International Journal of Electronic and Electrical Engineering.
ISSN 0974-2174 Volume 7, Number 9 (2014), pp. 979-985

http://www.irphouse.com

Text Extraction from Images

Usha Tiwari[1] Shivani Gupta[2] Nisha Basudevan[3] Piyush D. Shahani[4]

[1]Assisstant professor, Galgotias college of engg. & tech., Gr. Noida, U.P, India
Shivani.frnds2008@gmail.com Nnisha103@gmail.com
Shahani.piyush@gmail.com

Abstract

Handwriting recognition has been one of the active and challenging research areas in the field of image processing and also in context of pattern recognition. It has a number of applications which includes, reading aid for blind people, bank cheques and the conversion of any hand written data into structural text form. An attempt has been made in this paper to recognize handwritten characters for English alphabets and numbers using feature extraction of characterization loci. Each character data set contains 26 alphabets. A number of data sets are used for training the SVM Database. The trained network is used for classification and further for recognition. Then there is this proposed system in which each character is resized into 36x27 pixels, which is directly subjected to training, which basically means that each resized character contains 972 pixels and these pixels are taken as features for training the neural network. The results in the end show us that the proposed system yields good recognition rates which are comparable to other similar schemes for handwriting character recognition**.**

Keywords: OCR, Neural Network topologies, Back propagation algorithm, Feature extraction**.**

I. INTRODUCTION

Handwriting recognition is a wide area of research in the field of image processing and pattern recognition. With the growing computational power character recognition methodologies have been improved and increasing its demand in various applications. It is a difficult task to develop a practical system of handwritten character recognition with high accuracy of recognition. In the existing systems the accuracy of recognizing the text depends immensely on the quality of the input document. Optical character recognition (OCR) is usually referred to as an off-line character recognition process to mean that the system scans and recognizes static images of the characters [1].

Separate classifiers are used for upper and lower case English alphabets in order to increase the accuracy. Handwriting recognition can be divided into two of its kind, online and offline. As in [2] Online method is based on the pen trajectory data while offline method relies on the pixel data only. Now online method provides an advantage that the spatially overlapping characters does not create any problem in segmentation. On the other hand it poses a difficulty in offline method. First the handwritten or printed text is converted into the machine readable form with the help of Optical Character recognition system (OCR).

II. LITERATURE REVIEW

Earlier optical character recognition could be used for activities like expanding telegraphy and creating reading devices for all the blind people[3]. During 1914 a scientist named Emanuel Goldberg had developed a device that read characters and converts them into standard telegraph code. During that time, Edmund Fournier d'Albe was developing an optophone, a handheld scanner which when moved across a printed page, produced tones that helped in recognizing specific characters. But it failed to read non optical characters for which different researches took place. The development took place and ICR (Intelligent Character Recognition) was introduced by M. Sheppard in 1951[4]. Intelligent character recognition (ICR)is an advanced optical character recognition (OCR) or rather more specific handwriting recognition system that allows fonts and different styles of handwriting to be learned by a computer during processing to improve accuracy and recognition levels.

Most ICR software has a self-learning system referred to as a neural network, whose job is to automatically update the recognition database for the fresh handwriting patterns, thereby extending the usefulness of scanning devices for the purpose of document processing, from the printed character recognition (a function of OCR) to hand-written matter recognition, as this process is involved in the recognition of hand writing[5], sometimes the accuracy levels may not be very good but can achieve 97%+ accuracy rates in reading the handwritten content in structured forms. Mostly for achieving these high recognition rates several read engines are used within the software and each is given elective voting rights to determine the true reading of characters. In the numeric fields, engines that are designed to read numbers take preference, whereas in alpha fields, engines are designed to read hand written letters which have higher elective rights. When these are used in conjunction with a bespoke interface hub, the hand writing can be automatically be populated into a back office system avoiding laborious manual keying and can be more accurate than traditional human data entry Intelligent word recognition (IWR) can recognize and extract not only printed-handwritten information, but cursive handwriting as well [6].

III. BASIC PROCESS OF A CHARACTER RECOGNITION SYSTEM

A. Pre Processing

Preprocessing is the first and the major step of OCR software. At this stage certain operations are performed on the scanned image i.e. de-skew, converting an image

from color to black and white, cleans up non-glyph boxes and lines, identifies columns, paragraphs, captions as different blocks and normalization.

B. Segmentation

The aim of image segmentation is to provide label to each pixel in an image such that pixels with the same label share certain visual characteristics. Image segmentation is typically used to locate objects and boundaries (lines, curves etc) in images. The method of segmentation used in this is edge detection

C. Feature extraction

The aim of feature extraction is to capture the essential characteristics of the symbols, and it has been accepted that this is one of the biggest problems of pattern recognition. In this the approach is to extract certain features that characterize the symbols, but leaves out the unimportant attribute. The Selection of the appropriate feature extracting method is probably one of the most important factors in achieving high recognition performance [6].

D. Classification and recognition

The classifying and identifying of each character and assigning to it the correct character class is called classification. In this stage the decision making of a recognition system uses all the features extracted in the earlier stage[7].

E. Post processing

It is the final step of recognition system being discussed. It prints the corresponding characters which were recognized in the structured text form which is done by the calculation of equivalent ASCII value using recognition index of the test samples [8].

IV.ALGORITHMS IMPLEMENTED

A.EDGE DETECTION ALGORITHM

The Edge Detection Algorithm has a list which is called the traverse list, which is defined as the list of pixels that have been traversed by the algorithm

EdgeDetection (x, y, T_L);

1) Add the current pixel to T_L. The current position of pixel is at (x, y).
2) NewT_L= T_L + current position (x, y).
   If pixel at (x-1, y-1) then
   Check for it inthe T_L.

Edgedetection(x-1, y-1, NewT_L);
Endif

   If pixel at (x-1, y) then
   Check for it in the T_L.

Edgedetection(x-1, y, NewT_L);
Endif

```
    If pixel at (x-1, y) then
    Check for it in the T_L.
Edgedetection(x-1, y+1, NewT_L);
Endif

    If pixel at (x, y-1) then
    Check for it in the T_L.
Edgedetection(x, y-1, NewT_L);
Endif

    If pixel at (x, y+1) then
    Check for it in the T_L.
Edgedetection(x, y+1, NewT_L);
Endif

    If pixel at (x+1, y-1) then
    Check for it in the T_L.
Edgedetection(x+1, y-1, NewT_L);
Endif

Edgedetection(x+1, y, NewT_L);
Endif

    If pixel at (x+1, y+1) then
    Check for it in theTraverseList.
Edgedetection(x+1, y+1, NewT_L);
Endif.
```

3) Return;

The Edge Detection algorithm comes to an end when it has covered all the pixels of the character as every pixel's position would be in Traverse List so any further call to Edge Detection is prevented.

B. FEEDBACK BACK PROPOGATION ALGORITHM

It is the most common method of training artificial neural network. It requires a dataset of the desired output for numerous inputs which make up the training set. The back propagation (BP) algorithm's most specific feature is the error produced as the difference of real and desired output value that the neural network gets on its output. When using back propagation, we look at the weight/change, which mostly minimizes the error in the output. All the neurons use log- sigmoid transfer functions. The back propagation algorithm with momentum and adaptive learning rate is used to obtain the parameters of the network. The process of Back propagation learning algorithm is divided into 2 phases: propagation and weight update.

Propagation has following steps:

1. Forward propagation is of the training patterns input which went through the neural network in order to generate the propagation's output activations.
2. Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate the deltas of all output and hidden neurons.

Weight synapse procedure steps have been listed below :

1. Multiply its output delta and input activation to get the gradient of the weight.
2. Subtract a ratio of the gradient from the weight.

The neural network has three layers: an input layer consisting of 100 nodes (for the 10 by 10 letter input), a hidden layer which consists of 50 nodes, and an output layer with 26 nodes (one for each letter). The network makes use of the back-propagation method in addition to bias weights and momentum.

VII. CONCLUSION

We have worked on the classification and Recognition Techniques that are used for handwritten document Images. This detailed discussion will be beneficial insight into various concepts involved, and boost further advances in the area. The accurate recognition is directly depending on the nature of the material to be read and by its quality. Current research is not directly concerned with the characters, but also with words and phrases, and even the complete documents. Over here, we have used the word recognition distances for improving the word matching accuracy. From various studies we have seen that selection of relevant feature extraction and classification technique plays an important role in performance of character recognition rate. Artificial neural networks helped us in performing character recognition which was quite helpful due to its high noise tolerances. These systems have the ability to give excellent results. The feature extraction step of optical character recognition is the most important. We also learnt that a poorly chosen set of features will yield poor classification rates by any neural network. This method gives an estimate for the probabilities of word boundary segmentation using the distances between connected components and thereby combining the segmentation and recognition distances to create a probabilistic word matching similarity. A lot of Research is still needed for exploiting new features to improve the current performance. We also noticed that usage of some specific features that helped in increasing the recognition rate. To recognize strings in the form of words or sentences segmentation phase plays a major role for segmentation at character level and modifier level. So, there is still a need to do the research in this field of character recognition.

VI. RESULTS

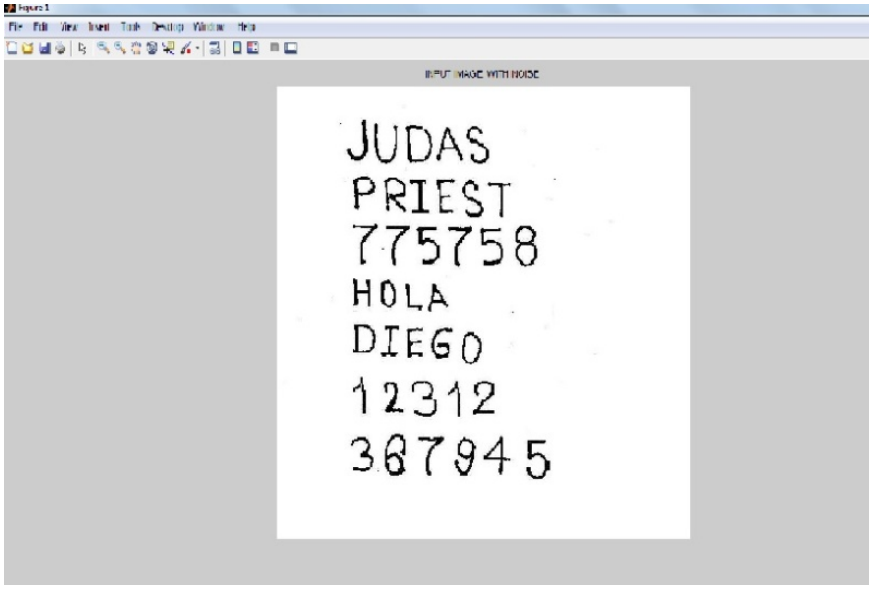

Fig3: Test image

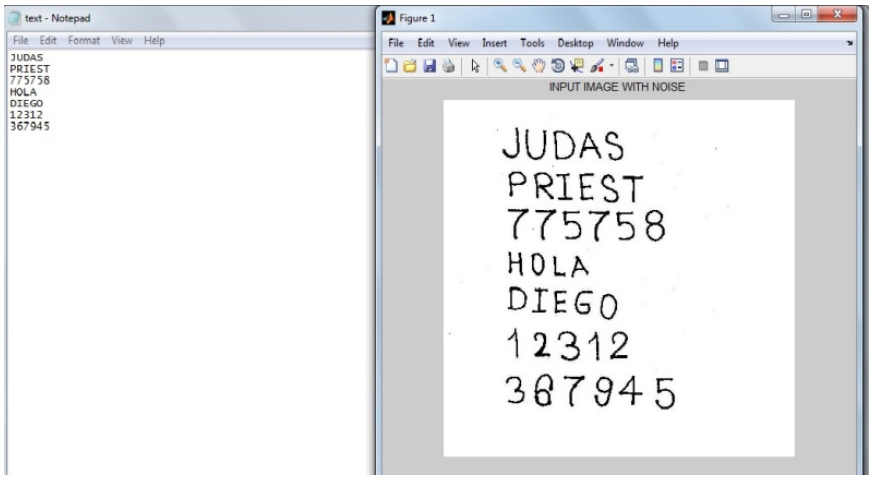

Fig4: Text extracted on notepad

REFERENCES

[1] Vijay Laxmi Sahu, Babita Kubde, ” Offline Handwritten Character Recognition Techniques using Neural Network: A Review “, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064.

[2] R. Seiler M. Schenkel E Eggimann, ”Off-Line Cursive Handwriting Recognition Compared with On-Line Recognition”, 1015-4651/96 $5.00 0 1996 IEEE Proceedings of ICPR '96.

[3] Nafiz Arica and Fatos T. Yarman-Vural, " An Overview of Character Recognition Focused on Off-Line Handwriting", IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 31, no. 2, May 2001.

[4] B. Verma, M. Blumenstein & S. Kulkarni, "Recent Achievements in Offline Handwriting Recognition Systems".

[5] Rejean Plamondon, and Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", IEEE transactions on pattern analysis and machine intelligence. Vol. 22, no. 1. January 2000

[6] J. Pradeep, E. Srinivasan, S. Himavathi, " Diagonal based feature extraction for handwritten alphabets recognition system using neural network", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[7] 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, vol. 31, no. 2, (2001), pp. 216 - 233.

[8] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre, " Character Recognition Using Matlab's Neural Network Toolbox" International Journal of u- and e- Service, Science and Technology Vol. 6, No. 1, February, 2013.

[9] J. pradeep, E. Srinivasan, S. Himavathi, " Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten ", International journal of Engineering. Vol. 25, No. 2, (May 2012) 99-106.

[10] Rokus Arnold, Poth Miklos, " Character Recognition Using Neural Networks", 11th IEEE International Symposium on Computational Intelligence and Informatics • 18–20 November, 2010.

[11] Anshul Gupta, Manisha Srivastava, Chitralekha Mahanta, " Offline Handwritten Character Recognition Using Neural Network, " International Conference on Computer Applications and Industrial Electronics(ICCAIE-2011).

[12] Ankit Sharma, Dipti R Chaudhary, " Character Recognition Using Neural Network", International Journal of Engineering Trends and Technology (IJETT) - Volume4Issue4- April 2013.

[13] Hong Lee and BrijeshVerma, "A Novel Multiple Experts and Fusion Based Segmentation Algorithm for Cursive Handwriting Recognition", 978-1-4244-1821-3/08/$25.00 c 2008 IEEE.

[14] Tasweer Ahmad, Ahlam Jameel, Dr. Balal Ahmad, "Pattern Recognition using Statistical and Neural Techniques", 978-1-61284-941-6/11/$26.00 <S> 20 11 IEEE.