IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 06, Issue 09 (Sep. 2016), ||V1|| PP 64-66

A research on artificial neural network based text retrieval system

[1]Er.Priyadarshni, [2]Dr.J.S.Sohal,
[1]Reasearch Scholar, IKGPTU
[2]Director, LCET Katani Kalan

Abstract:- An artificial neural network based text retrieval system which extract characters from an image file into an editable form, has been implemented using a free and open source computing software i.e Scilab. The proposed system is optimized using various topologies of neural networks. An unsupervised learning algorithm SOM and K means clustering is used to train the network. When comparing with existing neural network based systems our research shows that the proposed system has achieved a significant accuracy of 99-100% which is much higher, with an added advantage of economical availability of the software.

Keywords: *Artificial neural network, Text retrieval system, Self organizing maps, K means clustering, SciLab*

I. INTRODUCTION

Recognizing printed and handwritten text from image documents has been one of the demanding areas of research in the field of image processing and pattern recognition. To build a modern character recognition systems most of the efforts have been dedicated to develop the system using C language, Python[1], MatLab and other licensed software using various algorithm like Artificial neural networks (ANN), Principal component analysis (PCA)[2] ,Support Vector machines (SVM) etc. A Neural network based system using Back propagation algorithm with various preprocessing and feature extraction techniques has been extensively used [3]. The Main objective to design a Back Propagation (BP) neural network is to decide the number of neurons in each layer of network and number of hidden layers by taking into account the complexity and learning speed of the system [4]. Accuracy rate of BP neural network is higher than genetic algorithm (GA) albeit its complexity [5]. On the other hand, BP NN when combined with GA has fairly reduced the false recognition rate [6]. To meet the requirements of real time automatic number plate recognition (ANPR) system a feed forward ANN based OCR is an alternative to slower system [7]. Because many classifiers particularly ANN, are flexible in performance and are also affected by human factors so fair comparison of classifiers is inconsiderable [8]. To compare classifiers, standard pre-processing and feature extraction techniques should be applied to all the classifiers. Many techniques such as nonlinear normalization and direction feature extraction are variable in their performance details [9]. Due to drawbacks like slower convergence and longer training times associated with conventional back propagation algorithm, unsupervised Self-Organizing Map (SOM) which is a data-analysis method that visualizes similarity relations in a set of data items is preferred. For instance it has been applied to the comparison of enterprises at different levels of abstraction, to assess their relative financial conditions and to profile their products and customers in economy. Furthermore, in industry, the monitoring of processes, systems and machineries by the SOM method and in science and technology at large, there are number of tasks where the research objects must be classified on the basis of their intrinsic properties, to reveal the classification of proteins, genetic sequences and galaxies [10]. SciLab being a freely distributed and open source scientific software package is increasingly used in educational institutions, research centers and companies around the world, providing a powerful open computing environment for engineering and scientific applications [11]. Scilab is most proficient alternative for pattern recognition application as compare to MatLab and other licensed and expensive software [12].

II. METHODOLOGY

The proposed project, require SciLab an open source software version which can be downloaded freely from the internet. After installing the Scilab the required toolboxes Artificial Neural Network (ANN), Image Processing Design (IPD) and Scilab Image and Video Processing Toolbox (SIVP) are selected from module manager which are downloaded and installed. Using Scilab as a simulation tool an ANN model is made to recognize the characters from an image. Neural networks are particularly well suited for addressing non-linear problems like recognizing the characters and have proved themselves as proficient classifiers.Firstly character image which has a specific format such as JPEG, BMP etc is acquired by the recognition system in image acquisition process. Various experimental images with various characters having different font's style, size and different color backgrounds which are used for the proposed system are shown in Fig.1 to Fig.4.

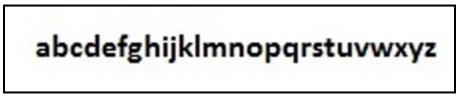

Figure1 : Standard lower case alphabets

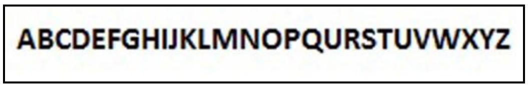

Figure 2: Standard upper case alphabets

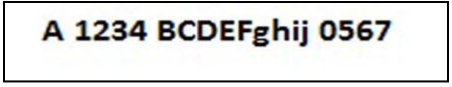

Figure 3 : Variable spacing Text

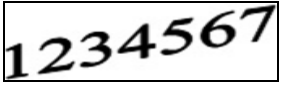

Figure 4:Tilted orientation Numerals

In pre-processing a chain of operations is performed on an input image to essentially upgrade the image to make it more suitable for segmentation. Various steps for preprocessing like resizing the given image to select number of horizontal and vertical pixels, converting the image into grayscale image using RGB2gray command, Converting the grayscale image into binary image by using a suitable threshold are used in SciLab 5.5.2.

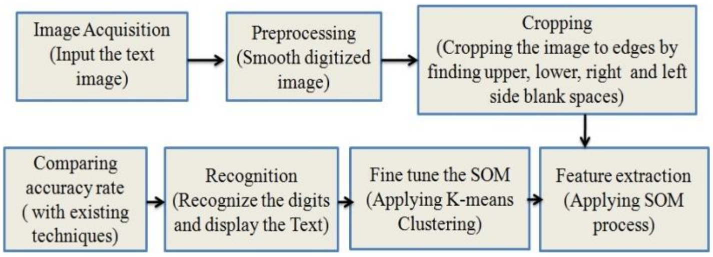
Image Acquisition (Input the text image)
Preprocessing (Smooth digitized image)
Cropping (Cropping the image to edges by finding upper, lower, right and left side blank spaces)
Feature extraction (Applying SOM process)
Fine tune the SOM (Applying K-means Clustering)
Recognition (Recognize the digits and display the Text)
Comparing accuracy rate (with existing techniques)

Figure 5 .Block Diagram of Proposed system

Fig.5 Block diagram of the proposed work interpret the string of operations performed to implement the proposed system. After acquiring the image from which text is to be extracted, the preprocessing is essential to improve the image with limited distortions. Blurring the background of image is done in smoothing process to extract the required information and making the image more informative as compared to original acquired image. After finding all the four boundaries image document is cropped. To achieve higher recognition accuracy and speed, self organizing Map (SOM) process of ANN is used.

III. IMPLEMENTATION AND EXPERIMENTATION

To solve the problem of classification of large data sets, determining the clusters using an unsupervised training paradigm is very important. This is done by capturing the essence of the resemblance exhibited by the stimuli, so that data of related type include one cluster and elements that are unrelated are assigned to different subsets. Kohenen Self organizing Map (SOM) invented by Teuvo Kohenen, one of the significant families of ANN is used to solve these problems and is an efficient way of representing multidirectional data in much lower dimensional spaces. Additionally, a network can be created that stores the information so that topological relationships are not altered within the training set.

SOM clustering of the input vector is done by calculating neuron weight vector according to Euclidean distance, which is used to find the similarity between the input vector and the map's node's weight vector. As the actual numerical distance from the input is not required so it is squared and some sort of uniform scale is made to compare each node to the input vector. Best Matching Unit (BMU) which is the winner output neuron with minimum distance is selected. The radius of the neighborhood of the BMU is a very large value in the initial stage and shrink with each time-step is calculated. Any nodes found within the radius of the BMU are modified to make them more similar to the input vector. The closer a node is to the BMU, the more its' weights are amended.After updating the weights of neurons and its neighborhood fine tuning the SOM is done by K means clustering algorithm which gives the benefit of reduction of the quantizing error and also results into more effective convergence.

IV. CONCLUSION

Several tests have been conducted on various character images, some of which were designed in MS paint itself for which the proposed system has achieved 100 % accuracy while for the images collected from web accuracy was comparatively lesser depending upon font style and presence of noise in the given image. This procedure allows the network to find its own solution, thus making it more efficient with pattern association. The recognition rate was considerable increased in the proposed method as compare to that of BP neural network (85%) and combination of BP and PCA algorithm (87.5%)[2].Moreover, Scilab being open source software has an added advantage of its cost-effectiveness.

Numerals with various font styles and size in gray scale as well as colored images are extracted successfully by implementing SOM process and K-means clustering algorithm using Scilab. Recognized characters appear in a picture box as well as in SciLab console window which is in editable format. The system has achieved accuracy of 99% to 100% which proved to be much higher than existing pattern classifier.

REFERENCES

[1] Shah.J, Gokani.V, Recognition System for Digits Recognition using the Pixel-Contour Features and Mathematical Parameters, *International Journal of Computer Science and Information Technologies, Vol. 5 (5)* , 2014, pp 6827-6830

[2] Liu. P, Guo J , Yu. Z, Li. H, Xian .Y, The design of digit recognition teaching experiment based on PCA and BP neural network' *Proc. . Conf. Chinese Control and Decision*,2015, pp 4132 - 4135

[3] Gao Y, Deng C, Guoxing, Application of BP neural network in the digital recognition system,*Proc Int. Conf., Jiang Computer Engineering and Technology (ICCET)*, 2010 , 6, pp V6-522 - V6-526

[4] Yu. H, Guo J, Yu. Z, Xian.Y, Chen. P, The Design of Digit Recognition Teaching Experiment Based on BP Neural Network, *25th Chinese Control and Decision Conference (CCDC)*, 2013, pp 5109 - 5113,

[5] Ali M.S, Mondal M.N.I, Character Recogntion System: Performance Comparison of Neural Networks and Genetic Algorithm, *Proc. Int. Conf. Computer and Information Engineering (ICCIE)*, 2015,pp 91 – 94

[6] Liu L, Qi H, Liu.J, The Research of Alphabet Identification Based on Genetic BP Neural Network, *Proc. Int. Conf.* ,2012, Volume: 1, pp 25 – 28

[7] Zhai.X, Bensaali.F, Sotudeh .R, OCR-based neural network for ANPR, *Proc. Int. Conf. on Imaging Systems and Techniques Proceedings*,2012, pp 393 – 397

[8] Duin R.P.W, A note on comparing classifiers, *Pattern Recognition Letters, Elsevier science*,1996, pp 529—536

[9] Liu.C.L , Fujisawa. H, Classification and Learning Methods for Character Recognition, *Advances and Remaining Problems by in Machine Learning in Document Analysis and Recognition*, Volume 90 ,pp 139-161

[10] Kohonen. T, *MATLAB Implementations and Applications of the Self-Organizing Map* (Unigrafia Oy, Helsinki, Finland, 2014).

[11] Tona. P, Teaching process control with Scilab and Scicos, *Proc. Int. Conf. American Control Conference,* June 2006, Minneapolis, MN Publisher

[12] Priyadarshni, Sohal J.S, Improvement of artificial neural network based character recognition system, using SciLab, *Optik - International Journal for Light and Electron Optics*, 127(22), November 2016, pp 10510–10518

[13] Gonzalez. C, Woods R.E, *Digital Image Processing* (Third Edition Rafael University of Tennessee, Pearson Education Upper Saddle River, New Jersey, 2008)