# A Comprehensive Survey on Kannada Handwritten Character Recognition and Dataset Preparation

Kiran Y. C
Research Scholar, Jain University
Associate Professor, Dept. of ISE
Dayananda Sagar College of Engineering
Bangalore, India

Lothitha B. J
M.Tech. Student Dept. of ISE
Dayananda Sagar College of Engineering, VTU Belgaum
Bangalore, India

***Abstract*- Recognition of handwritten text has been one of the active and challenging areas of research in the field of image processing and pattern recognition. Recognition of Kannada handwritten character is complicated compared to other languages. It has numerous applications which include postal mail application, reading aid for blind and conversion of any handwritten document into electronic form. There is no most robust dataset available for handwritten characters. This paper focuses on developing a dataset for offline handwritten Kannada character recognition and overview of the ongoing researches in this field.**

***Keywords: Optical Character Recognition, Kannada Scripts, Support Vector Machine, Fuzzy K- Nearest Neighbor***

## I. INTRODUCTION

Handwriting character recognition has always been a challenging and interesting task in the field of pattern recognition. An Optical Character Recognition (OCR) system is the process of transforming human readable and optically sensed data to machine understandable codes. The purpose behind an OCR is to identify and analyze a document image by dividing the page into line elements, further subdividing into words, and then into characters. Recognition of characters can be done either from printed documents or from hand written documents. The high performance of any recognition system depends on the detailed analysis of preprocessing and segmentation.

## II. KANNADA LANGUAGE

Kannada is a Dravidian language spoken primarily in Karnataka State in South India, and has a literature that dates from the ninth century. It is spoken not only in Karnataka, but to some extent in the neighbouring states of Andhra Pradesh, Tamil Nadu, and Maharashtra. Kannada is spoken by about 44 million people. The language has 47 characters in its alphabet set 13 vowels and 34 consonants. The characters called aksharas are formed by graphically combining the symbols corresponding to consonants, consonant modifiers and vowel modifiers using well defined rules of combination. The number of possible consonant-vowel combination is 34*15=544 and number of possible consonant-consonant-vowel combination is 34*34*16= 18511.

Characters can be in one of the following way

- A stand-alone vowel or a consonant
- A consonant modified by a vowel
- A consonant modified by one or more consonant and a vowel

Almost all South Indian languages are with curve shape. The curve shape and 18496 combinations of Kannada characters are made lot of difficulties for character segmentation as well as for character recognition. Some of the complex characters are listed below to show the complication of the segmentation (Characters of subscripts).

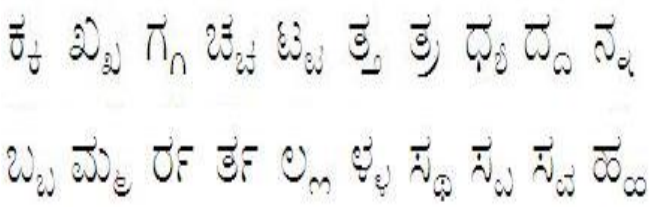

Fig.1.The conjunct consonant (Subscript/Vattu)

## III. DATABASE GENERATION DETAILS

### *A. Data Collection*

The most robust database for Kannada handwritten characters is not available therefore, our own database is created. Data samples are collected from persons of various ages, occupation and sex. A4 size sheet with boxes are drawn, and writers are requested to write using black gel pen especially concentrating on subscript (Vattu).The only restriction was that the character should not touch the boundary of the boxes on the sheet. We have taken dataset from 25 people. We have given two paper sheets for each person to take data from them. Each paper sheet contains 65 words along with subscript and person has to write each character at two times so that variation can be included in

that. Fig. 2 shows the sample of handwritten data written by a writer.

*B. Data Preparation*

The A4 size paper sheet having the data written by various writers (Fig. 2) is digitized using Canon Canoscan Lide 100 flatbed scanner at 300 dpi and stored that document in TIFF file format for the better resolution of image and store it in separate folder. Then we have cropped each word from the scanned document and saved cropped image in TIFF file format.

*C. Image Cropping*

For image cropping we have used Paint Tool. While working with Paint tool we have noticed that it is quite time consuming as each time image need to cropped then need to copy and then paste in new file. Finally need to save in TIFF format.

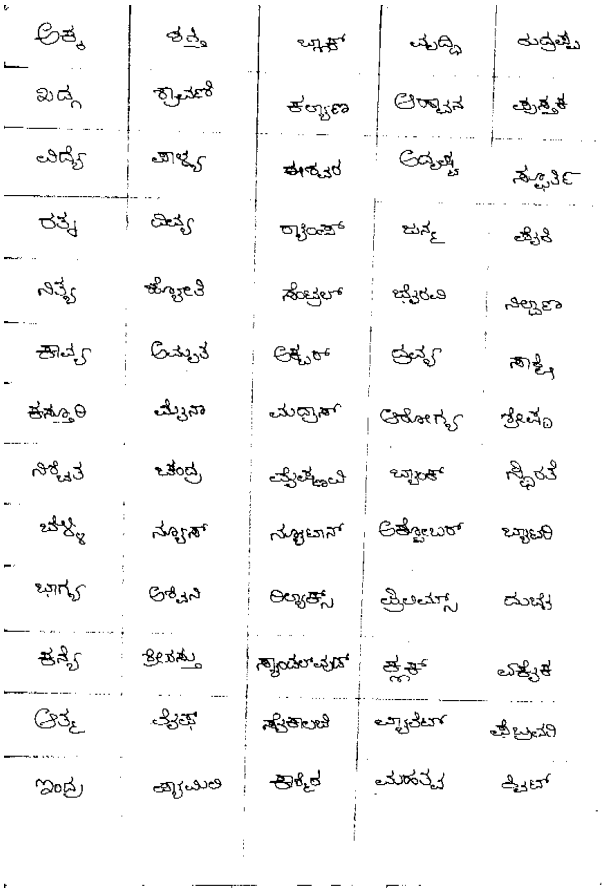

Fig.2. Scanned Document

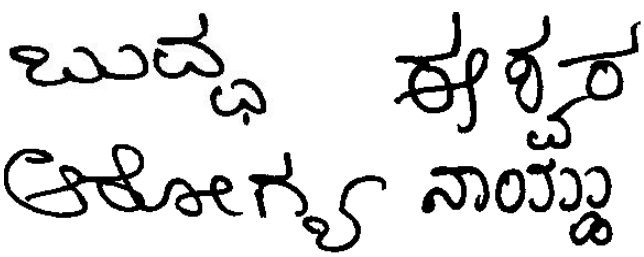

Fig.3. Cropped Images

## IV. PREPROCESSING

Preprocessing technique is used to do improvement of image data that enhances some image features required for processing and suppresses unwanted noise and distortion from image data and aims to correct degradation in an image.

*A. Binarization:* Binarization is the process of converting grayscale image in to binary (Black and White) image, so that image data will only contain 0 and 1.

*B. Noise Removal:* Digital image consist of variety of noises. These noises are required to be removed from an image for better processing. Morphological operation, Median filter and Weiner filter is used to remove noise from an image. Median filter reduces blurring of edges.

*C. Thinning and Filling:* Smoothing implies both filling and thinning. Thinning reduces width of character while filling eliminates gaps, small breaks and holes in digitized character.

*D. Normalization:* To obtain characters of uniform size, rotation and slant normalization is applied on images. It consist fragmentation. Fragmentation it reduce s the un-necessary part of the character where it has two types horizontal and vertical fragmentation.

*E. Skew Detection and Correction:* During the digitization of document page it is often that image is not align correctly or it may be happen by human while writing document. To make in correctly align skew detection and correction technique is used. Skew detection technique can be classified in to groups: Analysis of projection profile, Hough transform, clustering, connected component and correlation between line techniques.

## V. LITERATURE SURVEY

Dinesh Acharya U et.al [1] presented a automatic recognition of handwritten Kannada numerals based on structural features. Profile based 10-segment string, water reservoir; vertical and horizontal strokes, end points and average boundary length from the minimal bounding box are used as the structural features. These multiple features are combined to form a single feature vector. Fuzzy K-nearest neighbor (fuzzy KNN) classification is used in the recognition stage.

Manjunath Aradhya V N et.al [2] authors proposed unconstrained handwritten Kannada character recognition based on FLD (Fisher Linear Discriminate Analysis).the method extracts features from FLD, 2D-FLD and Diagonal FLD. For classification purpose, authors explored different distance measure techniques and tested there superiority on unconstrained handwritten Kannada characters. The system showed feasibility and effectiveness of method.

Sangame S.K et.al [3] reported an unconstrained handwritten Kannada vowel recognition based on the invariant moments. The proposed method extracts invariant moments feature from zoned images. KNN classifier and Euclidian distance criterion are the classifiers used to

recognize the handwritten Kannada vowels.1625 data samples are used for the experiment.85.53% accuracy is obtained.

Leena R Ragha et.al [4] to recognize a kagunita, authors needs to identify the vowel and the consonant dynamically preprocessed original image. Moments and statistical features are extracted from original images, directional images and cut images. These features are used for both vowel and consonant recognition on multi layer perceptron with back propagation neural network. The recognition result for vowels are average 86% and consonants are 65% when tested on separate test data.

Thungamani.M et.al [5] the objective of this paper is to describe an OCR system for handwritten text documents in Kannada. The system first extracts character and a set of features are extracted from the character image using Zernike moments. The recognition is achieved using Support Vector Machine (SVM). The recognition is independent of the size of the handwritten text and the system is seen to deliver reasonable performance. The recognition rate achieved 94 %.

B.V.Dhandra et.al [6] in this paper author proposed zone based features is extracted from handwritten Kannada vowels and English uppercase character images for their recognition. A total of 4,000 handwritten Kannada and English sample images are collected for classifications. The normalized images are divided into 64 zones and their pixel densities are calculated. These 64 features are submitted to KNN and SVM classifiers. The recognition accuracy of 92.71% for KNN and 96.00% for SVM classifiers are achieved in case of Kannada vowels and 97.51% for KNN and 98.26% for SVM classifiers are achieved in case English uppercase alphabets.

Suresh Kumar D.S et.al [7] in this paper, authors recognized handwritten Kannada characters using feed forward neural networks. A handwritten Kannada character is resized into 20x30 pixels. The resized character is used for training the neural network. Once the training process is completed the same character is given as input to the neural network with different set of neurons in hidden layer and their recognition accuracy rate for different Kannada characters has been calculated and compared.

present in the kagunita character image. In this paper, authors investigate the use of moment's features on Kannada kagunita. Kannada characters are curved in nature with some symmetry observed in the shape. So authors are finding 4 directional images using Gabor wavelets from the

Mamatha H R et.al [8] in this paper, a segmentation scheme for segmenting handwritten Kannada scripts into lines, words and characters using morphological operations and projection profiles is proposed. The method was tested on totally unconstrained handwritten Kannada scripts. Usage of the morphology made extracting text lines efficiently by an average extraction rate of 94.5%.

## VI. CONCLUSION

In this paper we made a survey on offline handwritten character recognition and dataset collection. We collected data samples from 25 writers of different age, sex, and group. Total of 1000 words are collected. Compared to other languages handwritten Kannada character recognition is very difficult task in that also recognizing the subscripts (Vattu) is very challenging. This survey provides brief information about Kannada handwritten character recognition and dataset preparation.

TABLE 1.Literature survey Summery

| Sl. No. | Preprocessing | Segmentation | Feature Extraction | Classification | Accuracy |
|---|---|---|---|---|---|
| 1 | It involves elimination of noise in the document | The original image is cut by some percentage from top, right and bottom directions | Moment features from original and directional images | Multi-layer perceptron with back propagation neural network is used | The recognition results for vowels are average 85% and that of consonants are 59% |
| 2 | RGB image is converted to gray scale and then gray scale image is converted to | Three classes, i.e. base characters, modifier Glyphs and subscripts which are recognized separately | K-means clustering is used for extraction | SVM approach is used for classification | The recognition rate is 94.76% in K-means |
| 3 | It involves noise reduction, slant correction, size normalization and thinning | The character/numeral image(50×500 is divided into 25 equal zones(10×10 | Zone and distance metric based feature extraction system | Feed forward back propagation neural network is used | 98% and 96% recognition rate for Kannada and Telugu numerals respectively |
| 4 | Histogram based global binarizing algorithm is used to convert gray image to two-tone images | Bounding box of a character is segmented into blocks | Directional chain code information of the contour points of the characters is used | Quadratic classifier based scheme | 97.87% and 98.45% recognition accuracy using 64 dimensional and 100 dimensional features respectively |
| 5 | Preprocessing stage involves noise reduction, slant correction, size normalization and thinning | Novel character segmentation method using Gabor filters method based on the analyzing the vertical projection of a character is developed to find column index | The features that are used to form a signature are direction of the stroke, density of the stroke and number of clicks for the character | KNN is used for classification | Accuracy is of 94.4% |

## REFERENCES

[1] Dinesh Acharya U., N.V. Subba Reddy, and Krishnamoorthi Makkithaya,"Multilevel Classifiers in Recognition of Handwritten Kannada Numerals", World Academy of Science, Engineering and Technology Vol. 2 2008-06-20

[2] Manjunath Aradya V.N, Kumar G.H, Noushath.S, Shivakumar P" Fisher Linear Discriminate Analysis Based Technique Useful for Efficient Character Recognition" Intelligent Sensing and Information Processing, 2006. Fourth International Conference on Oct. 15 2006-Dec. 18 2006

[3] Sangame S.K., Ramteke R.J., Rajkumar Benne," Recognition of Isolated Handwritten Kannada Vowels", Advances in Computational Research, ISSN: 0975–3273, Vol. 1, Issue 2, 2009, pp-52-55

[4] Leena Ragha, M. Sasikumar, "Adapting Moments for Handwritten Kannada Kagunita Recognition", Second Inter-national Conference on Machine Learning and Computing, IEEE Computer Society, Washington, DC, USA, pp. 125-129, 2010

[5] Thungamani M, Dr.RamakhanthKumar, P KeshavaPrasanna, Shravani Krishna Rau "Off-line Handwritten Kannada Text Recognition using Support Vector Machine using Zernike Moments", International Journal of Computer Science and Network Security, Vol. 11 No.7, July 2011

[6] B.V.Dhandra, Gururaj Mukarambi, Mallikarjun Hangarge," Kannada and English Numeral Recognition System", International Journal of Computer Applications (0975 – 8887) Vol. 26– No.9, July 2011

[7] Suresh Kumar D S, Ajay Kumar B R, K Srinivasa Kalyan "Kannada Character Recognition System Using Neural Network" International Journal of Internet Computing "ISSN No: 2231 – 6965, Vol. 1, ISS- 3 2012

[8] Mamatha H.R, Srikantamurthy K."Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document", International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Vol. 4– No.5, and October 2012

[9] Shraddha S. Gundal & P. P. Narwade, "Handwritten character recognition using neural network with four, eight & twelve directional feature extraction techniques", International Journal of Electronics, Communication & Instrumentation Engineering Research and Development (IJECIERD) ISSN(P): 2249-684X; ISSN(E): 2249-7951 Vol. 4, Issue 2, Apr 2014, 5-12

[10] S. V. Rajashekararadhya, P. Vanaja Ranjan, "Neural Network Based Handwritten Numeral Recognition of Kannada and Telugu Scripts", TENCON 2008 IEEE Region 10 Conference, 19-21 Nov 2008