken as the demarcation line to separate the top line from all the lines below it.

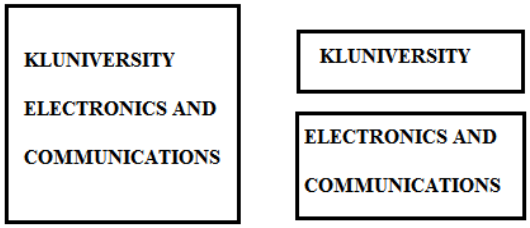

Fig 3. Dividing the image into lines

D.Remarks : This method has the same disadvantages as that of 'correlation method for single character recognition'.

IV. OCR USING ARTIFICIAL NEURAL NETWORKS

A.Artificial Neural Networks: Artificial Neural Networks (ANN) can be likened to collections of identical mathematical models that emulate some of the observed properties of biological nervous systems and draw on the analogies of adaptive biological learning. The key element of an Artificial Neural Network is its structure. It is composed of a number of interconnected processing elements tied together with weighted connections, which take inspiration from biological neurons. The ability to make decisions about imprecise input data makes it useful as a medical analysis tool. There is no need to provide a specific algorithm on how to identify the disease when using a neural network. Neural networks learn by example so the details of how to recognize the disease are not needed. What is needed

is a set of examples that is representative of all the variations of the disease. The quality of examples is not as important as the quantity.

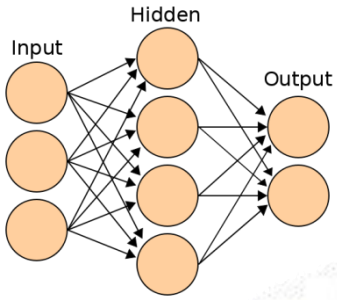

Fig 4. Artificial Neural Network

The Artificial Neural Network can be trained into two main groups that are supervised and unsupervised learning. In supervised learning, the network learns by example whereas in unsupervised method no target value or example is given. Unsupervised learning is very difficult and complex to implement [6] [7].

The neural network receives 35 Boolean values as 35-element input vector. It is then required to identify letter by responding with a 26-element output vector. The 26-elements of output vector each represent a letter. To operate correctly the network should respond with a '1' in position of letter being represented in network. All other values in output vectors should be '0'. In addition, the network should be able to handle noise. In practice, the network doesn't receive a perfect Boolean vector as input. Specifically, the network should make as few mistakes as possible when classifying vector with noise of mean 0 and standard deviation of 0.2 or less.

B.Architecture: The neural network needs 35 input and 26 neurons in its output layer to identify the letters. The network is a two layer "log-sigmoid" network. The log-sigmoid T.F is picked, as its output ranges from 0 to 1 is perfect for learning to output Boolean values. The hidden layer has 25 neurons. This number was picked by trial and error.

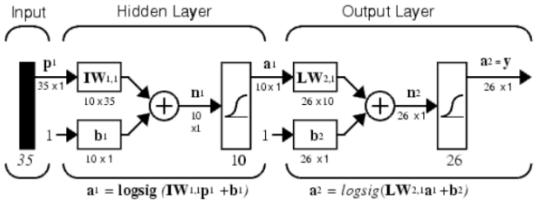

Fig 5. Architecture of back-propagation neural network

C.Training : A two-layer network is created and training is done with and without noise. All training is done using back propagation with both adaptive learning rate and momentum.

V. RESULTS

A. Correlation method for single character recognition:

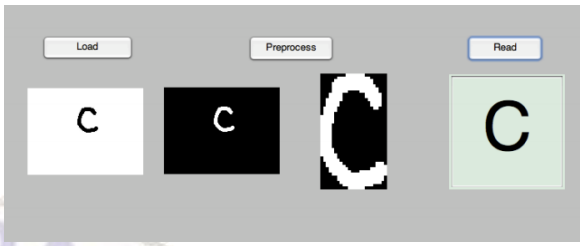

Fig 6. Single Character Recognition using correlation method

First an image which contains the character is loaded, converted to black and white image, cropped to meet the size of the template and then the resultant character is showed.

B. OCR using Neural Networks:

Step 1: Load an image with character from the hard disk and the noise is eliminated in the preprocessing step that is followed.

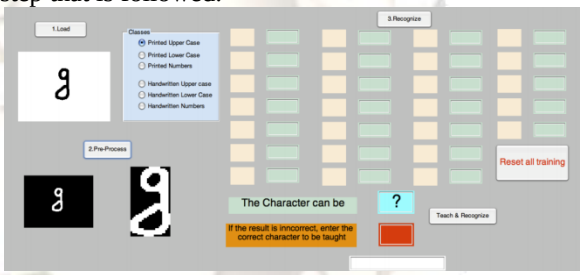

Fig 7. Demonstrating step 1

Step 2: The class of characters from the given set of classes is selected. This can be done by selecting suitable radio button in the classes window. This also fills up all the static text windows with the characters belonging to that class as shown in Fig 8.

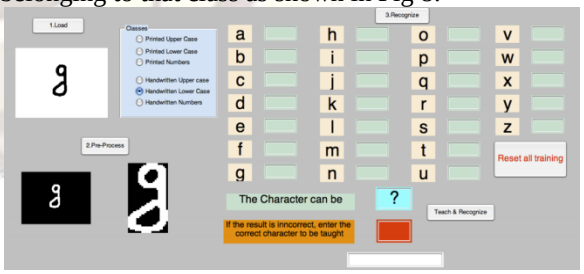

Fig 8. Demonstrating step 2

Step 3: Now the character can be recognized by pushing the Recognize button. This displays the percentage of closeness of the inputted character with characters in that class. The character with the maximum closeness is recognized as the input

character. The result is displayed in the edit text window below as shown in Fig 9.

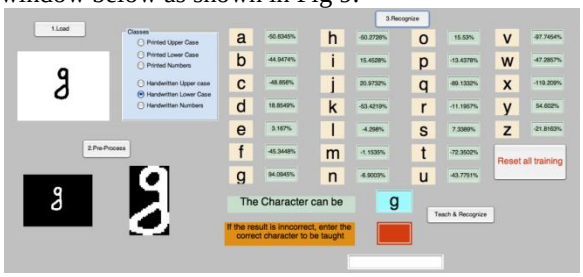

Fig 9. Demonstrating step 3

Step 4: If we get the right result, then the network can be trained to increase the efficiency of recognition. This can be done by pushing "Teach and Recognition" button as shown in Fig 10.

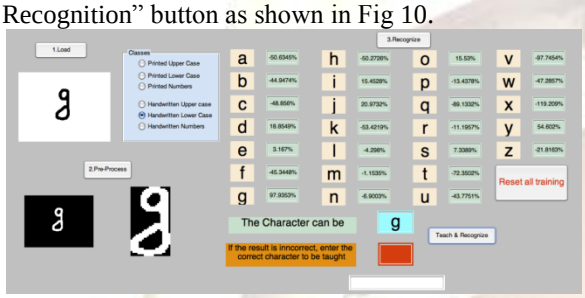

Fig 10. Demonstrating Step 4

It can be observed that the percentage of closeness for recognized character has increased. This shows that even the test data can be used as training data.

Step 5: Sometimes the result can be incorrect. In this situation the correct character is entered in the edit text box shown and the network can be trained by pushing "Teach and Recognize".

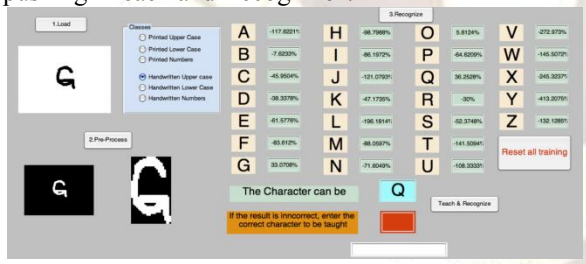

Fig 11. Demonstrating Step 5

When the "Teach and Recognize" is pushed, clearly an increase in percentage of closeness can be observed as shown in Fig 12.

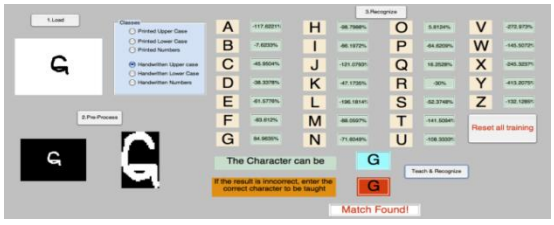

Fig 12. Demonstrating "Teach and Recognize"

Step 6 : If, by mistake the network is wrongly trained, then the training can be reset by pushing "Reset and Training" this turns the network into initial condition.

VI . CONCLUSION

A number of techniques that are used for optical character recognition have been discussed which uses correlation and neural networks. Much other advancement in Optical Character Recognition are being under development. The main research is currently going on in extending Optical Character Recognition to all the popular native languages of India like Hindi, Telugu, Tamil etc., Recognition system works well for simple language like English. It has only 26 character sets. And for standard text there are 52 numbers of characters including capital and small letters. But a complex but organized language like Telugu, OCR system is still in preliminary level. The reason of its complexities are its characters shapes, its top bars and end bars more over it has some modified, vowel and compound characters and also one of the important reasons for poor recognition in OCR system is the error in character recognition.

REFERENCES

[1] UNESCAP, Pop-IT project, 1997-2001
[2] S. Mori, C.Y. Suen and K. Kamamoto, "Historical review of OCR research and development," Proc. of IEEE, vol. 80, pp. 1029-1058, July 1992.
[3] S. Impedovo, L. Ottaviano and S.Occhinegro, "Optical character recognition", International Journal Pattern Recognition and Artificial Intelligence, Vol. 5(1-2), pp. 1-24, 1991
[4] R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey,"IEEE. Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, 2000.
[5] N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, 31(2), pp. 216 – 233
[6] Sang Sung Park,Won Gyo Jung,Young Geun Shin, Dong-Sik Jang, Department of Industrial System and Information Engineering, Korea University, South Korea, "Optical Character System Using BP Algorithm".
[7] Ahmad M. Sarhan, and Omar I. Al Helalat, "Arabic Character Recognition using Artificial Neural Networks and Statistical Analysis".
[8] Arun K Pujari, Prof. C Dhanunjaya Naidu, AI Lab, Un iversity of Hy derabad "An Ad aptive Character Recognizer for Telugu Scripts using Multiresolution Analysis a Associative Memory".
[9] "Telugu ", http://www.nriol.com/telugu-page.asp.