Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Diagonal Features and SVM Classifier for Handwritten Gurumukhi Character Recognition

Anita Jindal, Renu Dhir, Rajneesh Rani
Department of Computer Science &Engineering,
NIT Jalandhar
India

Abstract— **In this paper an isolated handwritten Gurumukhi character recognition system using SVM classifier is presented. Diagonal features are used for extracting the features of a character. Diagonal feature computation consists of two steps1) character image is divided into different zones2) Diagonal Features are computed for each zone of the character. Features are also computed from each image by considering zones in horizontal and vertical directions. Thus total 120 features are computed for each character. We have used in all 7000 images of Gurumukhi characters for the purpose of training and testing. The proposed system achieves a maximum recognition accuracy of 95.34% and 95.74% using diagonal features and SVM classifier with 5-fold and 10-fold cross validation respectively. The diagonal orientation for feature extraction is identified to be the most suitable method as it yields higher recognition accuracy.**

Keywords— **Handwritten Gurumukhi Character Recognition, Diagonal Feature, SVM classifier with RBF kernel.**

I. INTRODUCTION

Optical Character Recognition (OCR) systems aim at transforming large amount of documents, either printed or handwritten, into machine encoded text. Nowadays, although recognition of printed isolated characters is performed with high accuracy, recognition of handwritten characters still remains an open problem in the research area. A widely used approach in isolated character recognition is to follow a two step schema: a) represent the character as a vector of features and b) classify the feature vector into classes [1].

In general, handwriting recognition is classified into two types as off-line and on-line handwriting recognition methods. Off-line handwriting recognition involves automatic conversion of text into an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles. But, in the on-line system, the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available. The on-line methods have been shown to be superior to their off-line counterparts in recognizing handwritten characters due to the temporal information available with the former [2].

Any OCR problem consists of five stages such as Image preprocessing, segmentation, feature extraction, classification and recognition and post processing**.** All these steps are practiced in our approach. In our scheme we have implemented Diagonal feature extraction technique. In this methodology image is divided into 100 equal zones each of size 10*10 pixels. Features are extracted from pixels of each zone by moving along its diagonals.

In the following sections literature survey, pre-processing, and proposed methodology including feature extraction, result analysis and conclusion are discussed.

II. RELATED WORK

Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. A review of different feature extraction methods and the problem of choosing the appropriate feature extraction method for a given application is discussed in[3]-[5].these papers provide general discussion of Most oftenly used texture feature extraction methods in various research areas including biometric, medical research and printed character recognition. A study of methods in this paper is useful for appropriate use of existing methods as well as invention of new methods. In [6], U. Pal and R Jayadevan have provided a survey on all feature extraction techniques as well as training, classification and matching techniques used for recognition of machine printed and handwritten Devanagri state of the art from 1970s.In the literature survey[7] work done on different Indian language scripts is presented. It is found that a lot of work has been done in recognition of Devanagri and Bangla script characters, the two most popular languages in India, In case of Gurumukhi language work has been done by G.S. Lehal and Chandan Singh in[11]-[13] for printed Gurumukhi Script .But for Handwritten Gurumukhi Script work done is very less. In 2011, kartar Singh Siddharth and Mahesh Jangid have used different feature extraction techniques for Handwritten Gurumukhi character recognition such as zoning density, Projection Histograms ,distance Profiles and Background Directional Distribution(BDD) features. Highest accuracy obtained is 95.04% using Zoning Density and Background Directional Distribution features(BDD)[10].In zoning density image is divided into different zones. From each zone number of foreground pixels are calculated, then count of foreground

pixels is divided by the total number of pixels in that zone to obtain zoning density of each zone. For Background Directional Distribution features they have considered the directional distribution of neighboring background pixels to foreground pixels. They have computed 8 directional distribution features. To calculate directional distribution values of background pixels for each foreground pixel, masks for each directional values are used. In [8] Diagonal features are computed from each zone and for recognition purpose neural network classifier is used. Same technique has been implemented in [9] by Munish Kumar, M.K. Jindal and R.K. Sharma for Gurumukhi Characters using K-NN classifier and accuracy obtained is 94.12.In our recognition approach we have implemented the same technique on dataset of 7000 samples using SVM classifier with its Radial Basis Function (RBF) kernel with 5-fold and 10-fold cross-validation.

III. RECOGNITION SCHEME

Our Dataset consists of total of 7000 characters of Gurumukhi Script consisting of 200 samples of each of 35 characters. Each image consists of 10 samples of each of 35 characters. Thus total 20 images are used. Each image goes through a series of steps such as image acquisition, preprocessing, segmentation, feature extraction, classification and recognition and post processing. In post processing step we bind up our work to create complete digitized document after recognition process.

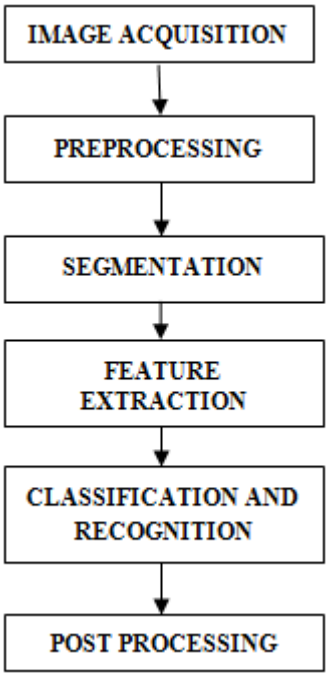

Figure 1.Basic Process of OCR

A. Image Acquisition

In Image acquisition, the recognition system acquires a scanned image as an input image. The image should have a specific format such as JPEG, BMT etc. This image is acquired through a scanner, digital camera or any other suitable digital input device.

B. Pre-processing

The pre-processing is a series of operations performed on the scanned input image. It essentially enhances the image rendering it suitable for segmentation. The following steps are used to create dataset. Entire work has been done in Matlab.

- Handwritten sample is scanned in RGB format.
- RGB image is converted into grayscale image.
- Grayscale image is converted into binary image by using a suitable threshold value by Otsu's method.
- Other preprocessing techniques like median filtration, dilation, some morphological operations are applied to join unconnected pixels, to remove isolated pixels, to set neighbor pixel values in majority and to remove the spur pixels.

C. Segmentation

In the segmentation stage, an image of sequence of characters is decomposed into sub-images of individual character [14]. In the proposed system, the pre-processed input image is segmented into isolated characters by assigning a number to each character using a labeling process. This labeling provides information about number of characters in the image. Each individual character is uniformly resized into 100x100 pixels for extracting its features.

D. Feature Extraction

Diagonal features are very important features in order to achieve higher recognition accuracy and reducing misclassification. These features are extracted from the pixels of each zone by moving along its diagonals as shown in Fig 2. Following algorithm describes the computation of Diagonal Features for each character image of size 100*100 pixels having 10*10 zones and thus each zone having 10*10 pixel size. Each of these zones are having 19 diagonals. The number of foreground pixels along each diagonal are summed up to get 19 features from each zone, then these features for each zone are averaged to extract a single feature from each zone.

Steps for computation of Diagonal Features

Step I: Divide the input image into *n* (=100) number of zones, each of size 10×10 pixels.

Step II: The features are extracted from the pixels of each zone by moving along its diagonals.

Step III: Each zone has 19 diagonals; foreground pixels present along each diagonal is summed up in order to get a single sub feature.

Step IV: These 19 sub-features values are averaged to form a single value and placed in corresponding zone as its feature.

Step V: Corresponding to the zones whose diagonals do not have a foreground pixel, the feature value is taken as zero.

Using this algorithm, we will obtain *100* features corresponding to every zone.

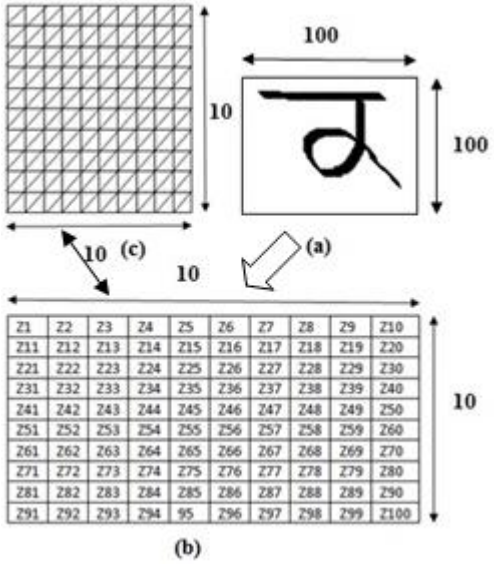

Fig 2.(a)Normalized Character Image(b)Image divided into 100 zones(c)Diagonal Feature Extraction in a zone each of size 10*10 pixels

IV. CLASSIFICATION

Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. The standard SVM classifier takes the set of input data and predicts to classify them in one of the only two distinct classes. SVM classifier is trained by a given set of training data and a model is prepared to classify test data based upon this model. For multiclass classification problem, we decompose multiclass problem into multiple binary class problems, and we design suitable combined multiple binary SVM classifiers. Our problem also needs to classify the characters into 35 different classes of Gurumukhi characters. We obtained such multiclass SVM classifier tool LIBSVM available at [18]. A practical guide for SVM and its implementation is available at [19]. According to how all the samples can be classified in different classes with appropriate margin, different types of kernel in SVM classifier are used. Commonly used kernels are: *Linear kernel*, *Polynomial kernel*, Gaussian *Radial Basis Function* (RBF) and *Sigmoid* (*hyperbolic tangent*).The effectiveness of SVM depends on kernel used, kernel parameters and soft margin or penalty parameter C.

The common choice is RBF kernel, which has a single parameter *gamma* (g or γ). We also have selected RBF kernel for our experiment. Best combination of C and γ for optimal result is obtained by grid search by exponentially growing sequence of C and γ and each combination is cross validated and parameters in combination giving highest cross validation accuracy are selected as optimal.

In *V*-fold cross validation we first divide the training set into V equal subsets. Then one subset is used to test by classifier trained by other remaining V-1 subsets. By cross validation each sample of train data is predicted and it gives the percentage of correctly recognized dataset.

V. EXPERIMENTS AND RESULTS

We have used 7000 samples of isolated handwritten Gurumukhi characters in our experiment. These samples are written by 20 different writers, each contributing to write 10 samples of each character out of 35 characters.

Initially a random sample out of total dataset was taken to train the SVM classifier and we optimized the parameters C and g (or γ). Further by training the system by whole dataset the optimization of these parameters was refined. We finally selected the optimal parameters combination giving highest cross validation accuracy. Thus selected optimal parameters are- C = 8 and g = 0.004. With these parameters 5-fold cross validation accuracy obtained is 95.34% while 10-fold cross validation accuracy obtained is 95.74%. The samples written by each writer are divided into data to train and test the system in the ratio of 4:1.

TABLE I
COMPARISON OF EARLIER METHODS

Method	Feature Extraction	Classification	Accuracy
Puneet Jhajj et al. [16]	zoning density	SVM with RBF kernel	73.83%
Ubeeka Jain et al. [17]	profiles, width, height, aspect ratio, neocognitron	Neocognitron Neural Network	92.78
J.Pradeep et al. [8]	Diagonal Features	KNN	94.12%
Kartar Singh Siddharth et al. [10]	zoning density and background directional distribution features	SVM with RBF kernel	95.04%
Our Method	Diagonal Featurs	SVM with RBF kernel	95.34%

REFERENCES

[1] A. S. Britto, R. Sabourin, F. Bortolozzi, C. Y. Suen, "Foreground and Background Information in an HMM- Based Method for Recognition of Isolated Characters and Numeral Strings", 9th International Workshop on Frontiers inHandwriting Recognition (IWFHR-9), 2004, pp. 371-376.

[2] N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.31 no.2, pp. 216 - 233. 2001.

[3] U. Akilandeswari, R. Nithya, B. Santhi," Review on Feature Extraction Methods in Pattern Classification" European Journal of Scientific Research,ISSN 1450-216X Vol.71 No.2 (2012), pp. 265-272© EuroJournals Publishing,Inc.2012,http://www.europeanjournalofscientificresearch.com.

[4] Oivind Due Trier,Anil K. Jain And Torfinn Taxt," Feature Extraction Methods For Character Recognition"--A Surveypattern Reco#nition, Vol. 29, No. 4, pp. 641-662, 1996,Elsevier Science Ltd,Copyright © I996 Pattern Recognition Society,Printed in Great Britain. All rights reserved.

[5] Brijmohan Singh, Ankush Mittal, Debashis Ghosh," An Evaluation of Different Feature Extractors and Classifiers for Offline Handwritten Devnagari Character Recognition, Journal of Pattern Recognition Research 2 (2011) 269-277,Received December 12, 2009. Accepted September 14, 2011.

[6] R. Jayadevan, Satish R. Kolhe,Pradeep M.Patil and Umapada Pal,"Offline Recognition of Devanagri Script:A Survey,IEEE TRANSACTIONS ON SYSTEMS,MAN, AND CYBERNETICS-PART C:APPLICATIONS AND REVIEWS,VOL.41,NO. 6,NOVEMBER 2011,1094-6977/$26.00©2010 IEEE.

[7] U.Pal,B.B.Chaudhuri,"Indian Script Character Recognition:A Survey".

[8] J.Pradeep,E.Srinivasan and S.Himavathi,"Diagonal Feature Extraction Based Handwritten character Using Neural Network" International Journal of Computer Applications (0975 – 8887)Volume 8– No.9, October 2010,**2011 IEEE.**

[9] Munish Kumar, M. K. Jindal and R. K. Sharma,"k-Nearest Neighbor Based Offline Handwritten Gurmukhi Character Recogniion" 2011 International Conference on Image Information Processing (ICIIP 2011).

[10] Kartar Singh Siddharth,Mahesh Jangid, Renu Dhir, Rajneesh Rani,"Handwritten Gurmukhi Character Recognition Using Statistical and background Directional Distribution Features" International Journal International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010Journal International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010.

[11] G **S** Lehal and Chandan Singh," A Gurmukhi Script Recognition System" 2000 IEEE.

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

[13] G.S.Lehal and Chandan Singh," A post-processor for Gurmukhi OCR".

[14] R.G. Casey and E.Lecolinet, "A Survey of Methods and Strategies in Character Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No.7, July 1996, pp. 690-706.

[15] Anuj Sharma, Rajesh Kumar, R.K. Sharma, "Online Handwritten Gurmukhi Character Recognition Using Elastic Matching", Conference on Image and Signal Processing, IEEE Computer Society, 2008.

[16] Puneet Jhajj, D. Sharma, "Recognition of Isolated Handwritten Characters in Gurmukhi Script", International Journal of Computer Applications (0975-8887), Vol. 4, No. 8, 2010.

[17] Ubeeka Jain, D. Sharma, "Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognition", International Journal of Computer Applications (0975-8887), Vol. 4, No. 8, 2010.

[18] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[19] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification", [Online]. Available:http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

