Gurmukhi Printed Character Recognition using Hierarchical Centroid Method and SVM

Sandeep Kaur
Punjabi University
Regional Centre for IT & Mgmt.
Mohali, India

Rekha Bhatia
Punjabi University
Regional Centre for IT & Mgmt.
Mohali, India

ABSTRACT

In this paper the system for the recognition of printed Gurmukhi character is proposed. Hierarchical centroid method is used for extracting the feature from images of printed characters. The main advantage of using this method is that it gives size invariant feature vector and therefore can play important role for manuscript recognition. The dataset used in this study consists of 29 different font styles of the printed characters. The classification is done by using Support Vector Machine. The performance of the classifier is determined by measuring accuracy using 10-fold cross validation procedure. The highest accuracy obtained on SVM is 97.87% with the combination of *nu*-SVC type and RBF kernel.

Keywords

Character Recognition, Support Vector Machine, Printed Gurmukhi.

1. INTRODUCTION

Multiple languages are used among different people in India which mainly depends on their geographical locations. Gurmukhi is one of the languages which is mostly used in northern region of India. This script was devised by first Sikh Guru, Sri Guru Nanak Dev ji in 16th century and popularized by second Sikh Guru, Sri Guru Angad Dev ji. Gurmukhi word is derived from two Sanskrit words Guru and Mukha which means the sayings coming from teacher's mouth. The Gurmukhi script consist of total thirty characters as shown in Fig. 1 and 10 numerals as shown in Fig. 2.

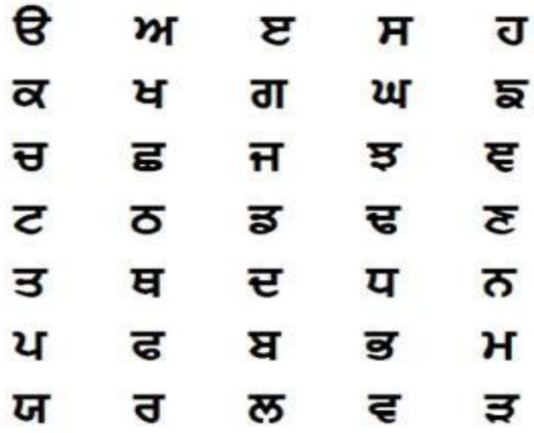


Fig.1 Thirty five Gurmukhi characters

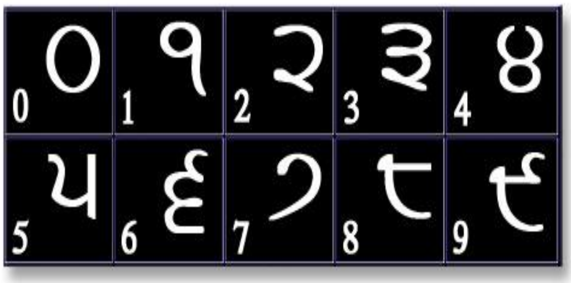


Fig. 2 Gurmukhi Numerals

Optical Character recognition (OCR) is a system which aims at transforming a document into computer readable form [1]. The document can be printed or handwritten form. OCR is further divided into subfields: - Online and Offline Character recognition and Offline character recognition further divided into two parts: - machine printed and handwritten character recognition [2]. Handwritten character recognition has many problems like different writing styles, variation in pen-tip, skewness in writing etc. OCR plays an important role in improvement of interface between computer and man [3]. Character recognition is one of fascinating area of artificial intelligence and pattern recognition.

Punjabi language, in which Gurmukhi script used, is 14th most widely spoken language in world. So designing an OCR for the recognition of the Punjabi language is an open and promising research topic. It may also help in the digitization of the old and valuable Punjabi literature which will further help in its conservation. The importance of OCR in transforming the documents into machine readable form, have led to great advancement in the field which are limited to the other languages like English, Arabic and Chinese languages [4-6].

In Punjabi, most of the work is focused on online or offline handwritten character recognition and very little work is done for printed character recognition. So, in this paper an algorithm has been proposed for the recognition of Gurmukhi printed character recognition using a large set of different types of fonts. Rest of the paper is structured as follows: work done so far in the field of OCR is reviewed in section II, details of the materials and methodology used in the proposed study is given in section III, in section IV results are discussed and the paper is concluded in section V.

2. RELATED WORK

In this section, work related to Gurmukhi character recognition done so far is briefly reviewed. During past 15 years, many researchers have tried to improve the OCR system for Gurmukhi language. Majority of the research is devoted to online and offline handwritten character/script recognition. Some of these studies are briefly reviewed as follows:

The zoning mechanism is used for the feature extraction which is used to train support vector machine (SVM). This paper present a technique for the recognition Gurmukhi characters which works in two phases: In first phase image was divided into different zones and in second phase diagonal features were computed for each zone of the character. Zones were also considered in both vertical and horizontal directions. 7000 images were used for the training and testing of the proposed method and total 120 features were extracted from each image. This method gives 95.34 % and 95.74% using diagonal features and SVM with 5 fold and 10 fold cross validation [1].

In [7] gradient information of an image is used as feature for the classification Gurmukhi characters and numerals. The two methods were proposed and both works on the blocks of an images. The information from each block was concatenated to form a single feature vector of dimensionality 2000. The method was tested on two different dataset each having 7000 and 2000 binary images. The recognition rate of Gurmukhi characters and numerals were found out to be 97.38% and 99.65%. The work was also extended to find out collective accuracy of methodology for both characters and numerals.

The main problem in Gurmukhi script recognition in case of machine printed characters faces major problems like unique characteristics of the script like connectivity of characters on the headline, a large number of similar characters and two or more characters in a word having intersecting minimum bounding rectangles. The segmentation algorithm break sentences into characters and then recognition take place. A hybrid classification is used classification is done by combining decision tree and nearest neighbor classifiers. The recognition rate and processing speed are found out to be 96.6% and 175 characters per second [3].

Munish et. al presented a work on recognition of Gurmukhi characters by employing k-NN classifiers. First Skelton of each character is prepared in order to extract its feature information. Diagonal and transition features were calculated on the bitmap images of the characters and then Euclidean distance between the testing point and reference point is calculated in order to find k-nearest neighbors. The dataset used in the study consist of 3500 images of Gurmukhi characters from nearly 100 different writers. The accuracy of 94.12% was achieved by this system.

The combination of horizontal and vertical projection based features along with K- nearest neighbor (k-NN) and Support Vector Machine (SVM) is used for the classification of Gurmukhi script. Total 3500 samples were collected from the 10 different people where each belongs to different age group. Three strategies were formed by dividing the whole dataset into the different propositions which were further used as the training and testing. Two features were extracted from each images and tested on the two classifiers SVM and k-NN. Linear and polynomial kernel functions were applied in case of SVM and k-NN were tested with 1, 3, 5 and 7 as different values for k [9].

The dataset consist of 7000 images of same size were collected from 20 writers of different age groups. In preprocessing stage, an image was converted to binary by using Otsu's method. The median filter and morphological operation was also applied on raw image as part of preprocessing. Iterative approach was used for the segmentation of the characters. Gabor based features named GABM and GABN were used to train SVM for the classification purpose. Fivefold cross validation method were used to evaluate and validate the methodology. The methodology gives 94.29% accuracy in Gurmukhi script classification with Gabor based features of dimensionality 200 [10].

Three different features and classifiers were tested for the recognition of Gurmukhi characters from documents. In first, 128 features comprised of distance profiles. Second feature set was different histogram projection of size 190 and in last zonal density and background directional distribution forming 144 features. These were applied on three classifiers names SVM, PNN and k-NN. The SVM was tested with radial basis kernel function. The different parameters were varied of each classifier to get best results. SVM gives best results with second parameters whereas with PNN and k-NN, third feature set gives best results. The 5 fold cross validation scheme is used for validation of results [11].

Lehal et.al worked on one another technique for the Gurmukhi script recognition. In this work, hybrid classification scheme was employed using binary decision tree and nearest neighbor classifier. The three stage methodology was developed. In first stage, characters were grouped into three sets on the basis of their position. In second, characters falling in middle zone set were further divided into smaller sets. In last stage, special features were employed to distinguish the characters with the help of nearest neighbor classifier.

As we see, most of the work is done for handwritten character recognition and very little work is done to provide an OCR system for printed character recognition. Gurmukhi language has gone through vast variations in terms of font styles. The variation is mainly due to different background and regions where it is spoken. So it is a challenging task to collect such a large data and then design an OCR system. In this paper an attempt has been made to provide an OCR system for printed character recognition. Details of the methodology adopted are discussed in next section.

3. MATERIALS AND METHOD

3.1 Dataset

The dataset created in the proposed study consists of 29 different font styles. Each character of every style is printed in three font sizes of 18, 24 and 30. There are total 35 characters in Punjabi and 10 numerals. So a dataset of 3915 images is created in the proposed study. Size of images varies from 20*21 pixels to 40*41 pixels with a bit depth of 1. The sample images of font styles used in our dataset are shown in Figure 3.

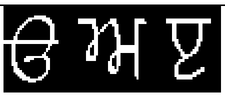
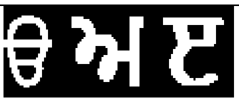
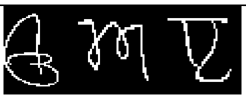
Style name	**Fonts sample**
Adami	
Rangdar	
Adhiapak	

Fig. 3 Sample Font Styles of our Dataset

3.2 Preprocessing

Images of different types of font styles used in Punjabi language were acquired using a scanner. Firstly the characters are typed and their printout is taken. Then, each character of each font style is scanned one by one. The digitized images obtained after scanning are stored in a PNG format. Since digitized images are in gray tone fashion, they are converted to binary images using simple thresholding method based on histogram.

3.3 Feature Extraction

Armon [12] proposed a hierarchical centroid method for features extraction from printed character images of Hebrew language. The method is based on recursive subdivisions of input binary image by measuring centroids at each division and outputs a fixed length features vector. Further, feature vector are normalized according the size of the input image. Therefore the size of images does not affect the final feature set. The method is a two-step procedure in which first includes feature extraction with respect to x-coordinate while second step extract feature with respect to y- coordinate. In the first step, for given input binary image, the first order moment (center of mass) is calculated for the x-coordinate and image is divided into two sub-images by x-coordinate. Then process is repeated recursively for newly generated sub-images by transposing them and features are computed by considering coordinate values relative to the original image. The length of the feature vector is give as: 2^d-1, where d is the parametric value which denotes the recursive depth or divisions. But it produces subdivisions along one axis twice as compare to other. Therefore, in order to balance output vector, features are also calculated along transposed image. The final feature vector is the concatenation of both vectors of original and transposed image. The length of final vector will be 2*(2^d-1). Figure 4 shows the divisions of input image at different depth levels of x-coordinate and y-coordinates.

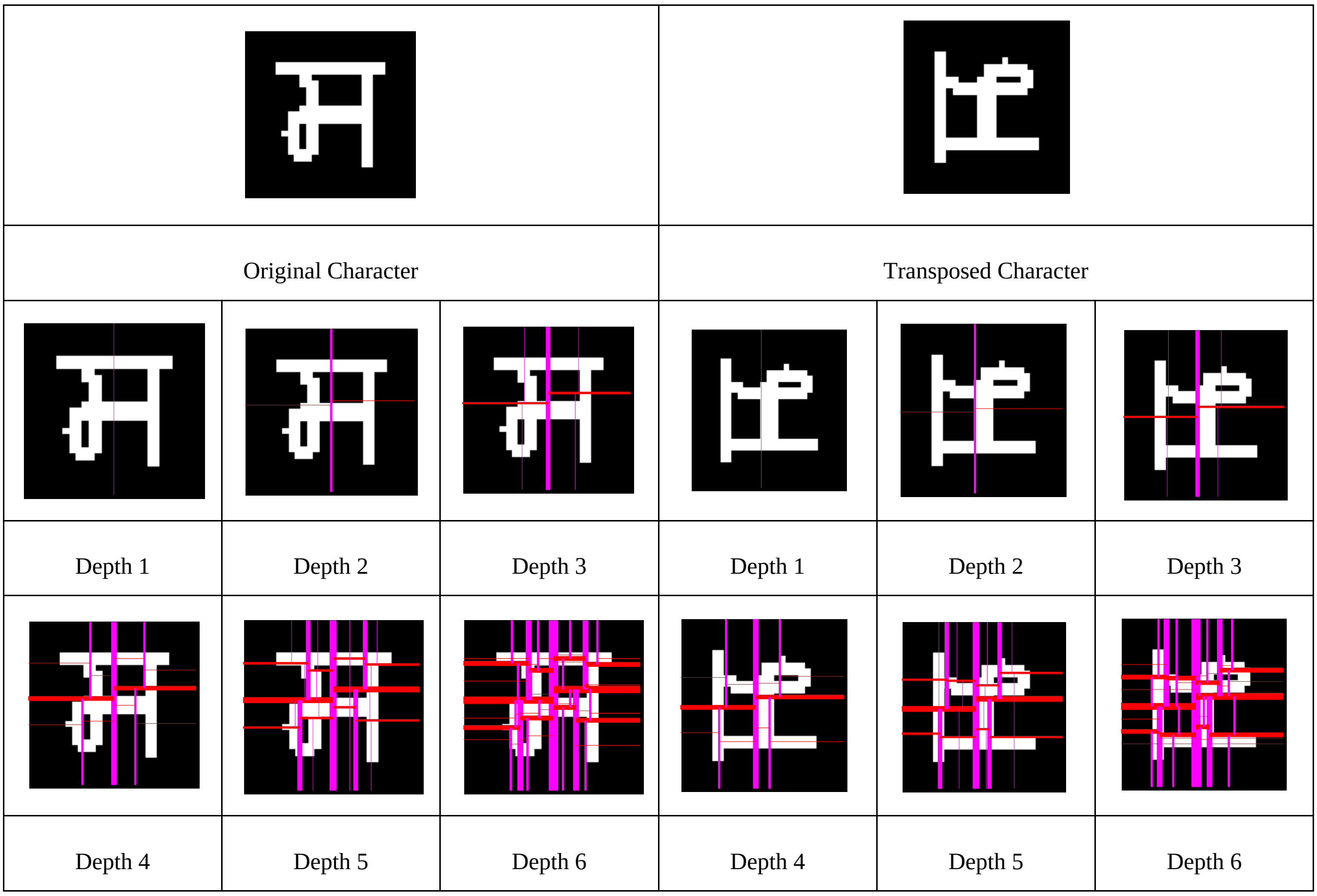


Fig. 4 Divisions of s at different depth levels along x and y coordinates

3.4 Classification

Support Vector Machine (SVM) [13] is a supervised learning technique used for classification or regression. SVM was initially designed for binary classification, but now extended for solving multi-class classification problems by decomposing problem into multiple binary classes. The discrimination (separation) between classes is done by constructing a hyper plane [15]. The main goal is to find an optimal hyper plane which expects to generalize the comparison to the others. The optimal hyper plane is one which maximizes a measure of the “margin” between such classes. The unknown data sample is then classified by the SVM according to the decision boundaries defined by these hyper planes. Hyper planes with maximum appropriate margins can be constructed using to different types of kernel functions in SVM classifier. The commonly used kernel functions are: Gaussian RBF (Radial Basis Function) kernel, Linear Kernel, Polynomial kernel and Sigmoid (hyperbolic tangent) kernel. The effectiveness of data classification depends upon kernel used, kernel parameters and soft margin or penalty parameter.

4. RESULTS AND DISCUSSION

The classification is done by dividing the Gurumukhi character dataset into two sets, training and testing. The efficiency of proposed classifier is obtained by using *n*-fold cross validation procedure. In this approach, initially the dataset of size P is divided into n disjoint groups known as test sets. Then the classifier is trained on rest of data by removing one test set. The process is repeated n-times and the average accuracies are measured. Since all the samples are used for testing and training, it has precedence over individual sets of training and testing. The value of n is set to 10 for experiments which mean 90% of dataset is used for training while rest 10% is used for testing for each time. The efficiency of the classifier is determined by considering performance measure, Accuracy.

Accuracy: It is the percentage of correctly classified characters from total number of characters in the test set. It is given as

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

The proposed character recognition system needs to classify the Gurumukhi characters into 45 different classes of characters including 35 alphabets and ten numeric characters. A multiclass SVM classifier tool LIBSVM (A Library for Support Vector Machines) has been used in this study.

The performance of classification on LIBSVM tool is measured by considering two types of SVM namely, C-SVC and nu-SVC and three types of kernels namely, Gaussian RBF (Radial Basis Function) kernel, Linear Kernel and Polynomial kernel.

Table 1: Classification results using different type of SVM and its Kernels

Sr. No.	SVM type	Kernel	Accuracy(%age)
1.	*nu*-SVC	RBF	97.87
2.	*nu*-SVC	Linear	96.85
3.	*nu*-SVC	Polynomial	96.59
4.	C-SVC	RBF	97.72
5.	C-SVC	Linear	97.18
6.	C-SVC	Polynomial	96.59

On observing the above table, it can be seen that the best classification results are obtained from RBF kernel, least accuracy from polynomial kernel and average results are obtained from linear kernel of both type of SVM. However, the results of C-SVC type of SVM are better than *nu*-SVC type of SVM. Further, it is observed that highest accuracy of 97.87% has been achieved from RBF kernel of *nu*-SVC type of SVM.

5. CONCLUSION

In this paper printed character recognition system for Gurumukhi is proposed. Hierarchical centroid method is used for feature extraction which does not require fixed size as compared to existing work. The obtained feature vector is then learned for classification using Support Vector Machine on the dataset created using different type of font sizes and types.

The performance is analyzed using different combinations of SVM type and its kernels on 10-fold cross validation procedure. The highest accuracy obtained on SVM is 97.87% with the combination of nu-SVC type and RBF kernel. In future, this work can be extended on handwritten Gurumukhi character recognition due to its high performance.

6. REFERENCES

[1] A. Jindal, R. Dhir, and R. Rani, “Diagonal Features and SVM Classifier for Handwritten Gurumukhi Character Recognition,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 2, no. 5, pp. 505–508, 2012.

[2] M. Jangid, R. Dhir, R. Rani, and K. Singh, “SVM Classifier for Recognition of Handwritten Devanagari Numeral,” in Image Information Processing (ICIIP), 2011, pp. 1–5.

[3] G. S. Lehal and C. Singh, “A Gurmukhi Script Recognition System,” in 15th International Conference on Pattern Recognition, 2000, no. 2, pp. 557–560.

[4] J. Mantas, “An overview of character recognition methodologies,” Pattern Recognit., vol. 19, no. 6, pp. 425–430, 1986.

[5] V. K. Govindan and A. P. Shivaprasad, “Character recognition - A survey,” Pattern Recognit., vol. 23, pp. 671–683, 1990.

[6] B. Al-Badr and S. a. Mahmoud, “Survey and bibliography of Arabic optical text recognition,” Signal Processing, vol. 41, no. 1, pp. 49–77, 1995.

[7] A. Aggarwal, K. Singh, and K. Singh, “Use of gradient technique for extracting features from handwritten gurmukhi characters and numerals,” Procedia Comput. Sci., vol. 46, pp. 1716–1723, 2015.

[8] M. Kumar, M. K. Jindal, and R. K. Sharma, “k -Nearest Neighbor Based Offline Handwritten Gurmukhi Character Recognition.,” in International Conference on Image Information Processing (ICIIP), 2011.

[9] M. K. Mahto, K. Bhatia, and R. K. Sharma, “Combined Horizontal and Vertical Projection Feature Extraction Technique for Gurmukhi Handwritten Character Recognition,” in International Conference on Advances in Computer Engineering and Applications (ICACEA), 2015.

[10] S. Singh, A. Aggarwal, and R. Dhir, “Use of Gabor Filters for Recognition of Handwritten Gurmukhi Character,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 2, no. 5, pp. 234–240, 2012.

[11] K. S. Siddharth, R. Dhir, R. Rani, M. Jangid, and K. Singh, “Comparative Recognition of Handwritten Gurmukhi Numerals Using Different Feature Sets and Classifiers,” in Proceedings of International Conference on Image Information Processing (ICIIP 2011), 2011.

[12] Armon, S., “Handwriting recognition and fast retrieval for Hebrew historical manuscripts”, Master Thesis, 2011.

[13] C. Cortes and V. Vapnik, “Support vector machine,” Machine learning, vol. 20, pp. 273-297, 1995

[14] D. Singh, B. Singh, A new morphology based approach for blood vessel segmentation in retinal images, in 2014 Annual IEEE India Conference (INDICON), 2014, pp. 1-6