A RECOGNITION SYSTEM FOR HANDWRITTEN GURMUKHI CHARACTERS

Mandeep Kaur[1], Sanjeev Kumar[2]
Department of Electronics and Communication
Amritsar College of Engineering and Technology
Amritsar-143001, Punjab, India

Abstract

This paper represent Handwritten Gurmukhi Character Recognition system using some statistical features like zone density, projection histograms , 8 directional zone density features in combination with some geometric features like area, perimeter, eccentricity, etc. The image document is first pre-processed by using many techniques like binarization, morphological operations (erosion and dilation) applied to remove noise and then segmented into isolated characters. The highest accuracy obtained by using these features and back propagation classifier is 98%.

1. Introduction

Since the inception of computers we are witnessing a great deal of research activities in the field of computer human interface. The natural handwriting is a very easy way of exchanging information between computers and human beings. The quick and natural way of communication between users and computers is inputting the data through handwritten documents. Variation in handwriting is one of the most prominent problems and achieving high degree of accuracy is a tedious task. These variations are caused by different writing styles.

Handwritten character recognition is mainly of two types online and offline. In online handwriting recognition, data is captured during the writing process with the help of a special pen on electronic surface. In offline handwriting recognition, prewritten data generally written on a sheet of paper is scanned. The process of handwritten character recognition of any script can be broadly classified into five stages i.e. Pre-processing, Segmentation, Feature Extraction, Classification and Post-processing. The first important step for recognition is pre-processing followed by segmentation and feature extraction. The selection of appropriate feature extraction method is probably the single most important factor in achieving high recognition performance. Artificial neural networks (ANN), Support Vector Machines (SVM), Structural pattern recognition, etc can be used for classification process. The neural networks have emerged as the fast and reliable tools for classification towards achieving high recognition accuracy.

Handwritten Character Recognition has applications in postal code recognition, automatic data entry into large administrative systems, banking, digital libraries and invoice and receipt processing.

2. Introduction to Gurmukhi Script

Gurmukhi script [1][11] is the most common script used for writing the Punjabi language. Gurmukhi was standardized by the second Sikh guru, Guru Angad Dev Ji. The whole of Sri Guru Granth Sahib Ji's1430 pages are written in this script. The name Gurmukhi is derived from the old Punjabi term "Guramukhi" meaning "from the mouth of the Guru". Gurmukhi script is 12th most widely used script in the world.

There is rich literature in this language in the form of scripture, books, poetry, etc. Gurmukhi is the first official script adopted by Punjab state. So it is important to develop a recognition system for such a rich and widely used language.

Writing style of Gurmukhi script is from top to bottom and left to right. In Gurmukhi script, there is no case sensitivity and most of the characters have a horizontal line at the upper part called headline and characters are connected with each other through this line. It consists of 41 consonants.

3. The Character Recognition System

The character recognition system consists of the different phase's digitization, pre-processing, feature

extraction and classification. The detail of each phase of handwritten Gurmukhi character recognition system is given in the proceeding sections. The block diagram of proposed recognition system is given in figure 1.

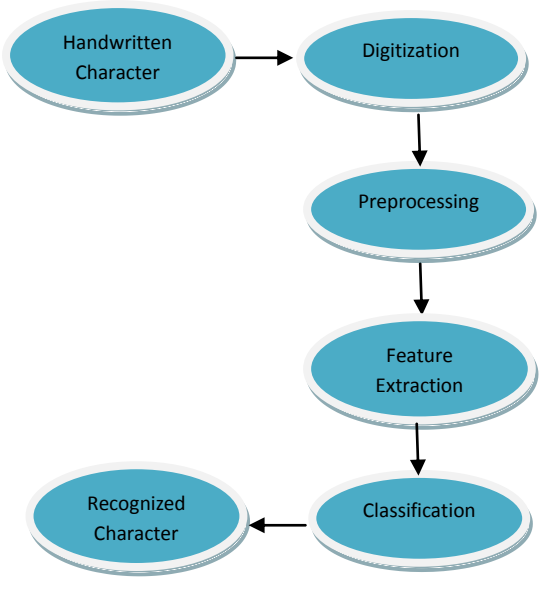


Figure 1: Block diagram of the character recognition system

4. Image Pre-Processing

In the pre-processing technique, the handwritten data is first converted into electronic form with the help of an optical scanner. This process is called digitization. Then the coloured image is first converted into gray-scale image and then gray-scale image is converted into a binary image using thresholding technique in which the information (object) of an image is separated from its background.

Digital images are prone to a variety of types of noise. Noise is the result of errors that occur during the image acquisition process. Noise is reduced by using some morphological operations such as dilation (to bridge unconnected pixels) and erosion (to remove isolated pixels and to remove spur pixels) to improve the quality of the document.

After the removal of noise the text image is segregated to form characters. The segmentation provides the separation of different logical parts, like lines of a paragraph, words of a line and characters of a word. The pre-processing of an image is shown in the figures below:

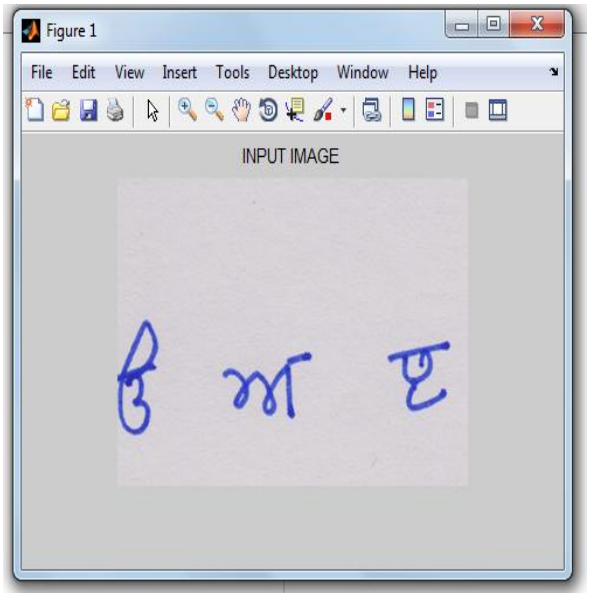


Figure 2: The scanned image of handwritten characters

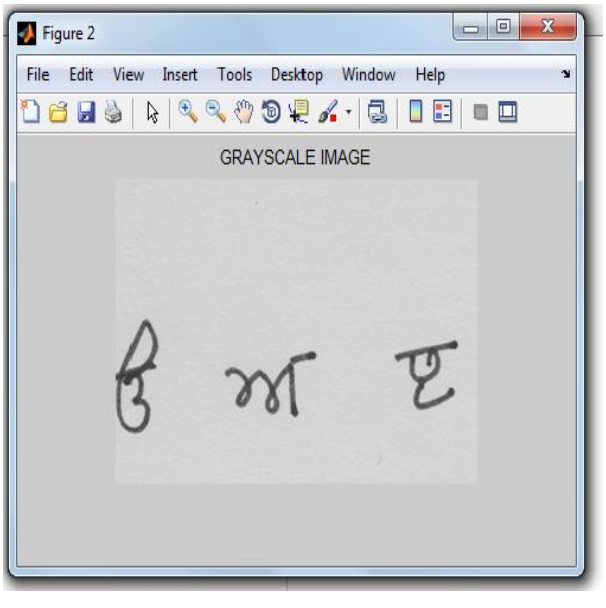


Figure 3: The coloured image is converted into gray-scale image

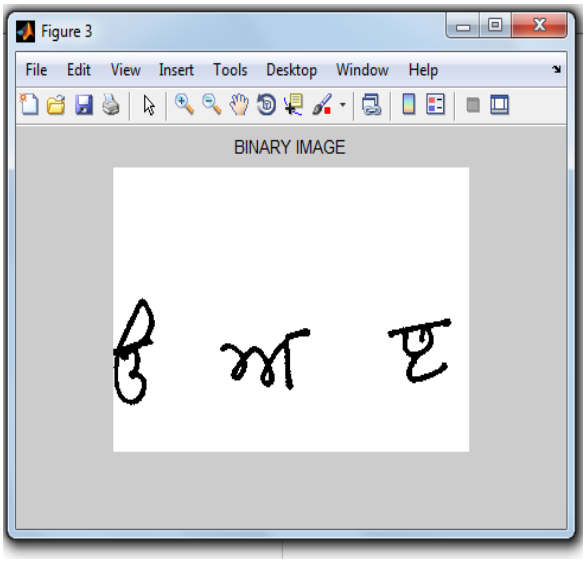


Figure 4: The gray-scale image is converted into binary image

5. Feature Extraction

Feature extraction is the important step which is used to extract the most relevant information which is further used to classify the objects. In our research work, we have used following listed features.

1) **Area:** Area of characters in a binary image, the number of non-zero pixels in a character.

2) Perimeter: Perimeter of characters in a binary image, the length of the smoothest boundary in pixels.

3) Area/Perimeter: The ratio of area to perimeter.

4) Minor Axis Length: Minor Axis Length is the length in pixels of the minor axis of the ellipse that has the same normalized second central moments as the region.

5) Major Axis Length: Major Axis Length is the length in pixels of the major axis of the ellipse that has the same normalized second central moments as the region.

6) Orientation: Orientation is the angle in degrees ranging from -90 to 90 degrees between the *x*-axis and the major axis of the ellipse that has the same second-moments as the region.

7) Eccentricity: The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1.

8) Zone density: In zoning, the character image is first resized into 32*32 pixel size and then divided into N*M zones where N is the number of rows and M is the number of columns. The density of each zone is obtained by dividing the foreground pixels in each zone by total number of pixels in each zone.

9) Projection Histogram: Projection histograms count the number of pixels in a particular direction and that direction can be horizontal, vertical or diagonal. In this approach, the character image is resized into 32*32 pixel size. The projection histograms are computed by counting the number of foreground pixels. In horizontal histogram these pixels are counted by row wise. Similarly, we can count the number of foreground pixels column wise and diagonal wise.

10) 8 -Directional Zone Density: The character is divided into 8 directions and we get 8 triangular zones. The density of each zone is then calculated by dividing the foreground pixels by the total number of pixels in each zone. Then this density of each zone is considered as a feature.

6. Classification

The type of neural network model used here is Feed Forward Multilayer Perceptron based. There are three layers namely the input, output and the hidden layer. Each element in the hidden and the output layer is fully connected to the elements in the input and hidden layer respectively. Feed-forward MLP neural network architecture is being used and trained with the error back propagation algorithm. Feed-forward, back-propagation networks are based on the Delta Rule. It basically states that if the difference (delta) between the users desired output and the network's actual output is to be minimized, the weights must be continually modified. The result of the transfer function changes the delta error in the output layer. The error in the output layer has been adjusted or fixed, and therefore it can be used to change the input connection weights so that the desired output may be achieved. This network consists of 3 layers-Input layers, Hidden layers and Output layers as shown in the figure 5.

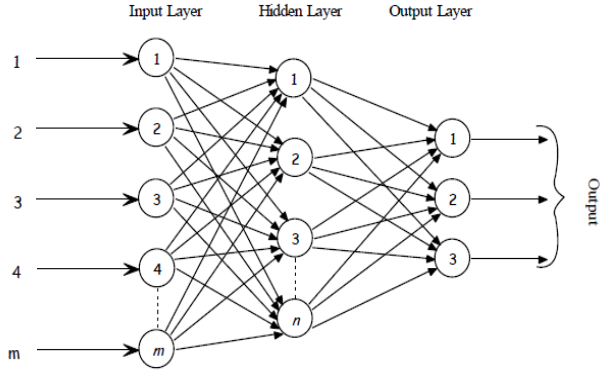


Figure 5: General Structure of a Neural

7. Results

In our work, we have used statistical features such as zoning, projection histogram, 8- directional zone density in combination with geometric features such as area, perimeter, major-axis length, minor-axis length, eccentricity, orientation of the character. In our research work, we have used Feed-forward MLP neural network architecture which is trained with the error back propagation algorithm. On the basis of these features, our neural network provides accuracy of 98%.The three layer Feed-forward MLP neural network is shown in figure 6.

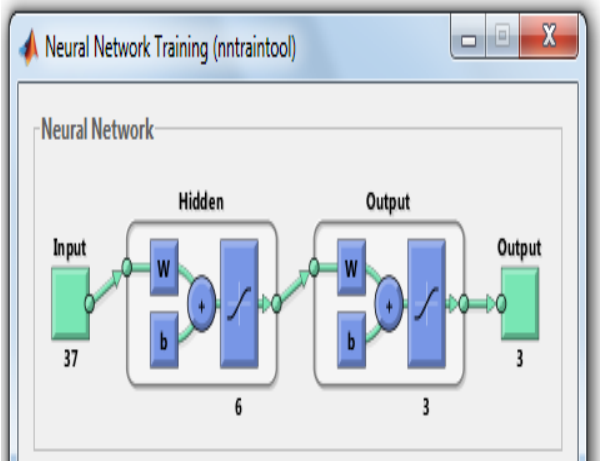
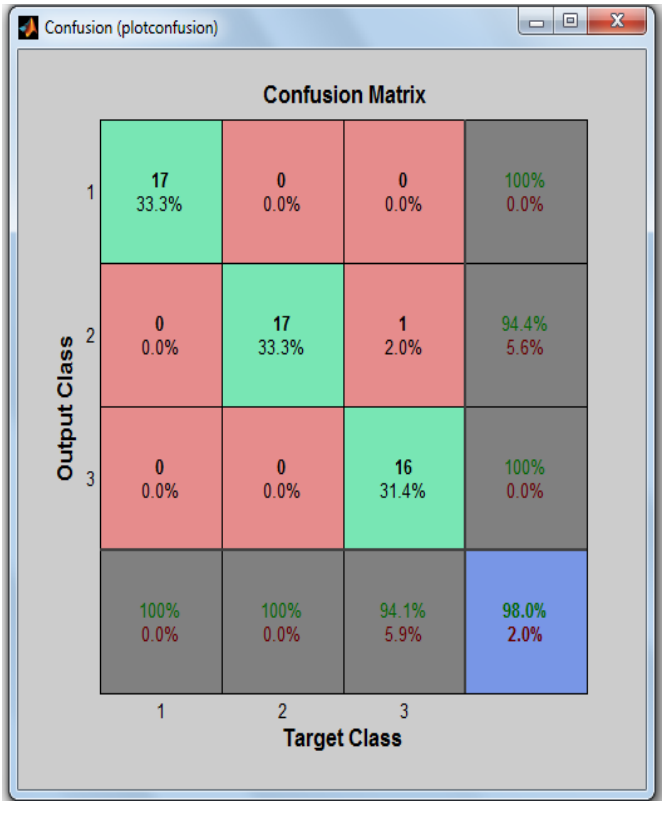


Figure 6: Three Layer Feed-forward MLP Classifier

7.1. Confusion Matrix

The performance of classifier is analyzed using confusion matrix which is also known as table of confusion. It displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The confusion matrix lists the correct classification against the predicted classification for each class. The confusion matrix shown below gives the classification results for 3 characters. From overall dataset 70% of dataset or a set of object used to design a classifier has been used for training phase.



The confusion matrix shows the percentages of correct and incorrect classifications. Correct classifications are the green squares on the matrices diagonal and the incorrect classifications are the red squares. The overall accuracy of the handwritten character recognition system is 98% as shown in the blue square.

8. Conclusion

In our thesis work we have recognized Handwritten Gurmukhi characters written by different writers. We have used three Gurmukhi characters written on plain paper with the help of different sketch colour pens. First the scanned image is pre-processed to get a cleaned image and the characters are isolated into individual characters. Statistical features like Zone density, projection histograms, and 8- directional zone density are used in combinations with geometric features like area, perimeter, eccentricity, major-axis length, minor-axis length, and orientation. The neural network is used as a classifier. The best recognition result obtained by using statistical features and geometric features through neural network is 98%.

9. References

[1] Kartar Singh Siddharth, Mahesh Jangid, Renu Dhir Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", *Proceedings of et al. /International Journal on Computer Science and Engineering*, Vol.3 No.6 June 2011.

[2] U. Pal, B.B. Chaudhury, "Indian Script Character Recognition: A Survey", *Pattern Recognition, Elsevier*, pp. 1887-1899, 2004.

[3] Vikas J Dungre et al., "A Review of Research on Devnagari Character Recognition", *International Journal of Computer Applications*, Volume-12, No.2, pp. 8-15, November 2010.

[4] Cao, L. J. Chong, W. K, "Feature extraction in support vector machine: a comparison of PCA, XPCA and ICA", *Proceedings of the 9th International Conference on Neural Information Processing*, Vol.2, PP.1001-1005, 2002.

[5] U. Pal, Wakabayashi, Kimura, "Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifiers", *10th International Conference on Document Analysis and Recognition*, pp. 1111-1115, 2009.

[6] Sandhya Arora, D. Bhattacharjee, M. Nasipuri, D.K. Basu, M. Kundu, "Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition", *Industrialand Information Systems, IEEE Region 10 Colloquium and the Third ICIIS*, pp. 1-6, December, 2008.

[7] L. Heutte, J. V. Moreau, T. Paquet, Y. Lecourttier and C. Olivier , "Combining structural and statistical features for the recognition of handwritten characters", *Proceedings of the 13th International Conference on Pattern Recognition*, Vol.2,pp.210-214,1996.

[8] G. Vamvakas, B. Gatos, S. Petridis, N. Stamatopoulos, "An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition", *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, Vol.2, pp.1073-1077, September 2007.

[9] Sarbajit Pal, Jhimli Mitra, Soumya Ghose, Paromita Banerjee, "A Projection Based Statistical Approach for Handwritten Character Recognition," *in Proceedings of International Conference on Computational Intelligence and Multimedia Applications*, Vol. 2, pp.404-408,2007.

[10] Wang Jin, Tang Bin-bin, Piao Chang-hao, Lei Gai-hui, "Statistical method-based evolvable character recognition system", *IEEE International Symposium on Industrial Electronics (ISIE)*, pp. 804-808, July 2009.

[11] G. S. Lehal, C. Singh, "A Gurmukhi Script Recognition System", *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 2, pp. 557-560, 2000.

[12] G. S. Lehal, C. Singh, "A Complete Machine printed Gurmukhi OCR", *Vivek*, 2006.

[13] G.S. Lehal, C. Singh, "Feature Extraction and Classification for OCR of Gurmukhi Script", *Vivek* Vol. 12, pp. 2-12, 1999.

[14] V. Goyal, G.S. Lehal, "Comparative Study of Hindi and Punjabi Language Scripts", *Nepalese Linguistics*, Vol. 23, pp. 67-82, 2008.

[15] Anuj Sharma, Rajesh Kumar, R. K. Sharma, "Online Handwritten Gurmukhi Character Recognition Using Elastic Matching", *Conference on Image and Signal Processing (CISP)*, Vol.2, pp.391-396, May 2008.