Multi-Layer Perceptron (MLP) Neural Network Technique for Offline Handwritten Gurmukhi Character Recognition

Gurpreet Singh
SLIET Longowal, Punjab
Punjab, India
gurpreet32@gmail.com

Manoj Sachan
Associate Professor
SLIET Longowal, Punjab
Punjab, India
manojsachan@gmail.com

***Abstract*— Machine vision researchers are working on the area of recognition of handwritten or printed text from scanned images for the purpose of digitizing documents and for reducing the errorless data entry cost. The classic difficulty of being able to correctly recognize language symbols is the complexity and the irregularity among the pictorial representation of characters due to variation in writing styles, size of symbols etc. Character recognition process depends on, how the input data is given to the system. Input data may be categorized as Online data or Offline data. Both the forms of data input have their own issues. In this paper, we are focusing on the Offline Gurmukhi character recognition from text image. There are lot of complexities associated with Gurmukhi Script. In this paper, we present a technique based on Multi Layer Perceptron (MLP) Neural Network model. Here we consider isolated handwritten Gurmukhi characters for recognition. MLP is used because it uses generalized delta learning rules and easily gets trained in less number of iterations. The proposed method in this paper detect graphical symbols by identifying lines and characters from the image. After that it analyzes the symbols by training the network using feed forward topology for a set of desired unicode characters. We achieve the performance rate of proposed system maximum up to 98.96% for recognition of symbols by using MLP neural network.**

Keywords— ***Digitizing documents, Offline recognition, Gurmukhi Script, MLP, Feed Forward topology, Unicode.***

I. INTRODUCTION

In modern era, we require all the useful data in digital form. The record keeping in digital form is necessary as the scanned data require large amount of storage space. Another advantage of digital data over scanned data is that, we can perform operations like editing, searching etc. over it. Because of these storage and manipulation issues, the scanned data must be converted to digital form. Digital data or data in electronic form takes less storage as compared to images or scanned records. The other part of consideration in case of data present in the scanned form is its mode of the input. So on the basis of mode of inputting characters, digital scanned document classification can be done as Printed Offline, Handwritten Offline and Handwritten Online [3]. In case of both offline printed and handwritten data, the characters first written on the paper before scanning. In case of Online process, the user input the characters with the help of some electronic device such as Digitizer Tablet and Pen [15, 16]. The process of conversion of data from scanned to digital form is either manual or automatic. In case of manual conversion, data entry operators are used. With manual task two problems are associated, one is the cost of data entry and second is the possibility of erroneous data entry by the operators. On the other hand, in case of automatic conversion process, digital data is constructed from scanned images by using Optical Character Recognition (OCR) methods. OCR systems are used for applications such as document automations, cheque verification etc. [2,5].

The task of recognizing characters from scanned images is very difficult. The text contained in the scanned images may be in the form of typed characters or handwritten characters [10,12]. The recognition process in case of handwritten characters again have to deal with some issues like variation in writing styles by different users, text written in different languages etc. The complexity issues related to different languages and their scripts also affect the recognition process [11,13].

Fig.1. Character set of Gurmukhi Script

978-1-4799-3975-6/14/$31.00 ©2014 IEEE

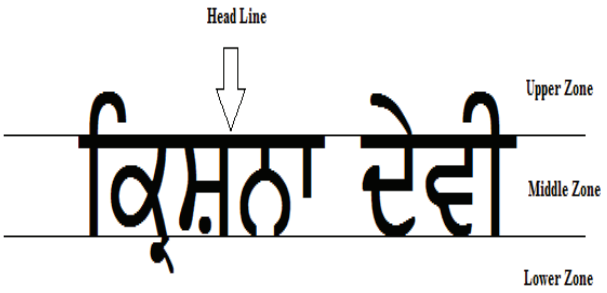


Fig. 2. Gurmukhi script word showing different zones

The OCR systems are made according to the choice of language whose characters we want to recognize from scanned image. One of the popular language of India is Punjabi. It is the 10th most spoken language in the world. The script used to write Punjabi language in India is Gurmukhi. The characters of Punjabi language are shown in Fig.1.

The recognition of characters from handwritten scanned documents of Punjabi language are difficult because of some properties of Gurmukhi script. In this script, the characters which form a complete word are joined with a horizontal line at the upper portion of these characters. This line is known as Head line. Some part of the word appear above the head line and other part appear below it. Also, the boundary boxes of two or more characters may overlap vertically. The words are considered to be exist in three different zones as upper, middle and lower zone [14]. Fig.2. shows a word written in Punjabi language using Gurmukhi script. This figure shows the upper, middle and lower zone where the part of the word can exist, this information is necessary for proper segmentation of word into characters.

In this paper, we have presented the recognition of offline Gurmukhi handwritten isolated characters. ANN is used because of its parallel execution and non-linear nature. In this paper, we have used the concept of Multi-Layer Perceptron (MLP) neural network for character recognition using feed forward topology.

In the next part of the paper, firstly, we go through the related work in the field of OCR for various Indian languages. After that we have the introductory portion for MLP neural network. Then the next section explain the algorithms, which we develop to deal with the problem under consideration. In the last section, performance of the proposed system is explained mathematically with the help of tables and graphs.

II. RELATED WORK

Bag et al. [1] used the structural characteristics of compound "Bangla" characters for their detection and recognition. Skeleton of the characters and straight lines are used by them to recognize the compound characters.

Razak et al. [3] concluded that character recognition for "Urdu" script based languages have some technical issues as compared to other languages, because of these issues OCR systems for Urdu script languages face low accuracy in character recognition process. Authors proposed two distinct methods, HMM and Fuzzy logic classifier for character recognition. Input strokes are divided into 62 classes, based on starting and ending styles of ligature using fuzzy rules. Their system works for both handwritten Nasta'liq and Nasakh font of "Urdu" and produced 87.6% and 74% classification results respectively.

Khobragade, Koli and Makesar [4] reviewed different sub tasks of Optical Character Recognition (OCR) such as pre-processing, segmentation, feature extraction, classification and matching techniques. They implement, the different techniques used to perform these tasks and observe different characteristics of considered techniques.

Rekha. [6] observed that by using SVM as a classifier in offline handwritten Gurmukhi characters and numeral recognition, accuracy of 95.05% is achieved using zoning and BDD features for character recognition and 99.20% accuracy is achieved in case of numeral recognition.

Bansal, Garg and Kumar [7] proposed a new technique of offline handwritten Gurmukhi character recognition. For classification they use three different classifiers SVM, MLP and Naive bayes with 5 fold and 10 fold cross validations. Authors achieved 91.95% accuracy with SVM, 87.30% with MLP and 77.70 with Naive bayes.

Bharath & Madhavanath [8] highlight the challenges in recognizing Indic scripts and present the overview of the state of the art approaches developed for isolated character recognition and word recognition.

Sharma & Singh [9] presented a survey for classification of various character recognition techniques by considering the factor that the variation in handwriting among different writers occur because of speed of writing, different styles, size or position of characters.

III. MULTI-LAYERED PERCEPTRON (MLP) NEURAL NETWORK

The area of Artificial Neural Network derives its basis from the way neurons interacts and function in human brain. The human brain is known to operate in a parallel manner for recognition, reasoning and damage recovery. Because of the ability of ANN to deal with above kind of processes, it can be used from simple applications to the complex applications like pattern recognition algorithms. One neuron in the network at one time is able to link with more than 10,000 other neurons to generate and share new knowledge. Neurons are linked with other neurons in the network through links known as synapse. Neurons in the network receive many inputs either from other neurons or directly as original data. Each neuron has a single threshold value also. Fig.3. represent the basic structure of ANN, where three layers input, output and hidden are present. Each layer contain some neurons and each neuron have some weight associated with it for further processing.

The activation of the neuron or PSP (Post Synaptic Potential) can be calculated as the weighted sum of the inputs subtracted by the threshold of the neuron. Network generates its output by passing activation signal through activation or transfer function [12]. A Feed-Forward network has neurons arranged in a distinct layer topology. The input layer introduces the values of input variables.

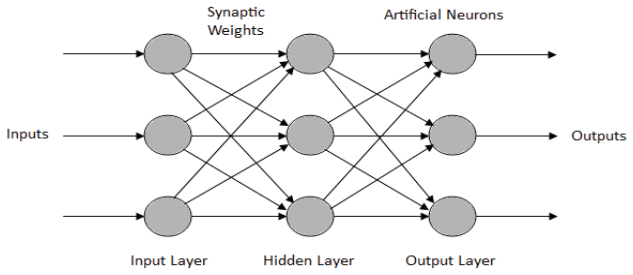


Fig. 3. Basic architecture of artificial neural network

The hidden and output layers have neurons connected to the neurons of their preceding layers. The network connection in this type of topology may be fully connected or partially connected. In MLP neural network, each unit perform a biased weighted sum of inputs to them and pass this activation level through a transfer function to generate output. The most common activation function in MLP are logistic and hyperbolic tangent sigmoid functions. In this paper, for the purpose of recognition we used hyperbolic tangent sigmoid function.

IV. METHODOLOGY

MLP implementation for the proposed technique in this paper is composed of three layers: Input layer, Hidden layer and Output layer. The symbols are represented after segmentation by a pixel matrix of order (10 X 15). So its pixel information is input to MLP network's input layer which contain total 150 neurons. The hidden layer consist 250 neurons for network training and output layer consist 16 neurons as it produces 16-bit unicode. The proposed method of isolated offline character recognition for Gurmukhi script is as per the following:

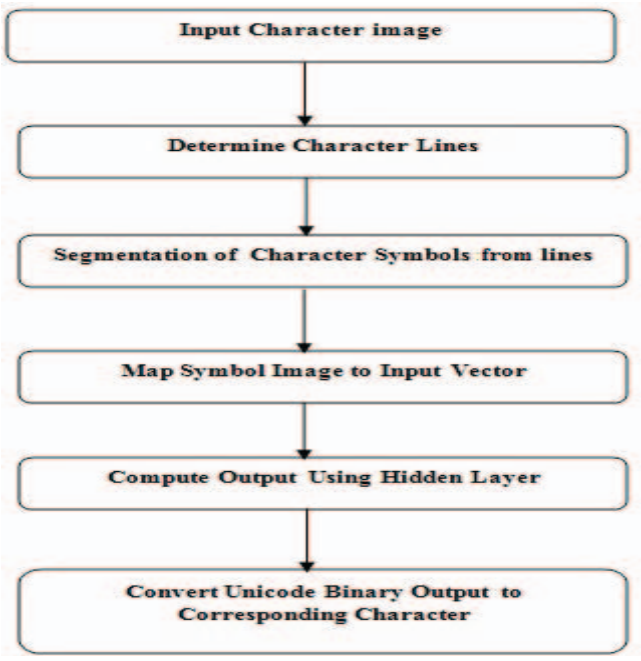


Fig. 4. Proposed system for character recognition

In our system, we input the scanned image. This image contain characters from Gurmukhi script. Next step is the identification of character lines. After identification of lines different character symbols from each line are extracted. Then the segmented character symbols are converted to the matrix form. Then the information of this matrix is passed to the MLP neural network as input for further classifications. MLP neural network had been trained by back propagation algorithm. Data had been collected from 25 different users. leaning rate is observed as 0.7. Next algorithm represents the extraction of lines and individual symbols from the image.

Algorithm for segmentation of character line and symbols:

Step-1. Start from the origin of the scanned image i.e. x=0,y=0 (f(x , y)=f(0 , 0)).
Step-2. Scan up to width of the image by keeping 'y' constant and incrementing 'x' by one each time.
Step-3. If a black pixel found, then register value of 'y' as top of line. If no black pixel found, Increment 'y' by one and reset the value of 'x' and rescan up to the width of image .
Step-4. Repeat the Steps 2 and 3 for new values of 'y' if in a scan no black pixel encountered then set y-1 as bottom of the line.
Step-5. Repeat step 1-4 to detect new line and symbols.
Step-6. Stop at the bottom of the image.

Table 1. shows a snapshot of the handwritten samples taken from various persons, where column 1 of the table represent the Gurmukhi character according to font Gurmukhi_Normal and other columns show the same characters written by the different users for checking the performance of the system.

TABLE1. Snapshot of Dataset

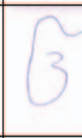
Character	Sample of Person_1	Sample of Person_2	Sample of Person_3
ਓ			
ਅ			
ਏ			
ਸ			
ਘ			

Fig. 5. Results of Segmentation phase

Fig.5. represents the results of the different characters segmented from the input image and also shows the conversion of these segmented character to the vector array for generating the weights of the neurons present in the first layer of ANN.

V. RESULTS & DISCUSSIONS

For checking the performance of the proposed system, we collected the handwritten samples of isolated Gurmukhi characters from 25 different persons. The system gives very good results with 98.96% recognition rate of correctly segmented Gurmukhi characters and if we include the segmentation error, still the system performs well and recognition rate becomes 97.71%, which is again noticeable one. Table 2. shows the results in tabular form, where total 481 characters are there for classification purpose. This table display results by considering both, the effect of segmentation error and without it. Fig. 6. represents the same results of Table 2. in graphical form.

TABLE 2. Performance results of proposed system

Number of Persons	Total Number of Characters	%age of Characters correctly recognized (Including segmentation error)	%age of Characters correctly recognized from correctly segmented characters
25	481	97.71	98.96

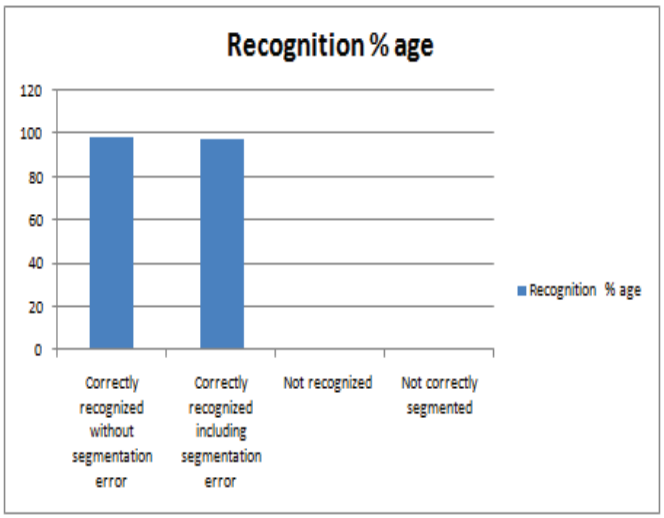


Fig. 6. Graph representing recognition rate

VI. CONCLUSION

In this paper, we have presented a new technique for the recognition of isolated offline handwritten characters of Gurmukhi script. For recognition purpose we have used the concept of MLP neural network because of its parallel execution capability to deal with the complexity associated with Gurmukhi script characters. We use the spatial information in the form of image coordinates for segmentation of character lines and character symbols from the offline handwritten data. 98.96% rate of recognition of Gurmukhi script characters shows the efficiency of the proposed system. In future we can work on the issue of Gurmukhi offline data containing words of Gurmukhi script with more complexities as compared to isolated characters.

REFERENCES

[1]. S. Bag, G. Harit, P. Bhowmick, " Recognition of bangla compound characters using structural decomposition" Elsevier, Pattern recognition (47), pp. 1187-1201, September 2013.

[2]. S. Naz, K. Hayat, M. I. Razak, M. W. Anwar, S. A. Madani, S. U. Khan, " The optical character recognition of Urdu like cursive script", Elsevier, Pattern recognition (47), pp. 1229-1248. October 2013.

[3]. M. I. Razak, F. Anwar, S.A. Husain, A. Belaid, M. Sher, "HMM and fuzzy logic: a hybrid approach for online urdu script-based language character recognition", Elsevier, Knowledge based system (23), pp. 914-823, July 2010.

[4]. R.N. Khobragede, N. A. Koli, M. S. Makesar, " A survey on recognition of devnagri script", International Journal of Computer Applications & Information Technology, Vol. 2, pp. 22-26, January 2013.

[5]. M. Kumar, M. K. Jindal, R. K. Sharma, " K-nearest neighbour based offline handwritten gurmukhi character recognition", International conference on Image Information processing, IEEE-ICIIP, 2011.

[6]. A. Rekha, " Offline handwritten character and numeral recognition using differen feature sets and classifiers- A survey", IJERA, Vol. 2, pp. 187-191, June 2012.

[7]. S. Bansal, M. Garg, M. Kumar, " A technique for offline handwritten charachter recognition", IJCAT, Vol. 1, Issue 2, pp. 2010-2015, March 2014.

[8]. A. Bharath, S. Madhavanath, "Online handwriting recognition for Indic scripts", HP laboratories, May 2008.

[9]. P. Sharma, R. Singh, " Survey and classification of character recognition system, International journal of engineering trends and technology, Vol. 4, Issue 3, pp. 316-318, 2013.

[10]. N.K. Garg, L. Kaur, M. Jindal, " Recognition of offline handwritten hindi text using SVM", International Journal of Image Processing(IJIP), Vol. 7, Issue 4, pp. 395-401, 2013.

[11]. G. S. Lehal, " A survey of the state of the art in punjabi language processing" Language in India, Vol. 9, pp. 9-23, October 2009.

[12]. K. S. Siddharth, R. Dhir, R. Rani, " Handwritten gurmukhi numeral recognition using different feature sets", International Journal of Computer Applications, Vol. 28, pp. 20-24, August 2011.

[13]. N. Garg, S. Kaur, " Improvement in efficiency of recognition of handwritten Gurmukhi script", International Journal of Computer Science and Technology (IJCST), Vol. 2, Issue 3, pp. 158-161, September 2011.

[14]. G. S. Lehal , C. Singh, " A gurmukhi script recognition system", 0-7095-0750-6,IEEE, pp. 557-560, 2000.

[15]. M. K. Sachan, G. S. Lehal, and V. K. Jain, “A Novel Method to Segment Online Gurmukhi Script”, Proceedings of International Conference on Information Systems for Indian Languages, ICISIL 2011, Patiala, Communications in Computer and Information Science, Vol. 139, Springer-Verlag Berlin Heidelberg, Germany, pp. 1-8, 2011.

[16]. M.K. Sachan, G.S. Lehal, V. K. Jain, ”A System for Online Gurmukhi Script Recognition “,Proceedings of International Conference on Information Systems for Indian Languages, ICISIL 2011, Patiala, Communications in Computer and Information Science, Vol. 139, Springer-Verlag Berlin Heidelberg, Germany, pp. 294-295, 2011.