STRUCTURAL FEATURES FOR RECOGNITION OF HAND WRITTEN KANNADA CHARACTER BASED ON SVM

S.A.Angadi [a] and Sharanabasavaraj.H.Angadi [b]

[a]Department of Computer Science and Engineering, Visvesvaraya Technological University,Belgavi, Karnataka,India
[b]Department of Computer Science and Engineering, Rural Engineering College,Hulkoti Karnataka,India.

ABSTRACT

Research in image processing involves many active areas, of these Recognition of Handwritten character holds lots of promises and is challenging one .The idea is to enable the computer to be able to recognize intelligibly hand written inputs In this paper, a new method that uses structural features and support vector Machine (SVM) classifier for recognition of Handwritten Kannada characters is presented. On an average recognition accuracy of 89.84 % and 85.14% for handwritten Kannada vowels and Consonants obtained with this proposed method, inspite of inherent variations.

KEYWORDS*—Handwritten Character recognition (HCR), Kannada script, preprocessing, feature extraction, SVM Classifier.*

1. Introduction

Character recognition is an important subtask of document image processing [1, 2]. It involves identifying the various characters that make up the text of the document. The field of character recognition has seen many reported works, most of which are on English and other foreign languages. It has been noticed that optical Character Recognition finds application in various areas and many researchers have shown interest in finding better accuracy of character recognition on foreign languages, but to the some extent researchers are working on Indian languages and finds difficulties on these Indian script because of multi lingual, hence there is very less work been seen in Kannada language, which means to say there's a scope of work to be done. One of challenging field is hand written Character Recognition within the applications of character recognition. Hand written information used as an important mode of communication between the people, since the beginning of mankind and will continue through many ages. Therefore handwritten recognition [3, 4] plays an important role in the fields of pattern recognition .Typical pattern recognition system operates in two phases, the very first is Training (learning) and second one is Testing (Recognition). In first phase that's training method system learns from large number of patterns of which the classes are known: in the recognition phase the system is required to classify patterns for which the classes are unknown. First method does consist of image preprocessing, feature extraction and feature storage. In the second i.e.

DOI : 10.5121/ijcseit.2015.5203

recognition phase preprocessing of unknown image is occurred then its features are obtained and compared to those learned in the training phase that's first phase.

Enormous work has been recently noted down on the development of handwritten character recognizers for English, Arabic, Chinese and other scripts. However, there is very less work has been noted down on developing OCR system especially for Indian languages like Tamil, Telugu, Gurumukhi, Oriya , because of India is a multi-lingual and multi script country. However, to the best of our knowledge very little work has been carried out with respect to Kannada language. Due to the impact and the advancement in the information Technology , more emphasis is being given in Karnataka to the use of Kannada at all levels and hence the use of Kannada in a computer system is also a necessity Therefore , efficient OCR system for Kannada are need of the day.

The proposed method works in two phases. The first phase(Training phase), preprocessing operations like resizing of image is done so as to make the recognition process independent of size then the normalization of the image is carried out to extract the structural features like perimeter, area, eccentricity etc from the set of samples and support vector machine is trained. During the second phase (testing phase) the set of samples are preprocessed and the structural features are obtained. The trained feature set along with the testing features is given to the recognition module which employs support vector machine classifier to recognize the characters. The method has been evaluated on hand written Kannada vowels and consonants collected from students of schools and colleges. These images are scanned by the flatbed scanner using 300 dpi. The system has given a recognition accuracy of 89.84 % and 85.14% for vowels and consonants respectively.

The remaining part of the paper is organized as follows Section 2 presents a survey of literature in recognition of Handwritten Indian language text. Section 3 describes Kannada character set and database. Section 4 describes the method for recognition of hand written Kannada character. Section 5 presents experimentation and Section 6 gives conclusions.

2. Literature Review

Handwriting recognition (or HWR) is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. Recognition of handwritten characters by a computer is a difficult problem due to the human handwriting variability, uneven skew, orientation writing habit, style. A few such recognition algorithms are described in the following

Amitabh Wahi, [5] et.al, presented a paper on Handwritten Tamil Character Recognition using Moments. Zernike moments and Legendre Polynomial features are used in the pattern recognition to extract the features of Tamil characters. Neural classifier has been used for the classification purpose. Handwritten Bangla characters features are extracted by local chain-code histograms using MLP classifier proposed by Bhattacharya et al [6] . Das et al [7] have extracted features based on quad tree, shadow, and longest run. And using these features multi-layer perceptron (MLP) and support vector machine (SVM) classifiers recognizes different groups of characters. Thechallenges involved in the identification of scripts particularly on handwritten documents are reviewed in [8]. Here described the challenges involved in script identification, the applications of script identification. Ashwin et al [9] have formed three basic Zones for the underlying character

image Support vector machines are employed for the classification of characters and have achieved an accuracy of 86.11%.

The Gaussian filters are used to down sample each segmented block to extract directional features for Kannada,Tamil and Telugu hand written characters recognition with aid of quadratic classifier is presentedin [10]. In [11] FLD based unconstrained handwritten Kannada character recognition by using Euclidean distance measure is described and the mean recognition accuracy of 68% is reported. Shreya N. Patankar et.al[12] have proposed a method that aims at recognizing Marathi language -Barakhadi characters by recognizing a vowel and a consonant separately using Invariant moment features with quadratic classifier. Recognition of isolated handwritten Kannada vowels is proposed in [13].The Invariant moments are used as features and K-NN classifier is used for classification. The recognition results for vowels are 85% on average. Mamatha et.al [14] describesatechnique to remove the noise induced in the handwritten Kannada documents. Tamil handwritten recognition system [15] having different font size and type extracts features from Zernike moments and Legendre Polynomial which have been used in pattern recognition with aid of Neural classifiers. Using Support Vector Machine (SVM) an attempt is made to recognize the similar looking Bangla basic characters, numerals and vowel modifiers [16]. From the literature survey, it is evident that there is still lot of on scope for research in handwritten Kannada character recognition.

In this paper, a SVM based approach for handwritten character recognition using structural features is proposed. As an initial attempt, the work is restricted to isolated and constrained vowels and consonant characters rather than considering entire character set.

3. Kannada Character Set and Database

Kannada script consists of 49 basic characters which are grouped into swaragalu (vowels),Vyanjanagalu-consonants and Yogavahakagalu..i.e, The Kannada script also has 10 numerals from 0 to 9 Kannada handwriting recognition is challenging task due to large character set, complex shape presence of compound characters and modifiers and similarity between characters etc. There is no standard dataset of handwritten Kannada texts available, hence we have collected handwritten text and built our own dataset for characters. These datasets were collected from students of primary schools and engineering college. For this purpose every individual was asked to write vowels and consonants on prescribed forms. The forms were scanned at 300 dpi. Asgray scale images, using a flatbed HP scanner. The characters were then manually extracted from the scanned Images. For each of the 49 symbols we selected 50 samples are used. In total there are 2490 samples in the whole dataset for training and testing. The detailed description of the feature extraction and classification of proposed methodology is given in the following section.

4. Proposed Methodology

Structural and topological features used by the proposed method for recognition of hand written Kannada vowels and consonants. Figure 3. Illustrates block schematic of the proposed method for recognition of handwritten Kannada character under which major steps included are preprocessing, feature extraction, knowledge base for training and classifier. SVM consists of training module (SVM_train) and classification module (SVM_test).The proposed methodology works as follows;

- Collect the data samples
- Scan with 300 dpi and store it as bmp format
- Normalize the image to the size of 30x30 pixels and apply thinning operation on it.
- Extract the structural features and store the features as vector
- Train SVM with the above features. The support vectors are stored as models
- Test the SVM with the features of unknown characters.

The detailed explanation of each stage is given in the following

Image acquisition

In image acquisition, the recognition system acquires a scanned image as an input in bmp format using flatbed scanner. The acquired image is given to the preprocessing phase.

Pre Processing

The images acquired may contain skewness, noises etc, which leads to the miss classification of characters. In order to reduce these factors, preprocessing is very essential.

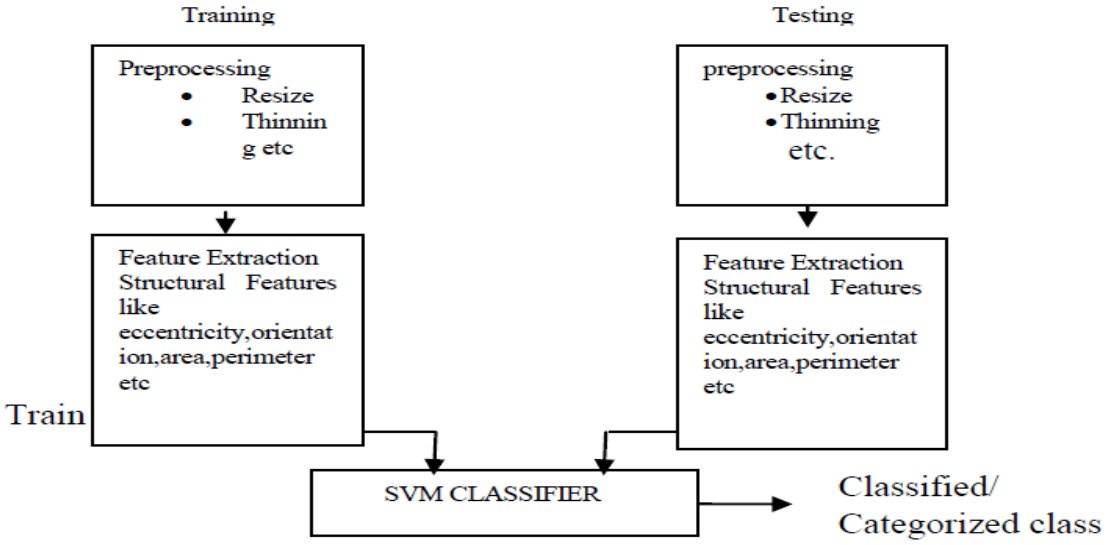

Figure 3.Proposed flow diagram of Handwritten Character Recognition System

The goal of preprocessing is to increase the quality of hand written data. The preprocessing stage performs size normalization, bounding box generation and further the thinning operation. The sequences of pre-processing steps are described below.

Normalization

It is the process of converting the random sized image into standard sized image. For this purpose the character images are resized to 30x 30 pixel size. The normalized image is subjected to the thinning process.

Thinning

To reduce the storage space and processing time thinning is carried out, without distorting shape. The thinning process is as follows afterapplying bounding box, it extracts the shape information of the character. Further thinning is carried out on the image which is cropped and the morphological operator is employed for the purpose. The thinned image is further processed for feature extraction.

Feature Extraction

The feature extraction [17, 18, 19, 20, 21, 22, 23, 24] stage captures the distinct characteristics of the digitized character for recognition. The main goal of feature extraction process is to find unique patterns in image discriminating pattern classes of images. In this phase for each character, a structural feature vector is extracted and it comprises of

The following structural/topological features. In this stage, the structural features are extracted from the images using regionprop( ) and are stored into a feature vector F as
F= [StructFeatures] StructFeature=[SFi] 1=<i<=43 Where SFi is the structural features of the character images. The dataset from the images are then applied to the training phase using svmtrain() .The training process trains images with the given structural feature vectors. These structural features are used by the classifier to categorize the character image.

Classification

Support vector machine (SVM) classifiers [25, 26, 27, 28, 29, 30, and 31] have gained prominence in the field of pattern recognition/classification. The SVM process generates the support vectors on margins on which data points lies. SVM approximates the function using the following form

$$f(x) = \mathrm{sgn}(w . \Phi(x) + b) \quad (1)$$

Where, w is the weight vector, b is a bias and Φ(x) represents a high-dimensional feature space which is nonlinearly mapped from the input space x. The coefficients w and b are estimated by minimizing the regularized risk function. The ONA approach is used in the proposed method for decomposition of the classification problem from N class pattern recognition into several two class classification problems. The explanation using the above mentioned techniques for Kannada character recognition is described in the following section.

5. Experimentation

The proposed methodology has been evaluated for hand written Kannada vowels and consonants which are collected from different persons. These images are having uneven thickness and other degradations. Therefore a bounding box is fit to extract the exact character image and the image is then resized to 30x30 pixel sizes and the resulting image is thinned.

Further the structural/topological features of the character like orientation, Filled Area, Perimeter, Eccentricity, EquiviDiameter, Convex Area and so on are extracted .

The recognition results are also brought out in Figure 5 and Figure 6 for the vowel and consonants. The overall recognition rates of vowels is 89.84 % and consonant is 85.14% The method is implemented on Pentium Processor T4300(2.1 GHz,800MHzFSB) System with 3GB RAM with Mat lab 7.8.0 and results as shown are satisfactory.

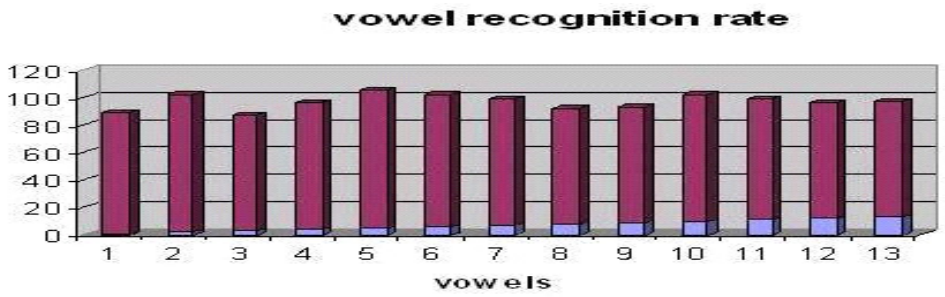

Figure 5. Vowel recognition performance

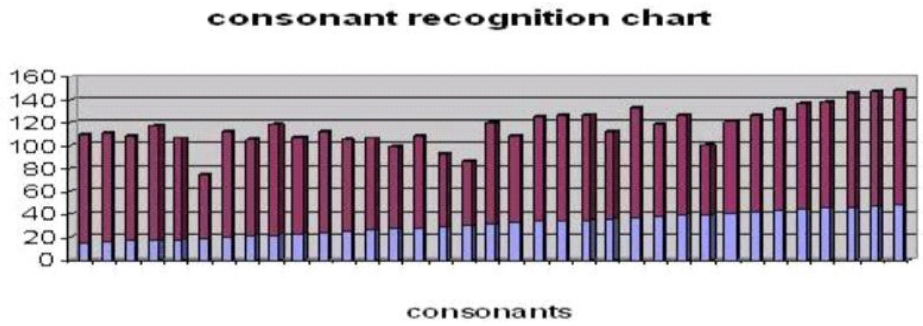

Figure 6. Consonant recognition performance

6. Conclusion

For hand written Kannada vowels and consonants recognition the proposed method is evaluated using structural features. The proposed method is capable of recognizing isolated and constrained handwritten Kannada vowels and consonants. The recognition system has training and testing phase. The characters are written from different persons are scanned using a flatbed scanner with 300 dpi. The images have uneven thickness and other degradations. Therefore preprocessing

operations is performed by bounding fitting, resizing the image to 30x30 and thinning of the image, so on. Further the structural features of the character are explored for construction of knowledge base. The test characters are categorized into vowel/consonant classes using the multiclass SVM classifier, obtaining 89.84 % efficiency in recognizing Kannada vowels (swaragalu) and 85.14% efficient in recognizing consonants. There is a scope for improving the methodology by using better features, which will be explored, in future works.

References

[1] J.Mantas. : An overview of character recognition methodologies, Pattern Recognition, 425–430, (1986)

[2] K. Govindan, A.P. Shivaprasad :Character recognition— a survey, Pattern Recognition vol 23 pp 671– 683, (1990).

[3] C. C. Tappert, C. Y. Suen and T. Wakahara, : The state of the art in online handwriting recognition, IEEETransaction on PAMI. Vol. 12, No. 1, pp. 787–808. (1990).

[4] R. Plamondon, S.N. Srihari, : On-line and off-line handwritten recognition: a comprehensive survey, IEEE Trans.Pattern Anal. Mach. Intell. pp 62–84 (2000).

[5] A. Dr.Amitabh Wahi, B. Mr.Su.Sundaramurthy, C. P.Poovizh,I, Handwritten Tamil Character Recognition using Moments, International journal of Computer Science & Netw ork Solutions - Volume 2.No3 (March.2014)

[6] Bhattacharya U, Shridhar M and Parui S K, : On recognition of handwritten Bangla characters, In Proceedings ofthe Indian Conference on Computer Vision, Graphics and mage Processing, 817–828. (2006).

[7] Das N, Das B, Sarkar R, Basu S, Kundu M and Nasipuri M : Handwritten Bangla basic and compound character recognition using MLP and SVM classifier, J. Comput. 2: 109–115, (2010).

[8] D S Guru,M Ravikumar, B S Harish. : A Review on Offline Handwritten Script Identification,InternationalJournal of Computer Applications (0975 – 8878) on National Conference on Advanced Computing and Communications –NCACC (April 2012).

[9] Ashwin T.V. and Sastry P.S.: A font and size- independent OCR system for printed Kannada documents using support vector machines, 27, (1), 35–58, (2002).[10] U. Pal, N. Sharma, T. Wakabayashi and F. Kimura, : Handwritten character recognition of popular south Indian scripts, Proceeding SACH'06 Proceedings of the conference on Arabic and Chinese handwriting recognitionSpringer-Verlag Berlin, Heidelberg, (2006).

[10] U. Pal, N. Sharma, T. Wakabayashi and F. Kimura, : Handwritten character recognition of popular south Indian scripts, Proceeding SACH'06 Proceedings of the conference on Arabic and Chinese handwriting recognition Springer-Verlag Berlin, Heidelberg, (2006).

[11] Niranjan S.K, Vijaya Kumar, Hemantha Kumar G, Manjunath Aradhya V N. : FLD based Unconstrained Handwritten Kannada Character Recognition, Second International Conference on Future GenerationCommunication and Networking Symposia (2008).

[12] Shreya N. Patankar Leena R. Ragha, : Zonal moments based Handwritten Marathi Barakhadi recognition, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, (August–2012).

[13] Sangame S.K., Ramteke R.J., Rajkumar Benne, : Recognition of isolated handwritten Kannada vowels, Copyright © 2009, Bioinfo Publications, Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, pp-52-55. (2009).

[14] Mamatha H.R, Sonali Madireddi, Srikanta Murthy K, : Performance analysis of various filters for De-noising of Handwritten Kannada documents, International Journal of Computer Applications (0975 – 888) Volume 48– No.12, (June 2012).

[15] Dr.Amitabh Wahi, Mr.Su.Sundaramurthy, P.Poovizhi, : Handwritten Tamil CharacterRecognition using Moments , International journal of Computer Science & Network Solutions, Volume 2.No3, (March.2014).

[16] Khondker Nayef Reza, Mumit Khan, : Grouping of Handwritten Bangla Basic Characters,Numerals and Vowel Modifiers for Multilayer Classification, International Conference Frontiers in Handwriting Recognition, (2012).
[17] L. Heutte , T. Paquet , J.V. Moreau , Y. Lecourtier , C. Olivier . : A structural/statistical feature based vector for handwritten character recognition, Pattern Recognition Lettersvol19, pp 629–641, (1998).
[18] Leena R Ragha and M Sasikumar, : Feature Analysis for Handwritten Kannada Kagunita Recognition, InternationalJournal of Computer Theory and Engineering, Vol.3, No.1, (February, 2011).
[19] Aditya Raj, Ranjeet Srivastava, Tushar Patnaik, Bhupendra Kumar, : A Survey of FeatureExtraction and Classification Techniques Used In Character Recognition for Indian Scripts, International Journal of Engineeringand Advanced Technology (IJEAT) ISSN: 2249–8958, Volume-2, Issue-3, (February 2013).[20]Mamatha H. R., Karthik S., SrikantaMurthy K., : Feature Based Recognition of Handwritten Kannada Numerals – A Comparative Study, International Conference onComputing, Communication and Applications (ICCCA), (Feb, 2012).
[21] Anil. K. Jain and Torfinn Taxt, : Feature extraction methods for character recognition-A Survey, PatternRecognition, vol. 29, no. 4.Newyork pp 641-662, (1996).
[22] B.V.Dhandra, R.G.Benne and Mallikarjun Hangarge, : Multi-font multi-size Kannada numeral recognition based on structural feature, eIT- 07,pp-193-199, 2nd National conference on Emerging trends in Informatiom Technology(eIT-2007). (2007).
[23] Kartar Singh Siddharth, Renu Dhir, Rajneesh Rani, : Handwritten Gurumukhi Character Recognition Using Zoning Density and Background Directional Distribution Features(IJCSIT), International Journal of Computer Science andInformation Technologies, Vol. 2 (3) , 1036-1041, (2011).
[24] R Sanjeev Kunte, R D Sudhaker Samuel,: An OCR System for Printed Kannada Text Using Two - Stage Multi-network Classification Approach Employing Wavelet Features, IEEE pp. 349 – 353. (Dec-2007).
[25] C. J. C. Burges, : A tutorial on support vector machines for pattern recognition., Dat aining and KnowledgeDiscovery, pp 121-167. (1998).
[26] Arvind C.S., Nithya E. And Nabanita Bhattacharjee, : Kannada Language Ocr System Using SVM Classifier, Journal of Information Systems and Communication ISSN: 0976-8742, E- ISSN: 0976-8750, Volume 3, Issue 1, pp-92-95, (2012).
[27] G. G. Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, : Printed and handwritten mixed Kannada numerals recognition using SVM, International Journal on Computer Science and Engineering Vol. 02, No. 05, 1622-1626. (2010).
[28] J. Manikandan , B.Venkataramani, : Study and evaluation of a multi-class SVM classifierusing diminishing learning technique, Neurocomputing 73 1676–1685, (2010).
[29] G. G. Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, : Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 05, , 1622-1626, (2010).
[30] S V. Rajashekararadhya, P. Vanaja Ranjan, : Neural Network Based Handwritten Numeral Recognition of Kannada and Telugu Scripts, TENCON 2008 - IEEE Region Conference, Hyderabad, (2008).
[31] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley- Newyork