OCR for Handwritten Gurmukhi Script using Daubechies Wavelet Transforms

Pritpal Singh
M.E. (ECE)
Department of Electronics and Communication Engineering, UIET, Panjab University, Chandigarh, INDIA

Sumit Budhiraja
Assistant Professor,
Department of Electronics and Communication Engineering, UIET, Panjab University, Chandigarh, INDIA

ABSTRACT
This paper presents an OCR (Optical Character Recognition) for handwritten Gurmukhi characters. Gurmukhi script is used for Punjabi language, which is the world's 14th most widely spoken language. The scanned images of the handwritten characters are used as input to this OCR system and the output is the computer editable format. The proposed method uses different Daubechies Discrete Wavelet Transforms for extraction of features and back propagation for recognition task. In this paper, approximately four hundred samples of first five characters of Gurmukhi script are used to train the back propagation neural network and approximately one hundred samples are used for testing. The recognition rates achieved using different Daubechies Discrete Wavelet Transforms are compared. An average accuracy of 93.41% is achieved.

General Terms
Gurmukhi script, features, back propagation, recognition rates.

Keywords
Optical character recognition; handwritten Gurmukhi script; feature extraction; Daubechies Discrete Wavelet Transform.

1. INTRODUCTION
Optical character recognition is the process of converting the scanned images of handwritten, typewritten or printed text into machine or computer editable text. OCR system is very useful to reduce the processing time , reduce storage because an image required more memory, fast retrieval of the data and helpful in various scanning corrections. OCR systems have applications in post offices for automatic reading of postal code , automatic data entry, reading cheques and forms in banks and other institutions , automatic processing of invoice and receipt. High accuracy and speed are the main characteristics of a good OCR system. OCR aims at enabling computers to recognize optical symbols automatically and needs no human intervention. Handwritten character recognition is a major area in pattern recognition [1]. A hand written character recognition (HCR) system has to address issues such as infinite variety of character shapes, similarity between characters, distorted and broken characters [2].
Gurmukhi script is used to write the Punjabi language. Gurmukhi script has basic 35 characters, 10 vowels and modifiers, 6 additional modified consonants (derived from basic characters). Writing style for Gurmukhi is from left to right. There is no concept of upper or lowercase characters. The basic 35 characters of Gurmukhi are shown in Fig 1:

ੳ	ਅ	ੲ	ਸ	ਹ
ਕ	ਖ	ਗ	ਘ	ਙ
ਚ	ਛ	ਜ	ਝ	ਞ
ਟ	ਠ	ਡ	ਢ	ਣ
ਤ	ਥ	ਦ	ਧ	ਨ
ਪ	ਫ	ਬ	ਭ	ਮ
ਯ	ਰ	ਲ	ਵ	ੜ

Fig 1: Gurmukhi Alphabet

Gurmukhi script OCR has few problems which are responsible for reduction in recognition accuracy of Gurmukhi OCR. The variability in the writing style, for both between different writers and between separate examples from the same writer overtime, is the main problem. The second problem is similarity of some characters. Low quality of text images and presence of noise are some other problems responsible for less recognition rate [3].
The organization of this paper is: Section 2 includes the previous work, section 3 includes proposed Gurmukhi OCR system, and section 4 includes the experimental results and section five includes the conclusion.

2. PREVIOUS RELATED WORK
A lot of research has been done in the field pattern recognition. Raju G. [2] has proposed an OCR system for Malayalam characters. Feature extraction has been done using different wavelet filters and MLP network has been used for classification. An average accuracy of 81.3% has been obtained. In case of handwritten Gurmukhi OCR systems not much work has been reported. For handwritten Gurmukhi character recognition three approaches have been reported. First one is proposed by Puneet Jhajj et al. [3], second one by Ubika Jain et al. [4] and third one by Kartar Singh Siddharth et al. [5]. Puneet Jhajj et al. has first resized the original image to 48*48 pixels normalized image and created 64 (8*8) zones.

Then the zoning densities of these zones have been taken as features. The SVM and K-NN classifiers have been used and compared the results. It has been observed that 72.83% was the highest accuracy with SVM (Support Vector Machine) with RBF kernel. Ubeeka Jain et al. has been created horizontal and vertical profiles for each chracter, stored height and width of each character and used neocognitron artificial neural network for feature extraction and classification. The accuracy of 92.78% has been obtained. Kartar Singh Siddharth et al. have been used statistical features e.g. zonal density, projection histograms (horizontal, vertical and both diagonal), distance profiles (from left, right, top and bottom sides). Also background directional distribution (BDD) features have been used. The images have been normalized to 32*32 sizes. SVM, K-NN and PNN classifiers have been used for classification. The highest accuracy of 95.04% has been obtained as 5-fold cross validation of whole database using zonal density and background distribution features in combination with SVM classifier used with RBF kernel. G S Lehal et al. [6] has given an OCR system for printed Gurmukhi script and obtained an accuracy of 96.6%. Kartar Singh Siddharth et al. [7] has proposed an OCR for handwritten Gurmukhi numerals using distance Profiles and Background Directional Distribution (BDD) as features .The SVM classifier with RBF (Radial Basis Function) kernel has been used for classification. The maximum recognition accuracy of 99.2% has been obtained.

3. THE PROPOSED GURMUKHI OCR SYSTEM

The process of optical character recognition of handwritten Gurmukhi script has mainly following stages [8]:

1. Image acquisition
2. Pre-processing
3. Feature extraction
4. Classification using neural network
5. Output result

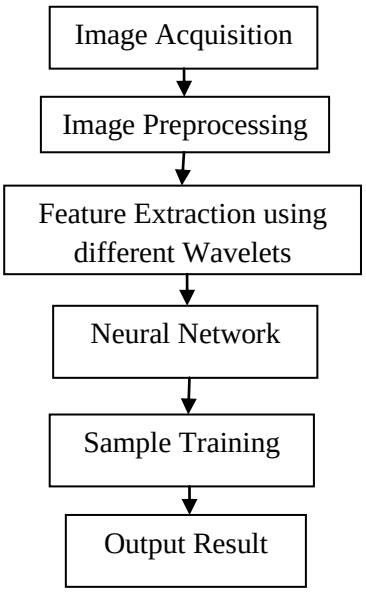

Fig 2: Block Diagram of Handwritten Gurmukhi OCR

3.1 Image Acquisition

The handwritten Gurmukhi character samples are taken from different writers at different times. In the proposed work the samples for first five characters of the Gurmukhi script has been taken. These samples are taken by scanning the handwritten Gurmukhi characters at 400 dpi. Total 392 samples are taken. Some of these samples are shown in Fig 3:

Fig 3: Few Samples of Gurmukhi Characters

3.2 Pre-Processing

In the pre-processing stage the OCR has given a raw scanned color image then following operations are applied on it:

1. Raw colour image is converted in to grey image.
2. Median filtering is applied to the image to remove noise.
3. The image then converted in to the binary image using thresholding.
4. The binary character image is normalized to 32*64.

3.3 Feature Extraction

Wavelets are localized basis functions which are translated and dilated versions of some fixed mother wavelet. They are used for representing signals, such as images, in a hierarchy of increasing resolutions. The decomposition of the image into different frequency bands is obtained by successive low-pass and high-pass filtering of the signal and down-sampling the coefficients after each filtering. In this work we have used various Daubechies discrete wavelet transforms [2]. The names of the Daubechies family wavelets are written ‘dbN’, where N is the order, and db the "surname" of the wavelet. The feature extraction is done by using the following algorithm:

For each pre-processed image repeat steps a to f

a. Number of black pixels along each row of the binarized image has been counted to form a 32 sized vector.
b. The 1D Daubechies wavelet transform on row count vector (two levels) has been applied.
c. Then the approximation (low frequency or average) coefficients have been directly taken as feature values.
d. Number of black pixels along each column has been counted to form a 64 sized vector.
e. The 1D Daubechies wavelet transform on column count vector (three levels) has been applied.
f. Then the approximation coefficients have been directly taken as next feature values.

Above explained steps are repeated with different Daubechies wavelet filters viz. db1, db2, db3, db4, db5. After the various Daubechies wavelet filters (db1, db2, db3, db4 and db5) are applied, the feature vectors lengths are summarized in the Table 1:

Table 1. Length of feature vectors

Daubechies Wavelet	Length of feature vector
db1	16
db2	20
db3	23
db4	27
db5	29

3.4 Classification

The back propagation neural network is used as classifier in this work. Back Propagation Neural Network (BPNN) is a Multilayer Neural Network which is based upon back propagation algorithm for training [9]. This neural network is based upon extended gradient-descent based Delta learning rule, commonly known as Back Propagation rule. In this network, error signal between desired output and actual output is being propagated in backward direction from output to hidden layer and then to input layer in order to train the network. Each neuron in one layer has directed connections to the neurons of the subsequent layer. The network structure we used includes input nodes equal to number of feature vector elements, one hidden layer with 30 nodes and 5 output nodes. Back Propagation Neural Network is explained as following: The training data is fed into the input layer. It is propagated to both the hidden layer and the output layer. In this stage, each node in the input layer, hidden layer and output layer calculates and adjusts the appropriate weight between nodes and generate output value of the resulting sum. The actual output values will be compared with the target output values. The error between these outputs will be calculated and propagated back to hidden layer in order to update the weight of each node again. This is called backward pass or learning. The network will iterate over many cycles until the error is acceptable. After the training phase is done, the trained network is ready to use for any new input data. During the testing phase, there is no learning or modifying of the weight matrices.

The testing input is fed into the input layer, and the feed forward network will generate results based on its knowledge from trained network.

3.5 Sample Training and Testing

In the experiment, the wavelet coefficients are given as input to the BPN network. For training 310 samples were used. Remaining 82 samples were used for testing. The testing and training samples are selected at random forming two data sets.

4. EXPERIMENTAL RESULTS

In the work two experiments have been performed. In both experiments various Daubechies Discrete Wavelet Transforms e.g. db1, db2, db3, db4 and db5 have been used to extract the features and to derive a feature vectors. The first experiment has used the feature vector in which aspect ratio is not considered. But in case of the second experiment the aspect ratio is considered as the additional feature. The outcomes of the both experiments have been summarized in Table 2 and compared in Fig 3. The average recognition accuracy in first experiment is 92.20% and in case of second experiment it is 93.41%.

Table 2. Comparision of Recognition Accuracy using Differrent Daubechies Wavelets

Daubechies Wavelet	Recognition Accuracy without Taking Aspect Ratio in Feature Vector	Recognition Accuracy Taking Aspect Ratio in Feature Vector
db1	91.46	93.90
db2	90.24	95.12
db3	90.24	93.90
db4	93.90	90.24
db5	95.12	93.90
Average	92.20	93.41

.

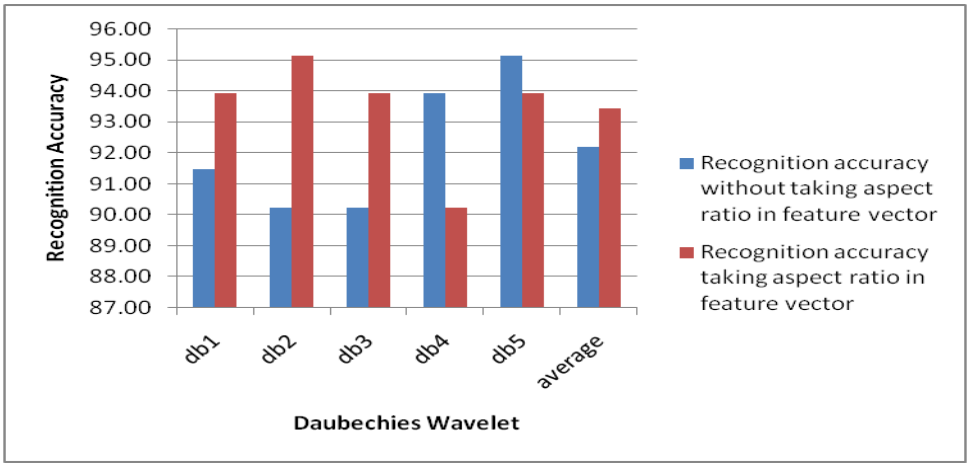

Fig.3 Recognition Accuracy using Daubechies wavelets

5. CONCLUSION

This OCR system is relatively faster as the feature vector has lesser elements as compared to other OCR systems developed so far. The result obtained is comparable with similar works reported earlier. Also this work shows that if aspect ratio is added as additional feature, then the recognition rate of the handwritten Gurmukhi OCR system may be further improved. As the size and quality of database is major factor influencing HCR systems, so relatively large database should be used in the future work. By adding some more features we can enhance the recognition accuracy. We have used only first five characters for training and testing purposes. It can be extended for all Gurmukhi characters. Also we can use different wavelet filters to design more accurate OCR system and the same will be done in near future so that an OCR system with nearly 100% recognition accuracy can be achieved.

6. REFERENCES

[1] O. D. Trier, A. K. Jain and T. Text, "Feature Extraction Methods for Character Recognition- A Survey", Pattern Recognition, Vol. 29, No. 4, pp. 641-662, 1996.

[2] Raju G., "Wavelet Transform and Projection Profiles in Handwritten Character Recognition – A Performance Analysis", IEEE, pp. 309-314, 2008.

[3] Puneet Jhajj, D. Sharma, "Recognition of Isolated Handwritten Characters in Gurmukhi Script", International Journal of Computer Applications, Vol. 4, No. 8, 2010.

[4] Ubeeka Jain, D. Sharma, "Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron", International Journal of Computer Applications, Vol. 10, No. 8, 2010.

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

[6] G S Lehal and Chandan Singh, "A Gurmukhi Script Recognition System", Proceedings of the International Conference on Pattern Recognition (ICPR'00), 2000.

[7] Kartar Singh Siddharth, Renu Dhir, Rajneesh Rani," Handwritten Gurmukhi Numeral Recognition using Different Feature Sets", International Journal of Computer Applications, Volume 28, No.2, August 2011.

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

[9] M Abdul Rahiman, M S Rajasree, "OCR for Malayalam Script Using Neural Networks", IEEE, 2009.