Volume 2, Issue 6, June 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Handwritten Devanagari Numeral Recognition Using Zonal Based Feature Extraction Method and SVM Classifier

Gita Sinha, Mrs. Rajneesh Rani, Prof. Renu Dhir
Department of Computer Science and Engineering
Dr B R Ambedkar National Institute of Technology
Jalandhar- 144011, Punjab (India)
gitawit321@gmail.com

ABSTRACT-The ability to identify machine printed characters in an automated or a semi-automated manner has obvious applications in numerous fields. Since creating an algorithm with a one hundred percent correct recognition rate is quite probably impossible in our world of noise and different font styles, it is important to design character recognition algorithms with these failures in mind so that when mistakes are inevitably made, they will at least be understandable and predictable to the person working with the program. In this paper we present an overview of Feature Extraction techniques for off-line recognition of isolated Devanagari numeral recognition. Our paper presents Zone based approach which is the combination of image centroid zone and zone centroid zone of numeral/character image. In image centroid zone character is divided into n equal zone and then image centroid and the average distance from character centroid to each zones/grid/boxes present in image is calculated. Similarly, in zone centroid zone character image is divided into n equal zones and centroid of each zones/boxes/grid and average distance from zone centroid to each pixel present in block/zone/grid is calculate. We have used SVM for subsequent classifier and recognition purpose. We obtained 99.11% recognition accuracy by SVM.

Keywords: Gurmukhi Script Image processing, pattern recognition SVM, ICZ, ZCZ.

1. INTRODUCTION

Optical Character Recognition (OCR) lets you convert images with text into text documents using automated computer algorithms.

OCR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition. The input of one step is the output of next step. Pre-processing consists of these operations slant correction, normalization and thicking and these have been adopted for the purpose of Feature Extraction the zone based method is used [2]. The last phase Classification SVM (Support Vector Machine) have been used as a classifier. The flow chart of a typical OCR can be shown as Figure 1.

Recognition of handwritten character is one of the most interesting topics in pattern recognition. Applications, of OCR is in different area especially digit recognition, which deals with postal mail sorting, bank check processing, form data entry, vehicle plate recognition, postal address block detection and recognition ,camera OCR etc. For these applications, the performance (accuracy and speed) of digit recognition is most important factor. While in pattern classification and machine learning communities, the problem of handwritten digit recognition is a good example to test the classification performance [3]. HCR is very valuable in terms of the variety of applications and also as an academically challenging problem. When HCR is used as a solution for inputting regional language data and also as a solution for converting paper information to soft form, the internet can be enriched with regional information, so that the digital divide can be minimized. It also facilitate solution to processing large volumes of data automatically, for example, in processing hand-filled application forms into machine printed character /number. Hence many research work are going on different scripts. But on Indian language scripts, very scanty literatures are available [4].

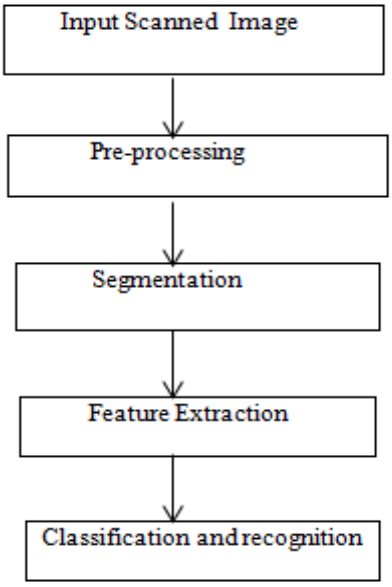

Figure 1: Phases in OCR

We have studied various research paper which reveals that the difficulty of the problem has two aspects. The first is attributed to the writer variations in style, size, shape, ink colour, ink flow and thickness, digitization imperfections etc. The second is the deficiencies of the particular method used for feature extraction. Indian scripts share a large number of structural features due to common Brahmi origin. The differences between the scripts are primarily on their form of writing. The written form has more curves than straight or slant lines and has many similarities among different character of the same scripts and also between the scripts of different languages. The scripts of some languages like Hindi, Marathi, Punjabi etc, are same and there is many similarities between Kannada and Telugu. We used Gurmukhi script for our experiments. Gurmukhi characters are curved in nature with some kind of symmetric structure observed in the shape. This information can be best extracted as a feature if we extract statistical features from the images.

In the literature survey we have found that numbers of authors have attempted to recognize the Handwritten Gurmukhi Character using different techniques.

In 2010 Bharath A. and Sriganesh Madhvanath [17] In this paper, they study the relevance of stroke size and position information for the recognition of online handwritten Devanagari words by comparing three different preprocessing schemes. Experimental results indicate that the word recognition accuracy achieved using a preprocessing scheme that completely disregards the original sizes and positions of the strokes In 2011 Mahesh Jangid have proposed these method for feature extraction: Zonal density, Projection histogram ,Distance Profiles, Background Directional Distribution (BDD) available [8], and SVM for classification and they have got 98%,99.1% and 99.2% of accuracy.

In 2010 Shailedra Kumar et. al.[12] using Support Vector Machine for Handwritten Devanagari Numeral Recognition. Moment Invariant and Affine moment Invariant techniques are used as feature extraction. This linear SVM produces 99.48% overall recognition rate which is the highest among all techniques applied on handwritten Devanagari numeral recognition system.

In 2009 S. V. Rajashekararadhya et.al.[10] have used the zone based feature extraction method on handwritten numeral/mixed numerals recognition of south-indian scripts. Feed forward back propagation neural network, and Support Vector Machine for classifier. They have obtained overall 98.9% of accuracy.

In 2008 S.V. Rajashekararadhya et. al [2] have used efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south indian scripts and Nearest neighbour ,Neural network as a classifier . They have got 98.5% of accuracy.

In 2007 M. Hanmandlu et.al [11] using Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals. This paper presents the recognition of Handwritten Hindi Numerals based on the modified exponential membership function fitted to the fuzzy sets derived from normalized distance features obtained using the Box approach method . They have obtained 95% recognition accuracy.

ZHAO Bin et. al. [18] they have used Support Vector Machine for classification and its Application in Chinese check recognition system. The experiment on NIST numeral database and the actual check samples shows that comparison with other classifiers, SVM possesses better generalization ability. They have got 99.95% accuracy.

Bikash Shaw et al [22] in this paper present offline handwritten devanagariword recognition: a segmentation based approach novel segmentation based approach is proposed for recognition of offline handwritten Devanagari words. Stroke based features are used as feature vectors A hidden Markov model is used for recognition at pseudocharacter level. The word level recognition is done on the basis of a string edit distance

2. Dataset by ISI

Devanagari Numerals dataset is not widely available for the research purpose only the Indian Statistical Institute (ISI), Kolkata provides the Devanagari Numerals Dataset after sending the proper request by Mahesh Jangid. All the experiment is done on the standard dataset provided by ISI. Dataset have 10 different folders which have collection of image of Devanagari Handwritten Numerals.

3. Devanagari Script

Devanagari also called Nagari (Nāgarī, नागरी, the name of its parent writing system), is an abugida alphabet of India and Nepal. It is written from left to right, does not have distinct letter cases, and is recognizable (along with most other North Indic scripts, with few exceptions like Gujarati and Oriya) by a horizontal line that runs along the top of full letters.

०	शून्य	shūnya
१	एक	Ek
२	दो	đo
३	तीन	tīn
४	चार	chār
५	पाँच	pāñch
६	छह	chhah
७	सात	sāt
८	आठ	Āth
९	नौ	nao/nau

Figure 2 Sample of devanagari numerals.

Devanāgarī is the main script used to write Standard Hindi, Marathi, and Nepali. Since the 19th century, it has been the most commonly used script for Sanskrit. Devanāgarī is also employed for Bhojpuri, Gujari, Pahari, (Garhwali and Kumaoni), Konkani, Magahi, Maithili, Marwari, Bhili, Newari, Santhali,

Tharu, and sometimes Sindhi, Dogri, Sherpa and by Kashmiri-speaking Hindus. It was formerly used to write Gujarati. The figure 2 shows sample of devanagari numbers from 0-9.OCR consist of following stage pre-processing, segmentation, feature extraction, classification and recognition.

4. PRE-PROCESSING

In preprocessing operations sample image are converted into gray scale .Then we have applied these techniques, Gray scale image are converted into binary image using threshold value obtained by Otsu's method, filtering operation, morphological operation, removal of noise having less than 30 pixels, Binarization, contour smoothing, skew detection, and skeletonization of a digital image so that subsequent algorithms along the road to final classification can be made simple and more accurate [14].

The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the HGCR system to operate accurately [6]. After pre-processing phase, a cleaned image is available that goes to the segmentation phase.

5. SEGMENTATION

Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images. Script segmentation is done by executing the following operations: Line segmentation, Word segmentation and character segmentation [7]

6. FEATURE EXTRACTION-

The selection of good feature set is the most the important aspect of handwritten character recognition. This method provide the ease of implementation and good recognition. We have used following sets of extracted features to recognize Devanagari numerals. In the next section we have define these algorithm step-by-step .The following paragraph explained the details about feature extraction method.

We have computed the centroid of image (numeral/character). This image is further divided into 100×100 equal zones where size of each zone is (10×10). Then we have computed the average distance from image centroid to each pixel present in the zones/block. We have got 100 feature vector of each image. Similarly in ZCZ we have divided image into n equal zones and calculated centroid of each zones. Then compute the average distance from the zone centroid to each pixel present in zones. There could be some zones that are empty then the value of that particular zone is assumed to be zero. We repeat these procedure for all zones present in image(numeral/character) [2].

We have used efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south Indian scripts that has been define in this paper [2]. But we have apply the same method on Devanagari numeral recognition. Algorithm 1 provides Image centroid zone (ICZ) based distance metric feature extraction system and Algorithm 2 provides Zone Centroid Zone (ZCZ) based Distance metric feature extraction system. Algorithm 3 provides the combination of both (ICZ+ZCZ) feature extraction system. The following are the algorithms to show the working procedure of our feature extraction methods and also in figure 3.

Algorithm 1: Image Centroid Zone (ICZ) feature extraction system.

Input: Image(character/numeral) are Pre-processed.

Output: Extract the Features for Classification and Recognition.

Algorithm

Step 1: Calculate centroid of input image.

Step 2: Division of input image in to **n** equal zones.

Step 3: Computation of the distance from the image Centroid to each pixel present in the zone.

Step 4: Repeat step 3 for the entire pixel present in the zone/boxes/grid.

Step 5: computation of average distance between these Points.

Step 6: Repeat this procedure sequentially for the entire zone present in the image.

Step 7: Obtaining **n** such feature for Classification and recognition process.

Ends.

Algorithm 2: Zone Centroid Zone (ZCZ) based feature extraction system.

Method Begins

Step 1: Division of input image in to **n** equal zones.

Step 2: Compute centroid of each zones.

Step 3: Compute the distance between the zone centroid to each pixel present in the zone.

Step 4: Repeat step 3 for the entire pixel present

in the zone/box/grid.

Step 5: computation of average distance is between these points present in image.

Step 6: This procedure are sequentially Repeat for the entire zone.

Step 7: obtaining **n** such features for classification and recognition.

Ends.

Proposed hybrid Algorithm 3: Which is the combination of both **(**ICZ + ZCZ) based Distance metric feature extraction system.

Input: Image(numeral/character) are Pre-processed.

Output: Extract Features for Classification and Recognition.

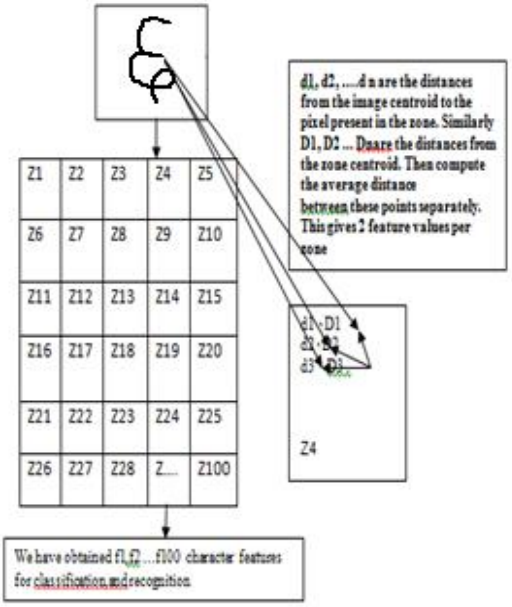

Figure 3 All procedure to extract feature from Devanagari numeral image "six".

Method Begins

Step 1: Compute Centroid of input image.

Step 2: Division of input image into **n** equal zones.

Step 3: Computation of the distance between the image centroid to each pixel present in the zone.

Step 4: Repeat step 3 for the entire pixel present in the zone.

Step 5: Computation of average distance between these all points in the image.

Step 6: Compute centroid of the zone/block .

Step 7: Computation of the distance between the zone centroid to each pixel present in the zone.

Step 8: Repeat step 7 for the entire pixel present in the zone.

Step 9: Computation of average distance between these points.

Step 10: Repeat the steps 3-9 sequentially for the entire zone.

Step 11: Obtaining **2×n** such features for classification and recognition.

Ends

7. CLASSIFICATION AND RECOGNITION

We have used Support Vector Machines (SVM) for the purpose of Classification and recognition.

SUPPORT VECTOR MACHINES

SVM is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input.

The process of rearranging the objects is known as mapping (transformation). Rearranging the object, using a set of mathematical functions, known as kernels. There are some common Kernel functions that include the linear kernel, the polynomial kernel, radial basis function (RBF) and sigmoid [9].

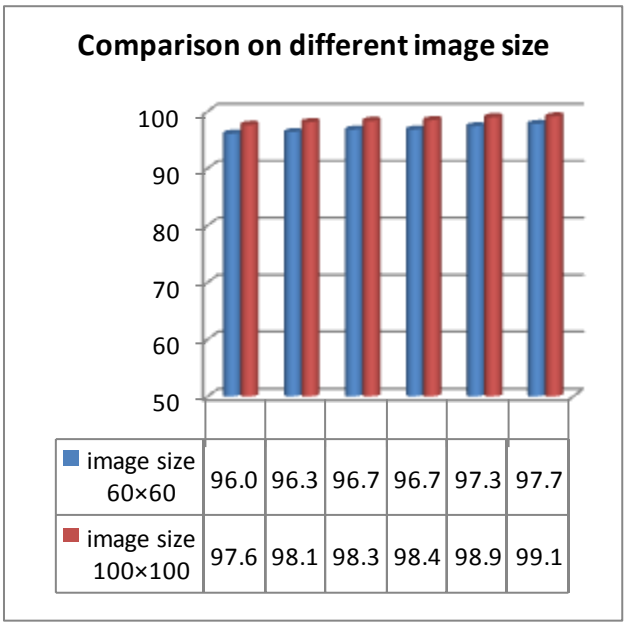

Figure 4. Devanagari numeral recognition using SVM.

We have obtained such multiclass SVM tool LIBSVM available at [1]. We have used RBF (Radial Basis Function) kernel which is also common choice, in our recognition. RBF has single kernel parameter gamma (g or γ).

SVM have proved to achieve good generalization performance by the use of concept of basis, without knowledge of the prior data [15]. In figure 4.5 we compare the recognition accuracy on different size of images. we got 97.79 highest accuracy on image size 60×60, and block size 6×6 on feature vector 100. Then on image size 100×100, block size 10×10 which also makes 100 feature vectors which produce 99.11% accuracy which is highest. This highest result is 99.11% which is obtained at γ = 0.001 with combination of C = 4, 8,32, 64 and 500

EXPERIMENTAL RESULT

We have applied above Recognition strategy on Devanagari numeral. The dataset consist of 22546 samples of Devanagari numeral. we have used Zone based Feature Extraction techniques on this samples and practiced on different size of image. We obtained different recognition accuracy on different size of image. The highest accuracy obtained from image size 100×100 and zone (10×10). We have obtained 200 Feature 100 Vector from both of the methods. The accuracy depends upon size of image and number of Feature vector we have obtained from the image. The recognition accuracy also depend on SVM parameter(C, γ). We have obtained 99.11% of accuracy on Devanagari numeral. Table 4.5 The recognition accuracy on image size 100×100.

S.No.	**Cost (*c*)**	Recognition accuracy
1	**1**	97.62 %
2	2	98.11 %
3	4	98.34 %
4	8	98.45 %
5	**16**	**98.95%**
6	**500**	**99.11 %**

CONCLUSION

In this paper we have used a Zone Based Feature Extraction Techniques which is the combination of Image Centroid Zone and Zone Centroid Zone by S.V. Rajashekararadhya and Dr P. Vanaja Ranja . This algorithm has shown a notable improvement for recognizing of handwriten Devanagari numeral.. Our Experimental Results proved that ZCZ method provide better recognition accuracy than ICZ.

ACKNOWLEDGMENT

We are very thankful to Mahesh Jangid for his sincere help towards the provision of collected dataset for our experiment.

REFERENCES

[1]. Chih-Chung Chang and Chih-Jen Lin LIBSVM : A Library of Support Vector Machine software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2]. S.V. Rajashekararadhya, Dr P. Vanaja Ranjan, . 2008 "efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south indian" journal of theoretical and applied information technology

[3]. Omid Rashnodi, , Hadieh Sajedi, Mohammad Saniee "using box box approach in persiam handwritten digit recognition" International Journal of Computer Applications (0975 – 8887) Volume 32– No.3, October 2011.

[4]. L R Ragha, M Sasikumar "using moment feature for gabor directional image for kannada handwritten character recognition" International Conference and Workshop on Emerging Trends in Technology (ICWET 2010) – TCET, Mumbai, India

[5]. Xuewen Wang Xiaoqing Ding ,Changsong Liu "Gabor filters- based feature extraction on character recognition" pattern recognition 38 (2005) 369-379.

[6]. Naveen Garg "Handwritten Gurmukhi Numeral Recognition using Nueral Network " M.tech Thesis, Thapar University, Patiala

[7]. Arif Billah Al-Mahmud Abdullah and Mumit Khan " a survey on script segmentation for bangla ocr" Working Papers 2004-2007.

[8]. Mahesh Jangid Kartar Singh, Renu Dhir Rajneesh Rani "Performance Comparison on Devanagari Handwritten Numeral Recognition" International Journel of Computer Application (0975-8887) volume-22 No.-1, May 2011 .

[9]. http://www.statsoft.com/textbook/support-vector machines/

[10]. V. Rajashekararadhya, P. Vanaja ranjan, "Handwritten Numeral/Mixed Numerals Recognition of South Indian: Zona based Feature Extraction Method" , 2005 - 2009 JATIT.

[11]. M. Hanmandlu, J. Grover, V. K.. Madasu, S. Vasikarla " Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals " International Conference on Information Technology (ITNG'07) 0-7695-2776-0/07 ,2007 IEEE.

[12]. Shailedra Kumar Shrivastava, Sanjay S. Gharde " Support Vector Machine for Handwritten Devanagari Numeral Recognion " International Journel of Computer Application (0975-8887) Volume 7-No. 11,October 2010

[13]. Kartar Singh Siddharth Renu Dhir Rajneesh Rani "Handwritten Gurmukhi Numeral Recognition using Different Feature Sets" International Journal of Computer Application (0975-8887) Vol. 28 No.-2 , August 2011

[14]. H. Swethalakshmi, Anita Jayaraman, V. srinivasa Chakravarthy , C. Chandra Sekhar , "Online Handwritten Recognition of Devanagari and Telgu Character using Support Vector machin".

[15]. Guca.sourceforge.net/.../introductiontogurmukhi.

[16]. Bharath A. and Sriganesh Madhvanathl " On the Significance of Stroke Size and Position for Online Handwritten" 2010 International Conference on Pattern Recognition

[17]. Devanagari Word Recognition: An Empirical StudyZHAO Bin, LIU Yong and XIA Shao- Wei "Support.

[18]. Vector Machine and its Application in Handwritten Numeral Recognition" 0-7695-0750-6/00 2000 IEEE

[21].http:saravananthirumuruganathan.wordpress.com2010/05/17/a-detailed-introduction-to-k-nearest- neighbor-knn-algorithm/

[22] Bikash Shaw et al [22] "offline handwritte Devanagari word recognition: a segmentation based approach. **"**978-1-4244-2175 6/08/$25.00 ©2008 IEEE

