Segmentation and Recognition of Handwritten Kannada Text Using Relevance Feedback and Histogram of Oriented Gradients – A Novel Approach

Karthik S
Research Scholar, Bharathiar University &
Department of ISE, P E S Institute of Technology-BSC
Bangalore, India

Srikanta Murthy K
Department of Computer Science & Engineering
P E S Institute of Technology Bangalore South Campus
Bangalore, India

***Abstract*—India is a multilingual country with 22 official languages and more than 1600 languages in existence. Kannada is one of the official languages and widely used in the state of Karnataka whose population is over 65 million. Kannada is one of the south Indian languages and it stands in the 33rd position among the list of widely spoken languages across the world. However, the survey reveals that much more effort is required to develop a complete Optical Character Recognition (OCR) system. In this direction the present research work throws light on the development of suitable methodology to achieve the goal of developing an OCR. It is noted that the overall accuracy of the OCR system largely depends on the accuracy of the segmentation phase. So it is desirable to have a robust and efficient segmentation method. In this paper, a method has been proposed for proper segmentation of the text to improve the performance of OCR at the later stages. In the proposed method, the segmentation has been done using horizontal projection profile and windowing. The result obtained is passed to the recognition module. The Histogram of Oriented Gradient (HoG) is used for the recognition in combination with the support vector machine (SVM). The result is taken as the feedback and fed to the segmentation module to improve the accuracy. The experimentation is delivered promising results.**

Keywords*—Optical character recognition; Histogram of oriented gradients; relevance feedback; segmentation; Support Vector Machine; handwritten Kannada documents

I. INTRODUCTION

Optical character recognition (OCR) refers to a process of transforming the images of either handwritten or printed document to a machine readable and editable format. In general, all OCR systems have the following stages: image preprocessing, segmentation, extraction of features and finally recognition of characters. The results of each of these stages are greatly affected by the performance of the previous stages. To make the results of the subsequent stages more accurate, segmentation plays an important role. The extraction of region of interest from the given image is termed as segmentation. In the segmentation of document images, first we extract the lines then the words and finally the characters. Segmentation of characters from a document is still a open challenge in the are of developing efficient OCR systems.

Because of the large dataset and structural complexity, the development of OCR for some of the Indian languages like kannada and telugu is considered to be a tedious task [1]. To add to these complexities in some cases the characters may overlap with each other. In spite of several attempt, the development a high accuracy OCR system for all the Indian languages is still a open challenge. The rest of the paper is organized as follows: In section II a brief discussion about the previous work is reported, the proposed method details can be found in section III. Section IV discusses the experiments and results followed by the conclusion in section V.

II. LITERATURE

In the recent past, due to the existence of digital library of India, the amount of document images for various Indian languages has grown tremendously. The library has taken care to collect documents from different sources and also retaining the original structure, size, font etc. Developing a robust OCR to handle all these issues is still a open challenge. In [1] the authors have highlighted the complexities involved in the segmentation of handwritten documents for some of the south Indian languages like tamil, telugu and malyalam. The existence of the curved characters poses special challenges in the segmentation process. Different strategies like Graph based, Hough transform based, and projection based techniques are proposed for the segmentation of the documents [2]. Arivazhagan et al. [3] proposed the projection-based algorithm in which first obtains candidate lines from the piece-wise projection profile of the document .The lines traverse around any obstructing handwritten connected component by associating it to the line above or below. The author claims that the proposed method is invariant to the skew present in the documents. A level set based new approach for the text line segmentation was proposed by Li et al [4].

In [5] a grouping approach for segmentation was suggested in which a block of connected components are grouped together to identify the characters in a text document. But this approach cannot be used on degraded documents as claimed by the authors. A combination of iterative hypothesis validation through hough transformation and connected components was proposed in [6]. This method is found to be effective in skewed

documents. A peak fringe number (PFN) based approach for segmentation is proposed in [7]. Here the author compute the fringe map for the text document and from that they calculate the PFN. This is used to perform the line segmentation. A method based on separation of header line, base line and contour is presented in [8] for the handwritten Hindi text documents. The authors have claimed that this method is invariant to non-uniform skew in the document. In [9&10] segmentation of English characters where proposed based on the skeletonization methods. A combination of horizontal and vertical projection profiles method is presented in [11] for the Gurumukhi handwritten characters.

It is observed from the literature survey that a lot of work is done for languages like Chinese and English where as a very few work is reported for some of the south Indian languages like kannada, tamil and telugu. This served as motivation for us to develop an efficient segmentation and recognition method for the handwritten kannada documents. A sample of kannada vowels and consonants are shown in figure 1 and 2 respectively.

ಅ ಆ ಇ ಈ ಉ ಊ ಋ ಎ ಏ ಐ ಒ ಓ ಔ

Fig. 1. Kannada vowel samples

ಕ ಖ ಗ ಘ ಙ

ಚ ಛ ಜ ಝ ಞ

ಟ ಠ ಡ ಢ ಣ

ತ ಥ ದ ಧ ನ

ಪ ಫ ಬ ಭ ಮ

ಯ ರ ಲ ವ ಶ ಷ ಸ ಹ ಳ

Fig. 2. Kannada consonants samples

III. Proposed Method

In this paper, we propose a relevance feedback based approach for the segmentation of handwritten kannada documents. Traditionally in all the optical character recognition system, the output of the segmentation process is fed as input for the recognition phase. If, the sample is wrongly segmented then the recognition system fails to recognize such samples. However, this information is not communicated back to the segmentation phase. In our proposed method we have attempted to fix this gap between the segmentation and recognition phases.

A. Segmentation module

Firstly, we extract the lines from the input document using horizontal projection profile method proposed in [12]. Here we compute the number of ON pixels along the row in the image. A very minimum number of ON pixels represent the rows with no contents. This will help us to identify starting and end of the lines and hence we can extract the lines from the documents. Once we extract the lines in the next step we try to extract the characters from each line. We have employed an adaptive window based technique to extract the characters. At the beginning, the width of the window is initialized to a predefined quantity. From empirical data we have identified this predefined width of characters. Using this window, we extract the character and this is passed to the recognition module for the feedback. If the recognition module can correctly classify the sample then, we consider the character is correctly segmented. If the recognition module is not able to identify the character, then this information is communicated back to the character segmentation phase and the window width will be increase by 'x' quantity and again the character segmentation will be done. This process will be continued till either the character is correctly classified or till the width of the window reaches the double the initial size. These steps are summarized in the following algorithm:

Input: Handwritten kannada document image of size mXn

Output: Set of segmented characters

Step 1: If the image is in RGB format then convert it to monochrome image using otsu's thresholding method

Step 2: Calculate the number of On pixels in every row. Any valvue close to zero represent the discontinuity. This will help to identify the start and end of height of a line. Using this information perform line segmentation.

Step 3: For every line do

Step 3.1: Initialize the window width to a predefined value

Step 3.2: Extract a character using this window

Step 3.3: Pass the segmented image to the recognize module

Step 3.4: If the root mean square error between the training and segmented sample is less than a predefined threshold then accept the segmented image else return negative acknowledgment to the segmentation module

Step 3.5: If negative acknowledgement is received then increase the window width by 'x' quantity. If window width is less than 2 times the original window width then proceed to step 3.2 else proceed to step 3.1.

B. The Recognition Module

The relevance feedback is provided by this character recognition module. The performance of the segmentation is dependent on the performance of this recognition module. The accuracy of the recognition system depends on the features that we extract from the image. To achieve high accuracy, we preferred to use histogram of oriented gradients, which is believed to be free from illumination changes and shading [13].

Navneed dallal et. al proposed histogram of oriented gradient descriptors which are widely used in various applications of image processing and computer vision. The basic concept behind the HoG is that the shape of the object within an image can be easily captured by the distribution of edge directions. To implement this, we divide the images into smaller connected regions called as cells. We calculate the gradient directions for each pixel in each cell. To enhance the performance and to make it invariant to illumination changes, we contrast normalize the local histograms by calculate the measure of intensity across a larger part of image known as blocks. The procedure to extract the HoG descriptor is

described in the next algorithm. We have used the support vector machines (SVM) for the classification of the samples.

Input: segmented sample

Output: HoG descriptor

Step1: Gradient computation: We can calculate the gradient magnitude and direction for a given image I, as

Gradient, $$G = \sqrt{I_x^2 + I_y^2} \quad (1)$$

And orientation of gradient, theta $$= \arctan \frac{I_x}{I_y} \quad (2)$$

Where Ix and Iy are obtained by convolving the given image I with the masks Dx = [-1 0 1] and

Dy= $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ respectively.

Step2: Orientation binning: In this step, we calculate the cell histogram. Each pixel that belongs to the cell, cast its weighted vote for the orientation based histogram. We have used unsigned gradiends and a total of 9 bins for the histogram channels. The weight of the vote depends on the gradient magnitude.

Step 3: Obtaining the HoG descriptor:

In order to nullify the effect of illumination and shading the cell histograms obtained in step 2 need to be normalized. The normalization is done based on the overlapping blocks. The normalized cell histogram values are represented in the form of a vector and this is called as HoG descriptor.

IV. Experiments and Results

The method was tested on the standard dataset, Kannada Handwritten Text Document (KHTD) Dataset which was proposed in [14]. The authors have considered four different category of kannada text. They are related to sports, medical documents, movies and general news. The data is collected from 51 individuals who belongs to different age groups and have different educational qualifications. The data was captured in unruled A4 size papers and the authors were free to choose the type of pens. On an average there were 21 lines per every document. The collected documents are then scanned using a flat bed scanner at a resolution of 300 dpi. We have considered 200 such documents for the experimentation. An average segmentation accuracy of 94% is achieved by the proposed approach. The comparison of the proposed method with few existing methods is shown in the table 1. An example of the original image, line segmented image and character segmented image is shown in figure 3,4 and 5 respectively.

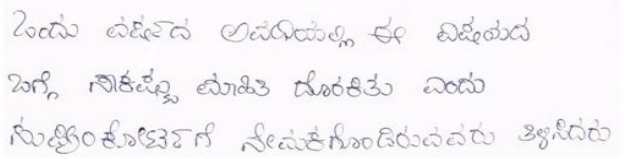

Fig. 3. Original handwritten kannada document image

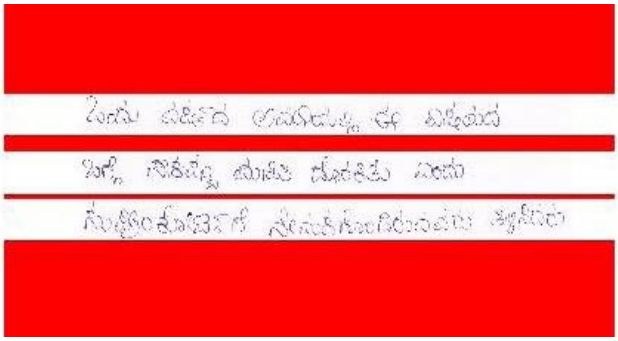

Fig. 4. Line segmented image

Fig. 5. Character segmentation

TABLE I. Comparison of the Method with the Existing Methods

Author	Method	Dataset size	Accuracy
Aradhya et al.,[15]	Component extension technique	250	Not specified
Alaei et al.,[14]	Potential Piece-wise Separation Line technique	204	94.98%
Mamatha et al [16]	Hough Transform	100	91%
Proposed Method	Adaptive window sizing and histogram of oriented gradient	200	94%

The performance of the histogram of oriented gradient descriptor was also evaluated on a test dataset consisting of 18800 samples. The recognition accuracy of vowels and consonants are shown in table 2 and 3 respectively. Table 4 indicates the comparison of the proposed method with the existing methods.

TABLE II. Performance of the Method on Vowels

Vowels in sequence as per Fig.1	Recognition Accuracy (%)	Vowels in sequence as per Fig.1	**Recognition Accuracy (%)**
V1	93.75	V8	93.75
V2	95	V9	96.25
V3	96.25	V10	97.5
V4	96.25	V11	96.25
V5	97.5	V12	96.25
V6	95	V13	97.5
V7	97.5		

TABLE III. PERFORMANCE OF THE METHOD ON CONSONANTS

Consonants in sequence as per Fig.2	Recognition Accuracy (%)	Consonants in sequence as per Fig.2	Recognition Accuracy (%)
C1	95	C18	97.5
C2	96.25	C19	93.75
C3	96.25	C20	96.25
C4	97.5	C21	97.5
C5	95	C22	96.25
C6	97.5	C23	96.25
C7	93.75	C24	97.5
C8	96.25	C25	95
C9	97.5	C26	96.25
C10	96.25	C27	96.25
C11	96.25	C28	97.5
C12	97.5	C29	95
C13	95	C30	97.5
C14	96.25	C31	97.5
C15	96.25	C32	96.25
C16	97.5	C33	97.5
C17	95	C34	97.5

TABLE IV. RESULT COMPARISON

Authors	No. of samples in data set	Feature extraction method	Accuracy (%)
Rajashekhara aradhya S V et.al[17]	1000	Vertical projection distance with zoning	93
R Sanjeev Kunte et al[18]	2500	Wavelet	92.3
V N Manjunath Aaradhya et al[19]	2000	Radon features	91.2
Dinesh Aacharya U et. al[20]	500	Structural features	90.5
Proposed	18,800	HOG	95.02

V. CONCLUSION

A new relevance feedback based approach for character segmentation and recognition in the handwritten kannada documents is proposed in this paper. The method was tested against a standard dataset called KHTD. We have achieved a segmentation accuracy of 94% and a recognition accuracy of 95.02% .The results are compared with the existing methods and found to be promising. However, the method does not address the segmentation of the touching characters in the document.

REFERENCES

[1] K.S. Sesh Kumar, A.M. Namboodiri, and C.V. Jawahar, “Learning Segmentation of Documents with Complex Scripts”, In Proc. of Computer Vision, Graphics and Image Processing 2006, LNCS 4338, pp. 749–760, 2006

[2] Z. Razak, K. Zulkiflee , M.Y. I. Idris, E. M. Tamil, M. N.M. Noor, R. Salleh, M. Y. Z.M. Yusof and M. Yaacob, ”Off-line Handwriting Text Line Segmentation : A Review “,In International Journal of Computer Science and Network Security, vol.8 ,no.7, pp 12-20, July 2008.

[3] M. Arivazhagan, H. Srinivasan and S. N. Srihari, “A Statistical Approach to Handwritten Line Segmentation”, In Proc. of SPIE Document Recognition and Retrieval XIV , San Jose, CA, February 2007.

[4] Y. Li, Y. Zheng , D. Doermann and S.Jaeger, “A New Algorithm for Detecting Text Line in Handwritten Documents”, In Proc. of the International Workshop on Frontiers in andwriting Recognition, pp. 35–40, 2006.

[5] L. Likforman-Sulem and C. Faure, “Extracting Text Lines in Handwritten Documents by Perceptual Grouping”,In Advances in Handwriting and Drawing: A Multidisciplinary Approach, pp. 21-38, 1994.

[6] L. Likforman-Sulem, A. Hanimyan and C. Faure, “A Hough based Algorithm for Extracting Text Lines in Handwritten Documents”, In Proc. of the Third International Conference on Document Analysis and Recognition, vol. 2, pp. 774- 777, August 1995.

[7] V. K. Koppula and A. Negi , “Using Fringe Maps for Text Line Segmentation in Printed or Handwritten Document Images” ,In Proc. of the Second Vaagdevi International Conference on Information Technology for Real World Problems, pp83-88,2010.

[8] N. K. Garg , L. Kaur and M. K. Jindal, “A New Method for Line Segmentation of Handwritten Hindi Text”, In Proc. of Seventh International Conference on Information Technology: New Generations, pp 392-397,2010

[9] Choudhary, A., Rishi, R. & Ahlawat, S., “ A New Character Segmentation Approach for Off-Line Cursive Handwritten Words”, Procedia Computer Science, Volume 17, pp. 88-95, 2013

[10] Lacerda, E. B. & Mello, C. A., “ Segmentation of connected handwritten digits using Self-Organizing Maps”, International conference on Expert Systems with Applications, November, pp. 586-587, 2013

[11] Mahto M K, Bhatia K, Sharma R K, “Combined horizontal and vertical projection feature extraction technique for Gurmukhi handwritten character recognition”, International conference on advances in computer engineering and applications, pp. 59-65, 2015

[12] Mamatha H R, Srikantamurthy K, “Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document”, International Journal of Applied Information Systems, Volume 4– No.5,pp. 13-19, October 2012

[13] N. Dalal, B. Triggs, “Histogram of oriented gradients for human detection”,International conference on computer vision and pattern recognition, pp. 886-893, 2005

[14] A. Alaei, P. Nagabhushan and U. Pal, “A Benchmark Kannada Handwritten Document Dataset and its Segmentation”, In Proc. of the International Conference on Document Analysis and Recognition, pp 141-145, 2011

[15] V. N. M. Aradhya and C. Naveena, “Text Line Segmentation of Unconstrained Handwritten Kannada Script”, In Proc. of the 2011 International Conference on Communication, Computing & Security, pp. 231-234, ACM, 2011

[16] H R Mamatha and K S Murthy, “Skew Detection, Correction and Segmentation of Handwritten Kannada Document”, In International Journal of Advanced Science and Technology, ISSN: 2005- 4238, Science and Engineering Research Support society, vol. 48, pp.71-88, November 2012.

[17] S.V. Rajashekararadhya P. Vanaja Ranja n,: Neural Network Based Handwritten Numeral Recognition of Kannada and Telugu scripts, TENCON, 2008

[18] R Sanjeev Kunte and Sudhakar Samuel R.D,: Script Independent Handwritten Numeral recognition, International conference on visual information engineering, pp 94-98, 2006

[19] V. N. Manjunath Aradhya, G. Hemanth Kumar and S. Noushath,: Robust Unconstrained Handwritten Digit Recognition Using Radon Transform, Proc. of IEEE International conference on signal processing, communication and networking , pp-626-629, 2007

[20] Dinesh Acharya U, N. V. Subba Reddy and Krishnamurthy,: Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster, pp.125 - 129, IISN, 2007