PARAGRAPH SEGMENTATION OF HANDWRITTEN DOCUMENT WITH SKEWED CONTENT

Veerappa B. Pagi[1], Bhargav H. K[2]. and Ramesh S. Wadawadagi[3]

1,3 Basaveshwar Engineering College, Bagalkot

2 SKSVM Agadi College of Engg & Technology, Lakshmeshwar.

Abstract

Kannada Handwritten document paragraph segmentation is a complex process which involves identification of size, style and pen used for writing. Therefore, it becomes crucial part of converting document into paragraph segmentation. A database of Kannada handwritten document is created and the same has been used for research purpose. In this paper,atmost focuss is laid on Kannada handwritten paragraph segmentation with deskewed content.Binarization method of preprocessing technique is used and Run Length Smoothing Algorithm(RLSA) to tackle the problem of segmenting Kannada handwritten documents.

Keywords: Segmentation,RLSA

1. Introduction

Kannada is thought to be one of the oldest language and It has long history associated with the same.It is described as queen of scripts.It is a language which is composed of rich poetry ,prose,drama, and criticism. Every language is expressed in the form of paragraph while communicating the content of the same.While expressing the content in handwritten form,the writer is bound to come in the way of skewed handwriting.In order to overcome the abouve mentioned problem of skewed handwriting paragraph segmentation with deskewed content is used. Segmentation is the process in which the lines are divided into words and the words further into characters its parts. The Run Length Smoothing Algorithm (RLSA) is a technique used for Kannada handwritten paragraph segmentation and text discrimination. The technique developed for the Document Analysis System consists of two steps. First, a segmentation procedure categorizes the area of a document into paragraphs, each of which should possesses only one type of data . Next, some basic features of these blocks are computed The RLSA is applied row-by-row as well as column-by-column to a Kannada handwritten document, producing two distinct bit-maps. Because spacings of document components tend to differ horizontally and vertically, different values of predefined limit is used for processing of row and columns. Then two bit-maps are then combined in a logical AND operation. Additional horizontal smoothing using the RLSA technique produces the paragraph segmentation with skewed content.

2. Literature Survey

Various Kannada document image segmentation techniques have been proposed in the literature. These techniques can be categorized based on the document image segmentation algorithm that they adopt. The most known of these segmentation algorithms are the following: X–Y cuts(projection profiles) based , Run Length Smoothing Algorithm (RLSA) , component grouping , document spectrum , whitespace analysis, constrained text lines, Hough transform , Voronoi tessellation and Scale space analysis. All of the above segmentation algorithms are mainly designed for contemporary documents. projection profiles, Run Length Smoothing Algorithm , Hough transform and scale space analysis algorithms are mainly used in Kannada handwritten document segmentation,. Table 1 categorizes all of the aforementioned segmentation algorithms and depicts the way they have been used in document processing.

SI.No	Segmentation Algorithm	Handwritten Document	Paragraph Segmentation	Text Line Segmentation
1	Hough transform			✓
2	X-Y Cuts(Projection Profiles)	✓	✓	✓
3	Scale Space Analysis	✓	✓	✓
4	RLSA	✓	✓	✓
5	Docstrum		✓	✓
6	White space Analysis		✓	✓
7	Constrained text lines		✓	✓
8	Voronai		✓	✓

.

Authors have focused on paragraph segmentation with deskewed content with different processing techniques.Some of these are discussed below.

Alireza Alaei.et.al [1] have introduced an unconstrained Kannada handwritten text database (KHTD) is introduced. The KHTD contains 204 handwritten documents of four different categories written by 51 native speakers of Kannada. Total number of text-lines and words in the dataset are 4298 and 26115, respectively. Banumathi. and Jagadeesh Chandra [2] have presented a segmentation system based on Projection Profile method for general handwritten text and historical handwritten text.In[5], Authors have proposed a method for the line extraction and skew correction of the extracted text lines uses a new cost function, which considers spacing between text lines and skew of each text line is used. The problem is formulated as an energy minimization of the cost function yields a set of text lines. In[6], Authors have strived towards the development of efficient techniques in order to segment document pages resulting from the digitization of historical machine-printed sources. These machine-printed documents often suffer from low quality and local skew, several degradations due to the old printing matrix quality or ink diffusion, and exhibit complex and dense layout.In[7],Authors have proposed a segmentation scheme for segmenting handwritten Kannada scripts into lines, words and characters using morphological operations and projection profiles. This proposed method was tested on totally unconstrained handwritten Kannada scripts, which pays more challenge and difficulty due to the complexity involved in the script.In[8] By making use of Classifiers authors focused on recognition of Karnataka district names in Kannada and 20 different English words from a given scanned word image.

In[10] By making use of bounding box technique, Hough transform and contour detection have proposed schemes for skew detection and correction, segmentation of handwritten Kannada document respectively.

In[11] By making use of Water flow technique is used in extraction of text lines authors have presented novel methodology for segmenting handwritten Malayalam documents into its constituent lines, words and characters addressing the issues mentioned.

3. Proposed Methodology

Figure 1 illustrates the proposed methodology for paragraph segmentation of Kannada handwritten document with skewed content.

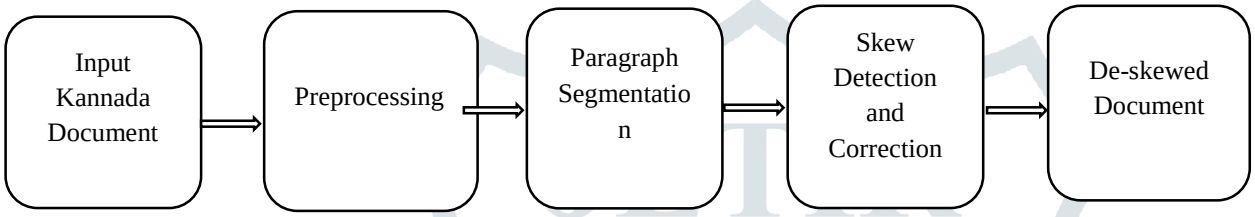
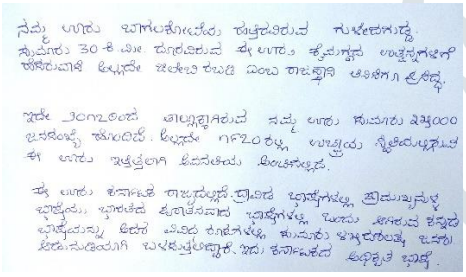
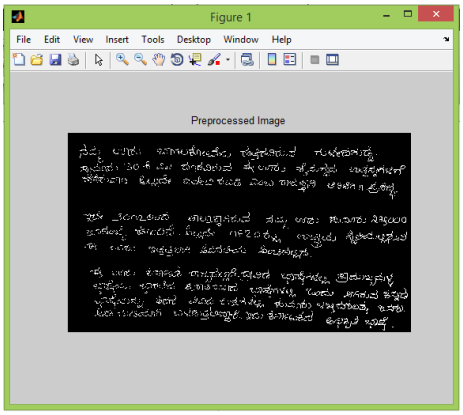
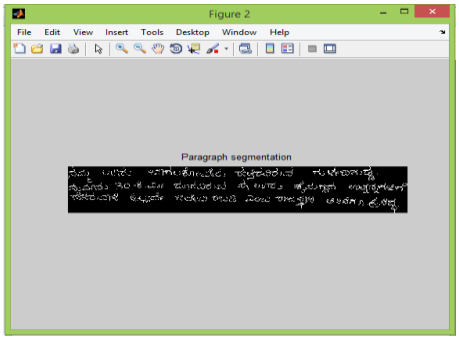


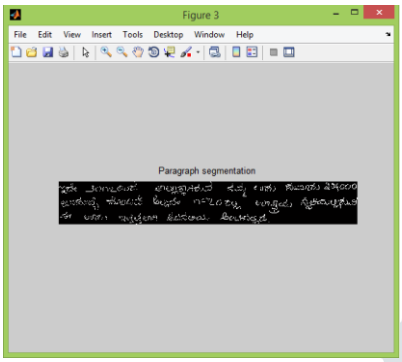
Figure 1: Proposed methodology for paragraph segmentation of Kannada handwritten document with skewed content.

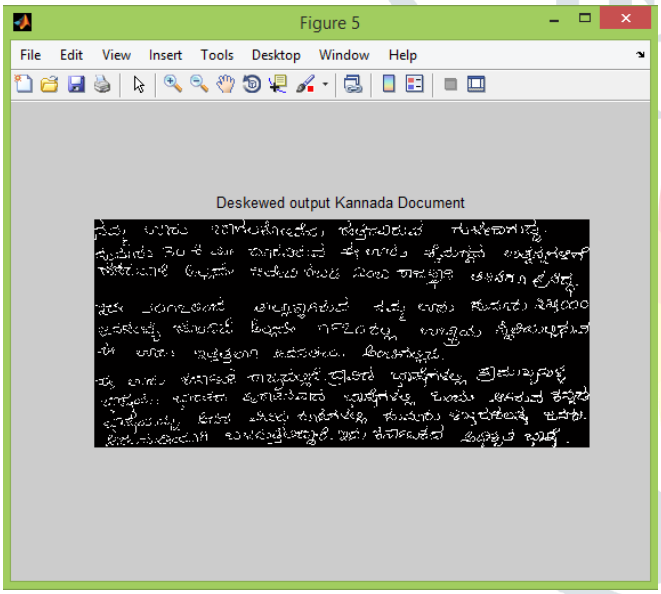
4. Results











5. Research Gap and future direction

The proposed work is able to segment the paragraph in which there is a large gap between the paragraphs.At the same time it is not performing satisfactorily in segmenting the paragraphs which are closely spaced.Future work can be done to segment paragraphs which are closely spaced.Future work can be done to segment paragraphs which are closely spaced using refined features of lines in the paragraph.

6.Conclusion

The algorithm is not performing satisafactorily in case of handwritten Kannada in which letters are closely spaced.It is performing with high efficiency aroung 90% where the letters of Kannada handwritten words are isolated clearly.

References

[1] Alireza Alaei, P. Nagabhushan, Umapada Pal," A Benchmark Kannada Handwritten Document Dataset and its Segmentation", International Conference on Document Analysis and Recognition,2011,pp.142-145

[2] Banumathi. K. L, Jagadeesh Chandra A P," Line and word Segmentation of Kannada HandwrittenText documents using Projection Profile Technique", International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT),2016,pp.196-201

[3] Priyanka Karmakar, Biswajit Nayak, Nilamani Bhoi,"Line and Word Segmentation of a Printed Text Document", International Journal of Computer Science and Information Technologies, 2014,Vol. 5 (1) ,pp.157-160

[4] Satadal Saha, Subhadip Basu, Mita Nasipuri, Dipak Kr. Basu," A Hough Transform based Technique for Text Segmentation", Journal of Computing, Vol. 2,no.2,pp.134-141

[5] Sunanda Dixit,Sures Hosalli Narayan, Mahesh Belur," Kannada Text Line Extraction Based On Energy Minimization And Skew Correction" , International Advance Computing Conference,2014, pp.62-67

[6] Nikos Nikolaou, Michael Makridis, Basilis Gatos, Nikolaos Stamatopoulos, Nikos Papamarkos,"Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths", Image and Vision Computing, 2010,pp.590-604

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

[8] Kumar B.Y , Dr. Keshava Prasanna , Dr Savitha.R ," A Novel Approach on Offline Kannada and English Handwritten Words", International Journal of Scientific Engineering and Applied Science, 2016, Volume-2,no. 8,pp.174-178

[9] Vijaya Kumar Koppula, Atul Negi," Segmentation of Closely set and Touching Lines in Handwritten document images using Fringe Maps", International Conference for Convergence of Technology,2014,pp.1-6

[10] Mamatha Hosalli Ramappa, Srikantamurthy Krishnamurthy," Skew Detection, Correction and Segmentation of Handwritten Kannada Document", International Journal of Advanced Science and Technology,2012,Vol. 48,pp.71-88

[11] Shahnaz Abubakker Bapputty Haji, Ajay James, Dr. Saravanan Chandran ," A Novel Segmentation and Skew Correction Approach for Handwritten Malayalam Documents", Procedia Technology International Conference on Emerging Trends in Engineering, Science and Technology,2015,pp.1341-1348

[12] Ms. Roopa Tonashyal and Mr. Y. C. Kiran, "Offline Handwritten Kannada Character Segmentation and Recognition based on Zoning",International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)Vol.3,2015,pp1107-1114

[13] C. Naveena, V.N. Manjunath Aradhya," Handwritten Character Segmentation for Kannada Scripts", World Congress on Information and Communication Technologies,Trivandrum,2012, pp.144-149.