International Journal of Engineering Research in Electronic and Communication Engineering (IJERECE) Vol 3, Issue 5, May 2016

Comparative Analysis of Segmentation Approaches for the Printed Documents

[1] Pruthvi B K, [2] Pooja A P
[1]PG Scholar, Signal Processing [2] Assistant Professor, Dept. of ECE,
[1][2]Vidya Vardhaka College Of Engineering
Mysore, India
[1]pruthvi.bk@gmail.com [2]poojavvce@gmail.com

***Abstract*—Segmentation is a vital procedure of any Optical Character Recognition (OCR) framework. It isolates the image content documents first into lines then to words lastly to characters. The accuracy of OCR framework for the most part relies on upon the segmentation algorithm being utilized. Segmentation of printed content of some Indian dialects like Kannada, Telugu and Assamese is troublesome when contrasted and Latin based dialects as a result of its auxiliary many-sided quality and expanded character set. It is be partitioned as vowels and consonants which can likewise contain subscripts and conjunct consonants. In spite of a few effective works in OCR everywhere throughout the world, advancement of OCR instruments in Indian dialects is still a progressing process. Character segmentation assumes a vital part in character acknowledgment in light of the fact that erroneously divided characters are unrealistic to be perceived accurately. In this paper, a segmentation plan for dividing printed Kannada scripts into lines and words utilizing Run Length Smoothing Algorithm (RLSA) and Variational Bayes (VB) strategies are proposed and their comparative analysis is carried out.**

Keywords: **Segmentation; OCR; RLSA; VB**

I. INTRODUCTION

Segmentation of content line, words and characters in document image processing is a mind boggling and critical step towards acknowledgment of either written by hand or printed document. Logged off printed content segmentation is a vital pre-essential of Optical Character Recognition (OCR). Segmentation blunders decrease the accuracy of OCR. The segmentation of both printed documents and written by hand content lines is now and again upheld various heuristics and presumptions. While different content line structure based documents can as a rule is effectively isolated content lines, ineffectively composed documents with fluctuating content line slant and covering ascenders or descanters are hard to portion into content lines. Printed document examination may take after numerous ways. In any case, isolated into content lines, then words and for certain situation character is required, with a specific end goal to concentrate acknowledgment components, for example, word length or word minutes. Acknowledgment can be word based or character based. Once the piece of content is isolated into content lines, word segmentation is significantly less demanding. Content line segmentation is typically consolidated with skew detection so writing study is ruined skew detection procedures. Most well known strategies in the writing for skew detection, consequently segmentation, are the strategies taking into account projection profile strategy, associated segment investigation, RLSA, Hough changes and Fourier change based algorithms. The technique displayed in this paper utilizes the RLSA of the content lines for introduction. In this paper, we propose a RLSA and Variation Bays segmentation algorithm, which can be utilized successfully on Kannada printed document images containing a wide range of sorts of covering and touching words in neighboring lines.

The proposed technique of segmentation of printed Kannada handwritten document majorly defines three phases. First phase involves classifying the Pre-Processing, Detection of the Skew and its Correction. The next phase is different levels of segmentation using two different approaches and the final phase is comparative analysis. The paper is categorically structured into sections. The work related to line segmentation of text documents are captured in Section

A brief about Kannada language is provided in section III. The core technique segmentation depicting the methods proposed and the dataset used for experimenting as well as Results comparative analysis is presented in Section IV. The conclusion and summarization for work is described in Section V.

II. RELATED WORK

The procedures of segmentation for documents that are in Asian dialects are moderately basic and numerous methods exist. The test is for portioning the documents that are in

South Indian dialects. If there should be an occurrence of disconnected from the net printed content documents the issue is various and is muddled exclusively by the way of various text style styles. So far numerous examinations works did for segmentation in English, Chinese, Arabic and Devanagari [3] [4] [5]. Zhang and Sang [6] proposed tensor votiing algorithm for Chinese written by hand content line segmentation. In this technique morphological processing is utilized for creating associated part. Evacuation of outliners and distinguishing proof of centroids of the associated part are done utilizing 2-D tensor voting guideline. The course of the vectors and in addition the saliency values portrays data for tensors which can be utilized for segmentation. Gupta et al proposed [10] a philosophy for making line segmentation of a content document thinking about the development of a pen tip. The procedure includes distinguishing all the associating segments and finding their centroids. The centriodal values so found are then connected in view of the guidelines of pen tip algorithm. 2D tensor voting algorithm evacuates plots. The algorithm additionally utilizes the inadequate information focuses to infer unending structures to be specific intersections and bends. The aforementioned strategies work better for content lines that are steady and parallel in a Predefined bearing. These methods have a confinement where the content lines are organized conflictingly and lines situated in various headings.

If there should arise an occurrence of touching content lines and differing skewness, the accuracy is peaceful less. Concentrated examination exertion in the region of line segmentation for literary transcribed documents exists for worldwide dialects English, Asian dialects Chinese and Japanese. According to the writing the majority of the written by hand segmentation work has been completed in Indian Languages like Bangla, Hindi and Gurumukhi. [7][8][9]. Few works have been proposed on south Indian dialects like Kannada and Tamil [11] [12].

III. CHARACTERISTICS OF KANNADA SCRIPT

In this segment, we portray quickly the center qualities in adherence to Kannada script in order to address the troubles of segmentation. Kannada otherwise called Canarese is famous in South India. It is generally talked and proclaimed as official dialect in the condition of Karnataka. Kannada is an antiquated dialect and is gotten from Dravidian scripts. The populace talking the dialect of Kannada adds up to 60 million especially in the Indian conditions of Karnataka. The principle challenges for dividing Kannada scripts are because of many-sided quality of characters. The characters can be any of the sort, standalone vowel, a standalone consonant and a consonant adjusted by a vowel or one or more consonant and a vowel. There are 16 vowels and 36 consonants including conjunct parts (subscript/vatthu) as appeared in Figure 1 and Figure 2 individually.

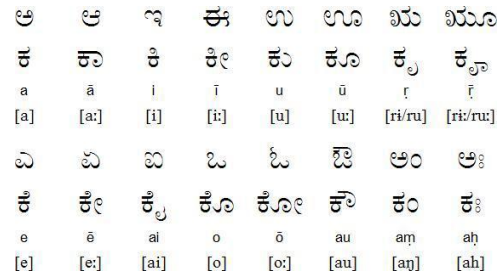


Fig. 1 Vowels and vowel diacritics with ka

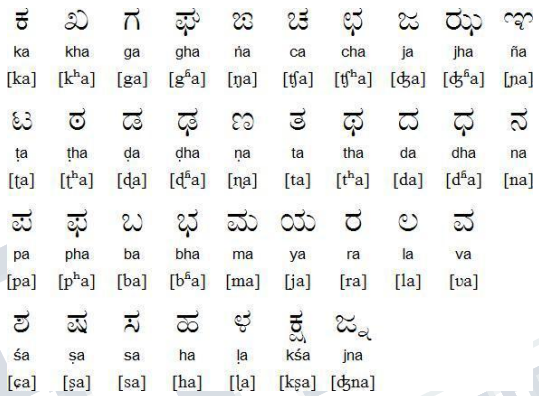


Fig. 2 Consonants

IV. PROPOSED METHODOLOGY

The block diagram in Figure 3 represents the algorithmic flow of the proposed segmentation technique for text documentation. The stages are clearly pipelined. The different process blocks are: (a) pre-processing, (b) skew detection and correction and (c) segmentation. The sub-section explains the process flow across each stage.

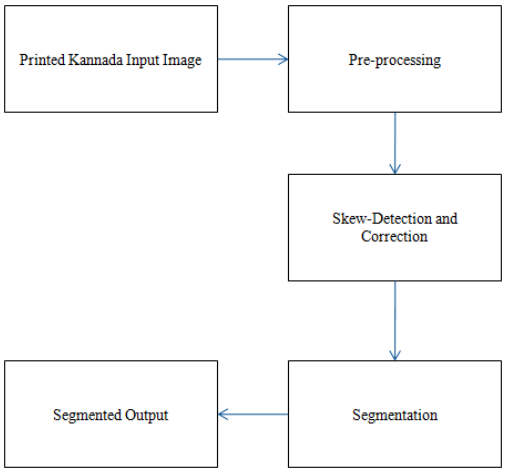
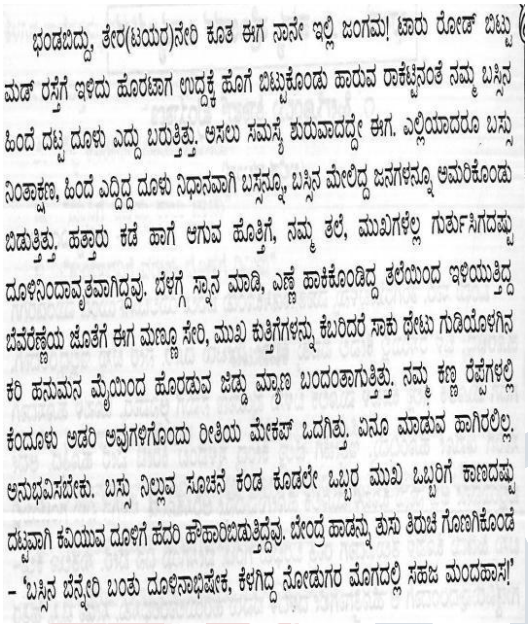


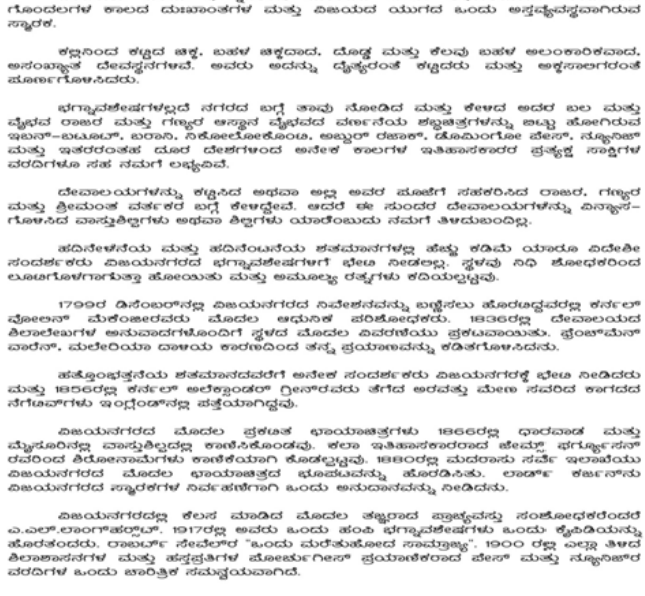
Fig. 3 Block diagram of the proposed algorithm

Firstly, different types of printed kannada database are collected using different sources like textbooks and through online documents. The above mentioned images are scanned and converted to the suitable form according to the requirement. These corresponding images are considered as the input image. The two different types of input images considered are shown in Figure 3 and Figure 4.



ಭಂಡಬಿದ್ದು, ತೇರ(ಟಯರ)ನೇರಿ ಕೂತ ಈಗ ನಾನೇ ಇಲ್ಲಿ ಜಂಗಮ! ಟಾರು ರೋಡ್ ಬಿಟ್ಟು
ಮಡ್ ರಸ್ತೆಗೆ ಇಳಿದು ಹೊರಟಾಗ ಉದ್ದಕ್ಕೆ ಹೊಗೆ ಬಿಟ್ಟುಕೊಂಡು ಹಾರುವ ರಾಕೆಟ್ಟಿನಂತೆ ನಮ್ಮ ಬಸ್ಸಿನ
ಹಿಂದೆ ದಟ್ಟ ದೂಳು ಎದ್ದು ಬರುತ್ತಿತ್ತು. ಆಸಲು ಸಮಸ್ಯೆ ಶುರುವಾದದ್ದೇ ಈಗ. ಎಲ್ಲಿಯಾದರೂ ಬಸ್ಸು
ನಿಂತಾಕ್ಷಣ, ಹಿಂದೆ ಎದ್ದಿದ್ದ ದೂಳು ನಿಧಾನವಾಗಿ ಬಸ್ಸನ್ನೂ, ಬಸ್ಸಿನ ಮೇಲಿದ್ದ ಜನಗಳನ್ನೂ ಆವರಿಸಿಕೊಂಡು
ಬಿಡುತ್ತಿತ್ತು. ಹತ್ತಾರು ಕಡೆ ಹಾಗೆ ಆಗುವ ಹೊತ್ತಿಗೆ, ನಮ್ಮ ತಲೆ, ಮುಖಗಳೆಲ್ಲ ಗುರ್ತಿಸಿಗದಷ್ಟು
ದೂಳಿನಿಂದಾವೃತವಾಗಿದ್ದವು. ಬೆಳಗ್ಗೆ ಸ್ನಾನ ಮಾಡಿ, ಎಣ್ಣೆ ಹಾಕಿಕೊಂಡಿದ್ದ ತಲೆಯಿಂದ ಇಳಿಯುತ್ತಿದ್ದ
ಬೆವರೆಣ್ಣೆಯ ಜೊತೆಗೆ ಈಗ ಮಣ್ಣೂ ಸೇರಿ, ಮುಖ ಕುತ್ತಿಗೆಗಳನ್ನು ಕೆಬರಿದರೆ ಸಾಕು ಥೇಟು ಗುಡಿಯೊಳಗಿನ
ಕರಿ ಹನುಮನ ಮೈಯಿಂದ ಹೊರಡುವ ಜಿಡ್ಡು ಮ್ಯಾಣ ಬಂದಂತಾಗುತ್ತಿತ್ತು. ನಮ್ಮ ಕಣ್ಣ ರೆಪ್ಪೆಗಳಲ್ಲಿ
ಕೆಂದೂಳು ಅಡರಿ ಅವುಗಳಿಗೊಂದು ರೀತಿಯ ಮೇಕಪ್ ಒದಗಿತ್ತು. ಏನೂ ಮಾಡುವ ಹಾಗಿರಲಿಲ್ಲ.
ಅನುಭವಿಸಬೇಕು. ಬಸ್ಸು ನಿಲ್ಲುವ ಸೂಚನೆ ಕಂಡ ಕೂಡಲೇ ಒಬ್ಬರ ಮುಖ ಒಬ್ಬರಿಗೆ ಕಾಣದಷ್ಟು
ದಟ್ಟವಾಗಿ ಕವಿಯುವ ದೂಳಿಗೆ ಹೆದರಿ ಹೌಹಾರಿಬಿಡುತ್ತಿದ್ದೆವು. ಬೇಂದ್ರೆ ಹಾಡನ್ನು ತುಸು ತಿರುಚಿ ಗೊಣಗಿಕೊಂಡೆ
- 'ಬಸ್ಸಿನ ಬೆನ್ನೇರಿ ಬಂತು ದೂಳಿನಾಭಿಷೇಕ, ಕೆಳಗಿದ್ದ ನೋಡುಗರ ಮೊಗದಲ್ಲಿ ಸಹಜ ಮಂದಹಾಸ!'

Fig. 3 Input Image



ಹಂಪಿಯು ಒಂದು ಭಗ್ನಾವಶೇಷ ಹೊಂದಿದ ನಗರ. ಅಭೂತಪೂರ್ವ ರಾಜಕೀಯ ಗೊಂದಲಗಳ ಕಾಲದ ಮಜಾಂತಗಳ ಮತ್ತು ವಿಜಯದ ಯುಗದ ಒಂದು ಅಸ್ಪಷ್ಟವಸ್ಥವಾಗಿರುವ ಸ್ಮಾರಕ.

ಕಲ್ಲಿನಿಂದ ಕಟ್ಟಿದ ಚಿಕ್ಕ, ಬಹಳ ಚಿಕ್ಕದಾದ, ದೊಡ್ಡ ಮತ್ತು ಕೆಲವು ಬಹಳ ಅಲಂಕಾರಿಕವಾದ, ಅಸಂಖ್ಯಾತ ದೇವಸ್ಥನಗಳವೆ. ಅವರು ಅದನ್ನು ದೈತ್ಯರಂತೆ ಕಟ್ಟಿದರು ಮತ್ತು ಅಕ್ಷರಶಾಲಗರಂತೆ ಪೂರ್ಣಗೊಳಿಸಿದರು.

ಭಗ್ನಾವಶೇಷಗಳಲ್ಲದೆ ನಗರದ ಬಗ್ಗೆ ತಾವು ನೋಡಿದ ಮತ್ತು ಕೇಳಿದ ಅದರ ಬಲ ಮತ್ತು ವೈಭವ ರಾಜರ ಮತ್ತು ಗಣ್ಯರ ಆಸ್ಥಾನ ವೈಭವದ ವರ್ಣನೆಯ ಶಬ್ದಚಿತ್ರಗಳನ್ನು ಬಿಟ್ಟು ಹೋಗಿರುವ ಇಬನ್-ಬಟೂಟಾ, ಬರಾನಿ, ನಿಕೋಲೋ ಕೊಂಟಿ, ಅಬ್ದುರ್ ರಜಾಕ್, ಡೊಮಿಂಗೋ ಪೇಸ್, ನ್ಯೂನಿಜ್ ಮತ್ತು ಇತರರಂತಹ ದೂರ ದೇಶಗಳಿಂದ ಅನೇಕ ಕಾಲಗಳ ಇತಿಹಾಸಕಾರರ ಪ್ರತ್ಯಕ್ಷ ಸಾಕ್ಷಿಗಳ ವರದಿಗಳೂ ಸಹ ನಮಗೆ ಲಭ್ಯವಿವೆ.

ದೇವಾಲಯಗಳನ್ನು ಕಟ್ಟಿಸಿದ ಅಥವಾ ಅಲ್ಲಿ ಅವರ ಮೂರ್ತಿ ಸಹಕರಿಸಿದ ರಾಜರು, ಗಣ್ಯರು ಮತ್ತು ಶ್ರೀಮಂತ ವರ್ತಕರ ಬಗ್ಗೆ ಕೆಳಪ್ಪಿವೆ. ಆದರೆ ಈ ಸುಂದರ ದೇವಾಲಯಗಳನ್ನು ವಿನ್ಯಾಸ-ಗೊಳಿಸಿದ ವಾಸ್ತುಶಿಲ್ಪಿಗಳು ಅಥವಾ ಶಿಲ್ಪಿಗಳು ಯಾರೆಂಬುದು ನಮಗೆ ತಿಳಿದುಬಂದಿಲ್ಲ.

ಹದಿನೇಳನೆಯ ಮತ್ತು ಹದಿನೆಂಟನೆಯ ಶತಮಾನಗಳಲ್ಲಿ ಹೆಸ್ಸು ಕಡಿಮೆ ಯಾರೂ ವಿದೇಶಿ ಸಂದರ್ಶಕರು ವಿಜಯನಗರದ ಭಗ್ನಾವಶೇಷಗಳಿಗೆ ಭೇಟಿ ನೀಡಲಿಲ್ಲ. ಸ್ಥಳವು ನಿಧಿ ಹೂಡಕರಿಂದ ಲೂಟಿಗೊಳಗಾಗುತ್ತಾ ಹೋಯಿತು ಮತ್ತು ಅಮೂಲ್ಯ ರತ್ನಗಳು ಕದಿಯಲ್ಪಟ್ಟವು.

1799ರ ಡಿಸೆಂಬರ್‌ನಲ್ಲಿ ವಿಜಯನಗರದ ನಿವೇಶನವನ್ನು ಬಯ್ಟಿಸಲು ಹೊರಡಲ್ಪಟ್ಟರಲ್ಲಿ ಕರ್ನಲ್ ವೋಲನ್ ಮೆಕೆಂಜರವರು ಮೊದಲ ಆಧುನಿಕ ಪರಿಶೋಧಕರು. 1836ರಲ್ಲಿ ದೇವಾಲಯದ ಶಿಲಾಲೇಖಗಳ ಅನುವಾದಗಳೊಂದಿಗೆ ಸ್ಥಳದ ಮೊದಲ ವಿವರಣೆಯು ಪ್ರಕಟವಾಯಿತು. ಫ್ರೆಂಚ್‌ಮನ್ ವಾರೆನ್, ಮಲೇರಿಯಾ ಬಾಧೆಯ ಕಾರಣದಿಂದ ತನ್ನ ಪ್ರಯಾಣವನ್ನು ಕಡಿತಗೊಳಿಸಿದನು.

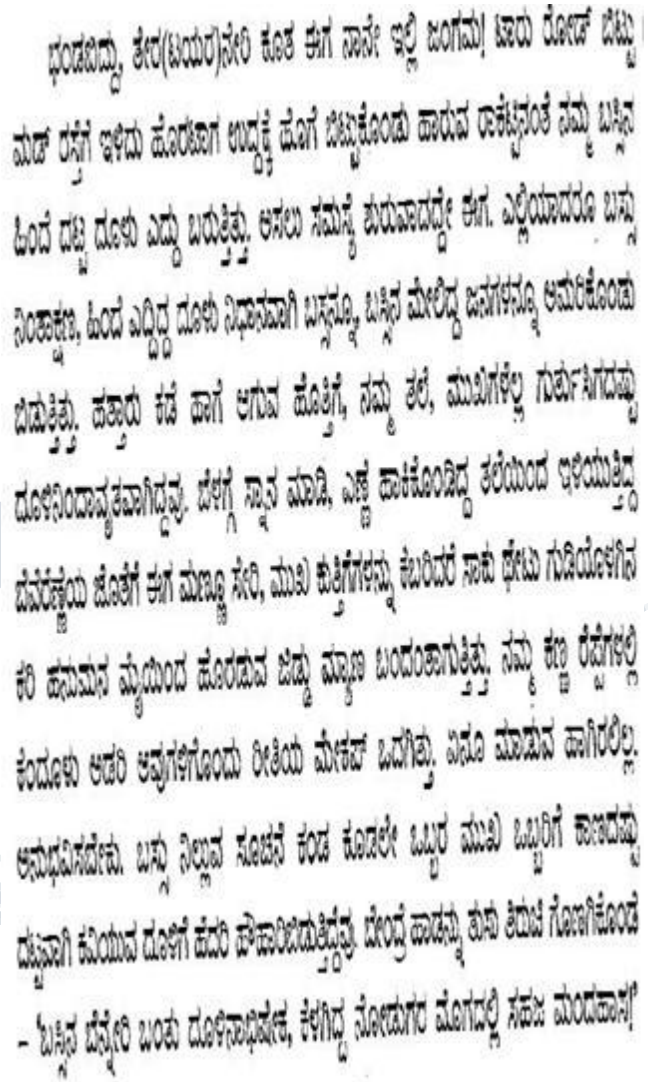
ಹತ್ತೊಂಭತ್ತನೆಯ ಶತಮಾನದವರೆಗೆ ಅನೇಕ ಸಂದರ್ಶಕರು ವಿಜಯನಗರಕ್ಕೆ ಭೇಟಿ ನೀಡಿದರು ಮತ್ತು 1856ರಲ್ಲಿ ಕರ್ನಲ್ ಅಲೆಕ್ಸಾಂಡರ್ ಗ್ರಿಣ್ಣಾವರು ತೆಗೆದ ಅರವತ್ತು ಮೇಣ ಸವರಿದ ಕಾಗದದ ನೆಗೆಟಿವ್‌ಗಳು ಇಂಗ್ಲೆಂಡ್‌ನಲ್ಲಿ ಪತ್ತೆಯಾಗಿದ್ದವು.

ವಿಜಯನಗರದ ಮೊದಲ ಪ್ರಕಟಿತ ಭಾವಚಿತ್ರಗಳು 1856ರಲ್ಲಿ ಧಾರವಾಡ ಮತ್ತು ಮೈಸೂರಿನಲ್ಲಿ ವಾಸ್ತುಶಿಲ್ಪದಲ್ಲಿ ಕಾಣಿಸಿಕೊಂಡವು. ಕೆಲವು ಇತಿಹಾಸಕಾರರಾದ ಜೇಮ್ಸ್ ಫರ್ಗ್ಯುಸನ್ ರವರಿಂದ ತಿರುಚಲಾಪಮೆಗಳು ಕಾಣಿಕೆಯಾಗಿ ಕೊಡಲ್ಪಟ್ಟವು. 1860ರಲ್ಲಿ ಮದರಾಸು ಸರ್ವೇ ಇಲಾಖೆಯು ವಿಜಯನಗರದ ಮೊದಲ ಭಾವಾಚಿತ್ರದ ಭೂಪಟವನ್ನು ಹೊರಡಿಸಿತು. ಲಾರ್ಡ್ ಕರ್ಜನ್‌ನು ವಿಜಯನಗರದ ಸ್ಮಾರಕಗಳ ಸರ್ವಹಣೆಗಾಗಿ ಒಂದು ಅನುದಾನವನ್ನು ನೀಡಿದನು.

ವಿಜಯನಗರದಲ್ಲಿ ಕೆಲಸ ಮಾಡಿದ ಮೊದಲ ತಜ್ಞರಾದ ಪ್ರಾಧ್ಯಾಪನ್ನು ಸಂಶೋಧಕರೆಂದರೆ ಎ.ಎಲ್.ಲಾಂಗ್‌ಹರ್ಸ್ಟ್. 1917ರಲ್ಲಿ ಅವರು ಒಂದು ಹಂಪಿ ಭಗ್ನಾವಶೇಷಗಳು ಒಂದು ಕೈಪಿಡಿಯನ್ನು ಹೊರತಂದರು. ರಾಬರ್ಟ್ ಸೆವೆಲ್‌ರ "ಒಂದು ಮರೆತುಹೋದ ಸಾಮ್ರಾಜ್ಯ". 1900 ರಲ್ಲಿ ಎಲ್ಲಾ ತಿಳಿದ ಶಿಲಾಶಾಸನಗಳ ಮತ್ತು ಹಸ್ತಪ್ರತಿಗಳ ಪೋರ್ಚುಗೀಸ್ ಪ್ರಯಾಣಿಕರಾದ ಪೇಸ್ ಮತ್ತು ನ್ಯೂನಿಜ್‌ರ ವರದಿಗಳ ಒಂದು ಚಾರಿತ್ರಿಕ ಸಮನ್ವಯವಾಗಿದೆ.

Fig. 4 Input Image

Since the input images considered are scanned images it contains a type of noise called as clutter noise. This is removed a using a suitable technique called as binarization and thresholding. The noise removed image using the above mentioned technique is shown in Figure 5.



ಭಂಡಬಿದ್ದು, ತೇರ(ಟಯರ)ನೇರಿ ಕೂತ ಈಗ ನಾನೇ ಇಲ್ಲಿ ಜಂಗಮ! ಟಾರು ರೋಡ್ ಬಿಟ್ಟು
ಮಡ್ ರಸ್ತೆಗೆ ಇಳಿದು ಹೊರಟಾಗ ಉದ್ದಕ್ಕೆ ಹೊಗೆ ಬಿಟ್ಟುಕೊಂಡು ಹಾರುವ ರಾಕೆಟ್ಟಿನಂತೆ ನಮ್ಮ ಬಸ್ಸಿನ
ಹಿಂದೆ ದಟ್ಟ ದೂಳು ಎದ್ದು ಬರುತ್ತಿತ್ತು. ಆಸಲು ಸಮಸ್ಯೆ ಶುರುವಾದದ್ದೇ ಈಗ. ಎಲ್ಲಿಯಾದರೂ ಬಸ್ಸು
ನಿಂತಾಕ್ಷಣ, ಹಿಂದೆ ಎದ್ದಿದ್ದ ದೂಳು ನಿಧಾನವಾಗಿ ಬಸ್ಸನ್ನೂ, ಬಸ್ಸಿನ ಮೇಲಿದ್ದ ಜನಗಳನ್ನೂ ಆವರಿಸಿಕೊಂಡು
ಬಿಡುತ್ತಿತ್ತು. ಹತ್ತಾರು ಕಡೆ ಹಾಗೆ ಆಗುವ ಹೊತ್ತಿಗೆ, ನಮ್ಮ ತಲೆ, ಮುಖಗಳೆಲ್ಲ ಗುರ್ತಿಸಿಗದಷ್ಟು
ದೂಳಿನಿಂದಾವೃತವಾಗಿದ್ದವು. ಬೆಳಗ್ಗೆ ಸ್ನಾನ ಮಾಡಿ, ಎಣ್ಣೆ ಹಾಕಿಕೊಂಡಿದ್ದ ತಲೆಯಿಂದ ಇಳಿಯುತ್ತಿದ್ದ
ಬೆವರೆಣ್ಣೆಯ ಜೊತೆಗೆ ಈಗ ಮಣ್ಣೂ ಸೇರಿ, ಮುಖ ಕುತ್ತಿಗೆಗಳನ್ನು ಕೆಬರಿದರೆ ಸಾಕು ಥೇಟು ಗುಡಿಯೊಳಗಿನ
ಕರಿ ಹನುಮನ ಮೈಯಿಂದ ಹೊರಡುವ ಜಿಡ್ಡು ಮ್ಯಾಣ ಬಂದಂತಾಗುತ್ತಿತ್ತು. ನಮ್ಮ ಕಣ್ಣ ರೆಪ್ಪೆಗಳಲ್ಲಿ
ಕೆಂದೂಳು ಅಡರಿ ಅವುಗಳಿಗೊಂದು ರೀತಿಯ ಮೇಕಪ್ ಒದಗಿತ್ತು. ಏನೂ ಮಾಡುವ ಹಾಗಿರಲಿಲ್ಲ.
ಅನುಭವಿಸಬೇಕು. ಬಸ್ಸು ನಿಲ್ಲುವ ಸೂಚನೆ ಕಂಡ ಕೂಡಲೇ ಒಬ್ಬರ ಮುಖ ಒಬ್ಬರಿಗೆ ಕಾಣದಷ್ಟು
ದಟ್ಟವಾಗಿ ಕವಿಯುವ ದೂಳಿಗೆ ಹೆದರಿ ಹೌಹಾರಿಬಿಡುತ್ತಿದ್ದೆವು. ಬೇಂದ್ರೆ ಹಾಡನ್ನು ತುಸು ತಿರುಚಿ ಗೊಣಗಿಕೊಂಡೆ
- 'ಬಸ್ಸಿನ ಬೆನ್ನೇರಿ ಬಂತು ದೂಳಿನಾಭಿಷೇಕ, ಕೆಳಗಿದ್ದ ನೋಡುಗರ ಮೊಗದಲ್ಲಿ ಸಹಜ ಮಂದಹಾಸ!'

Fig. 5 Noise Removed Image

4.1 Skew Detection and Correction

Along with the noise the input image also contains a deviation known as skew which should be detected and eliminated using a suitable algorithm. This is successfully done using RLSA (run length smoothing algorithm). The steps of this algorithm in detail are given below where,

Input: Binary image, Is, of skewed document

Output: Binary image, Ic, of corrected document

Procedure:

1. Run-length value is calculated for Is.
2. Connected components are extracted.
3. Sum of angles and counter is set.
4. Skew angle is calculated.
5. Image is rotated.

6. Corrected image is obtained, Ic.

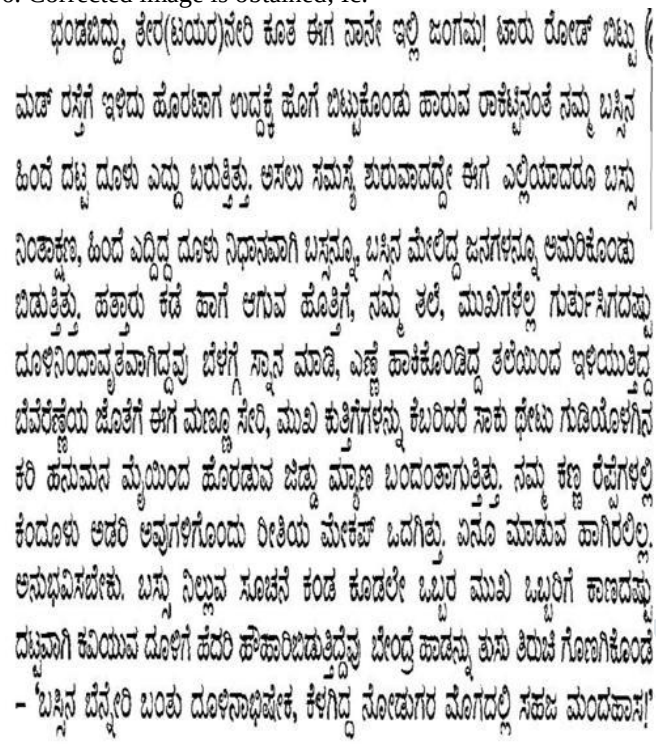

ಭಂಡಬಿದ್ದು, ತೇರ(ಟಯರ)ನೇರಿ ಕೂತ ಈಗ ನಾನೇ ಇಲ್ಲಿ ಜಂಗಮ! ಟಾರು ರೋಡ್ ಬಿಟ್ಟು
ಮಡ್ ರಸ್ತೆಗೆ ಇಳಿದು ಹೊರಟಾಗ ಉದ್ದಕ್ಕೆ ಹೊಗೆ ಬಿಟ್ಟುಕೊಂಡು ಹಾರುವ ರಾಕೆಟ್ಟಿನಂತೆ ನಮ್ಮ ಬಸ್ಸಿನ
ಹಿಂದೆ ದಟ್ಟ ದೂಳು ಎದ್ದು ಬರುತ್ತಿತ್ತು. ಅಸಲು ಸಮಸ್ಯೆ ಶುರುವಾದದ್ದೇ ಈಗ ಎಲ್ಲಿಯಾದರೂ ಬಸ್ಸು
ನಿಂತಾಕ್ಷಣ, ಹಿಂದೆ ಎದ್ದಿದ್ದ ದೂಳು ನಿಧಾನವಾಗಿ ಬಸ್ಸನ್ನೂ, ಬಸ್ಸಿನ ಮೇಲಿದ್ದ ಜನಗಳನ್ನೂ ಅಮರಿಕೊಂಡು
ಬಿಡುತ್ತಿತ್ತು. ಹತ್ತಾರು ಕಡೆ ಹಾಗೆ ಆಗುವ ಹೊತ್ತಿಗೆ, ನಮ್ಮ ತಲೆ, ಮುಖಗಳೆಲ್ಲ ಗುರ್ತುಸಿಗದಷ್ಟು
ದೂಳಿನಿಂದಾವೃತವಾಗಿದ್ದವು ಬೆಳಗ್ಗೆ ಸ್ನಾನ ಮಾಡಿ, ಎಣ್ಣೆ ಹಾಕಿಕೊಂಡಿದ್ದ ತಲೆಯಿಂದ ಇಳಿಯುತ್ತಿದ್ದ
ಬೆವರೆಣ್ಣೆಯ ಜೊತೆಗೆ ಈಗ ಮಣ್ಣೂ ಸೇರಿ, ಮುಖ ಕುತ್ತಿಗೆಗಳನ್ನು ಕೆಬರಿದರೆ ಸಾಕು ಥೇಟು ಗುಡಿಯೊಳಗಿನ
ಕರಿ ಹನುಮನ ಮೈಯಿಂದ ಹೊರಡುವ ಜಿಡ್ಡು ಮ್ಯಾಣ ಬಂದಂತಾಗುತ್ತಿತ್ತು. ನಮ್ಮ ಕಣ್ಣ ರೆಪ್ಪೆಗಳಲ್ಲಿ
ಕೆಂದೂಳು ಅಡರಿ ಅವುಗಳಿಗೊಂದು ರೀತಿಯ ಮೇಕಪ್ ಒದಗಿತ್ತು. ಏನೂ ಮಾಡುವ ಹಾಗಿರಲಿಲ್ಲ.
ಅನುಭವಿಸಬೇಕು. ಬಸ್ಸು ನಿಲ್ಲುವ ಸೂಚನೆ ಕಂಡ ಕೂಡಲೇ ಒಬ್ಬರ ಮುಖ ಒಬ್ಬರಿಗೆ ಕಾಣದಷ್ಟು
ದಟ್ಟವಾಗಿ ಕವಿಯುವ ದೂಳಿಗೆ ಹೆದರಿ ಹೌಹಾರಿಬಿಡುತ್ತಿದ್ದೆವು ಬೇಂದ್ರೆ ಹಾಡನ್ನು ತುಸು ತಿರುಚಿ ಗೊಣಗಿಕೊಂಡೆ
- 'ಬಸ್ಸಿನ ಬೆನ್ನೇರಿ ಬಂತು ದೂಳಿನಾಭಿಷೇಕ, ಕೆಳಗಿದ್ದ ನೋಡುಗರ ಮೊಗದಲ್ಲಿ ಸಹಜ ಮಂದಹಾಸ!'

Fig. 6 Skew Corrected Image

4.2 Segmentation Using RLSA Algorithm

Run Length Smoothing Algorithm (RLSA) has been used previously in text/non-text segmentation, a combination of RLSA in horizontal and vertical direction to segment blocks of text and non-text. Afterwards, the text blocks are analyzed for the extraction of words. In our case, instead of segmenting the document image into blocks of text and non-text, we segment the image directly into words and graphics. For that we apply RLSA only in horizontal direction. The basic RLSA is applied to a binary sequence in which white pixels are represented by 0's and black pixels by 1's.

4.2.1 Text-Line Segmentation using RLSA

Text line segmentation is performed by applying the RSLA algorithm to the original image, after the noise and Skew Elimination. Furthermore, RLSA is constrained by text line and column obstacles. Therefore, RLSA is performed only in cases where a background (white) pixel sequence does not include pixels that they have been also detected as obstacles. Constant b (eqn. 1) is set to d (eqn. 2), which is a relative large value, so as distant parts of the same text line can be linked if the other three conditions of the RLSA are satisfied. Obstacles prevent parts of different text lines to be linked despite of the large value of b.

$T1 = b.\min\{h1, h2\}$ (1)

$T2 = d.\min\{h1, h2\}$ (2)

Steps for line segmentation algorithm:

It is nothing but Separate line from the text document.
We compute the run length of the document image box.
1. Use the RLSA method for segmenting lines from text.
2. Count the white pixel in each row.
3. Find minimum and maximum values of the rows
4. Find minimum and maximum values of the columns
5. The values of rows and columns give no white pixels
6. Replace all such rows and columns by 1
7. Invert the image to make empty rows as 0 and text lines will have original pixels.
8. Crop the line from the min and max values of rows and columns.

The output obtained for the RLSA text line segmentation is shown in the Figure 7.

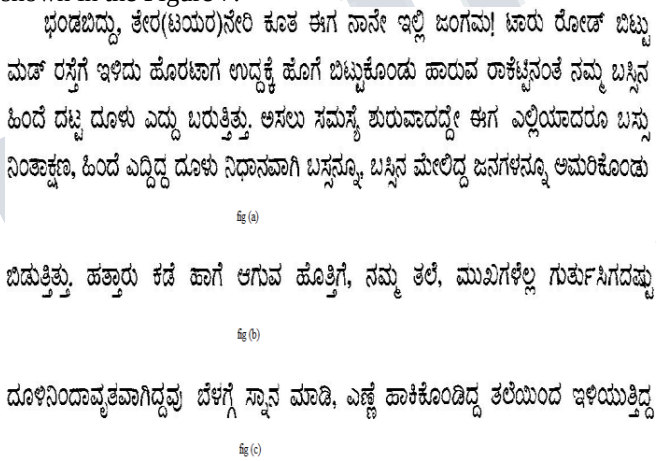

ಭಂಡಬಿದ್ದು, ತೇರ(ಟಯರ)ನೇರಿ ಕೂತ ಈಗ ನಾನೇ ಇಲ್ಲಿ ಜಂಗಮ! ಟಾರು ರೋಡ್ ಬಿಟ್ಟು
ಮಡ್ ರಸ್ತೆಗೆ ಇಳಿದು ಹೊರಟಾಗ ಉದ್ದಕ್ಕೆ ಹೊಗೆ ಬಿಟ್ಟುಕೊಂಡು ಹಾರುವ ರಾಕೆಟ್ಟಿನಂತೆ ನಮ್ಮ ಬಸ್ಸಿನ
ಹಿಂದೆ ದಟ್ಟ ದೂಳು ಎದ್ದು ಬರುತ್ತಿತ್ತು. ಅಸಲು ಸಮಸ್ಯೆ ಶುರುವಾದದ್ದೇ ಈಗ ಎಲ್ಲಿಯಾದರೂ ಬಸ್ಸು
ನಿಂತಾಕ್ಷಣ, ಹಿಂದೆ ಎದ್ದಿದ್ದ ದೂಳು ನಿಧಾನವಾಗಿ ಬಸ್ಸನ್ನೂ, ಬಸ್ಸಿನ ಮೇಲಿದ್ದ ಜನಗಳನ್ನೂ ಅಮರಿಕೊಂಡು
fig (a)
ಬಿಡುತ್ತಿತ್ತು. ಹತ್ತಾರು ಕಡೆ ಹಾಗೆ ಆಗುವ ಹೊತ್ತಿಗೆ, ನಮ್ಮ ತಲೆ, ಮುಖಗಳೆಲ್ಲ ಗುರ್ತುಸಿಗದಷ್ಟು
fig (b)
ದೂಳಿನಿಂದಾವೃತವಾಗಿದ್ದವು ಬೆಳಗ್ಗೆ ಸ್ನಾನ ಮಾಡಿ, ಎಣ್ಣೆ ಹಾಕಿಕೊಂಡಿದ್ದ ತಲೆಯಿಂದ ಇಳಿಯುತ್ತಿದ್ದ
fig (c)

Fig. 7 Output for Line Segmentation using RLSA algorithm

Text line segmentation output for the input image i.e., Figure 5 using the RLSA algorithm is shown in the below Figure 8.

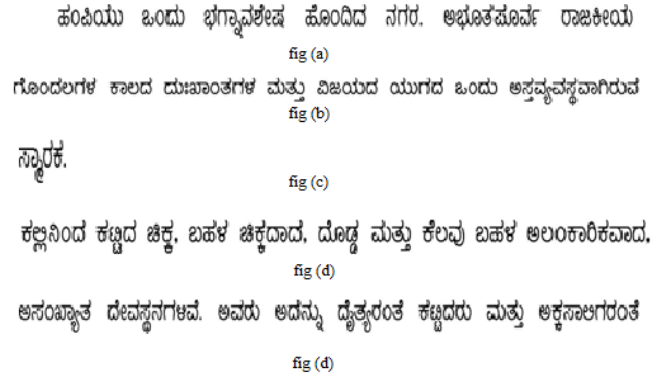

ಹಂಪಿಯು ಒಂದು ಭಗ್ನಾವಶೇಷ ಹೊಂದಿದ ನಗರ. ಅಭೂತಪೂರ್ವ ರಾಜಕೀಯ
fig (a)
ಗೊಂದಲಗಳ ಕಾಲದ ದುಃಖಾಂತಗಳ ಮತ್ತು ವಿಜಯದ ಯುಗದ ಒಂದು ಅಸ್ತವ್ಯವಸ್ಥವಾಗಿರುವ
fig (b)
ಸ್ಮಾರಕ.
fig (c)
ಕಲ್ಲಿನಿಂದ ಕಟ್ಟಿದ ಚಿಕ್ಕ, ಬಹಳ ಚಿಕ್ಕದಾದ, ದೊಡ್ಡ ಮತ್ತು ಕೆಲವು ಬಹಳ ಅಲಂಕಾರಿಕವಾದ.
fig (d)
ಅಸಂಖ್ಯಾತ ದೇವಸ್ಥನಗಳವೆ. ಅವರು ಅದನ್ನು ದೈತ್ಯರಂತೆ ಕಟ್ಟಿದರು ಮತ್ತು ಅಕ್ಕಸಾಲಿಗರಂತೆ
fig (d)

Fig. 8 Output for Line Segmentation using RLSA algorithm

4.2.2 Word Segmentation

The word segmentation methodology of the proposed procedure is connected autonomously to every content line identified from the past phase of the algorithm. Every associated segment of a content line Li are initially sorted by x coordinate and the histogram Hd of the level

separations between nearby jumping boxes is built. A negative quality for a separation of vertically covered bouncing boxes is thought to be zero. It is only separate word from the line.

Steps for word segmentation Algorithm:

1. Label and count connected components
2. Use the RLSA method for segmenting word from each line.
3. Count the white pixel in each row.
4. Find minimum and maximum values of the rows
5. Find minimum and maximum values of the columns
6. The values of rows and columns give no white pixels
7. Replace all such rows and columns by 1
8. Invert the image to make empty rows as 0 and text lines will have original pixels.
9. Crop the word.
10. Save the word in the file.

The output obtained for the RLSA word segmentation is shown in the below Figure 9.

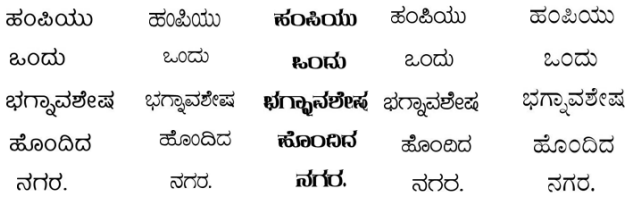


Fig. 9 Output for Word Segmentation using RLSA algorithm

4.3 Segmentation Using VB Method

This is another segmentation approach in view of blend thickness estimation utilizing the Variational Bayes (VB) system. Seeing the document image as a dispersion of pixels, every content line can be demonstrated as bivariate Gaussian circulation and the document is a blend of Gaussians. The VB strategy can consequently decide the quantity of parts without extra overhead of model request determination. For processing document images, we have augmented the VB strategy such that it can part segments and in addition take out excess segments and specifically control the introduction of content lines. The output obtained for the VB method text line and word segmentation is as shown in the below Figure 10 and Figure 11 respectively.

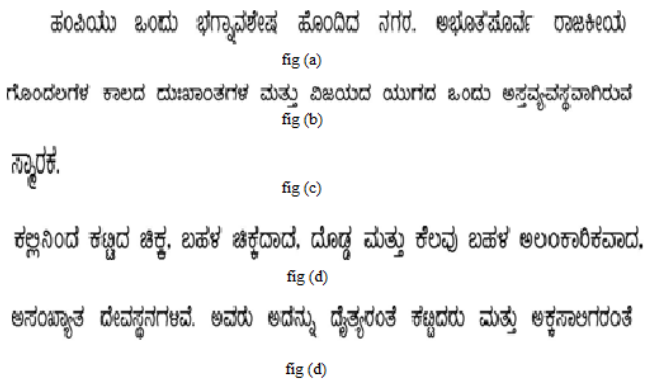


Fig. 10 Output for Line Segmentation using VB method

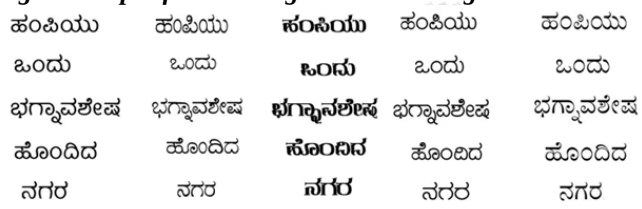


Fig. 11 Output for Word Segmentation using VB method

V. COMPARATIVE ANALYSIS OF RESULTS

Based on the results obtained for both the RLSA and VB methods a comparative analysis is done and a particular algorithm is considered as the best algorithm. A slight change in the results is observed in both text line and word segmentation techniques of the proposed methods. Since it is the segmentation of the documents which are the most important either for the personal or official use, a slight change in the results can even change the actual meaning itself. Comparative analysis of the text line segmentation results is as shown in the below Figure 12 which indicates that there is a dilation of the last sentence font size in the RLSA method and the same sentence result is displayed perfectly in VB method.

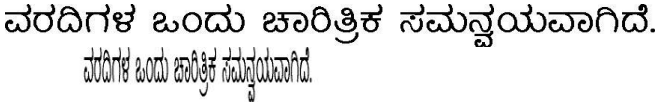


Fig. 12 Comparative Analysis of the Line Segmentation Results

As mentioned earlier a slight change in the sentence can change the actual meaning itself, the same is depicted in Figure 13 where because of the use of the clip command in the RLSA method, the subscripts are clipped for a line which has led to the change in the meaning of the actual sentence itself. The same sentence is displayed perfectly in the VB method. The same analysis is done for the word segmentation results, the same segmented words are displayed which is clearly shown in the Figure 14. The only difference observed here is the display of special

characters in the RLSA method along with word which is not in the VB method. From all the above analysis among the proposed methods Vartional Bayes (VB) method is concluded as the best method for the Segmentation of printed documents

Fig. 13 Problem observed in RLSA Line Segmentation Result

Fig. 14 Comparative Analysis of the Word Segmentation Results

VI. CONCLUSION

The paper proposes techniques of line and word segmentation scheme along with the noise removal and skew detection and correction for documents that are printed in Kannada Language scripts. The comparative analysis of the results of line and word segmentation clearly indicates that among the proposed methods Variational Bayes method is efficient compared to the RLSA method. The proposed system also works well for the different font styles for both line and word segmentation.

REFERENCES

[1] Zaidi Razak, Khansa Zulkiflee, Mohd Yamani IdnaIdris, Emran Mohd Tamil, Mohd Noorzaily, Mohamed Noor, Rosli Salleh, MohdYaakob, Zulkifli Mohd Yusof, and Mashkuri Yaacob, "Off-line Handwriting Text Line Segmentation: A Review", 2008.

[2] S. Nicolas, T. Paquet, L. Heutte, "Text line segmentation in handwritten document using a production system", 2004.

[3] F. Yin and C L Liu, "A Variational Bayes Method for Handwritten Text Line Segmentation", 2009.

[4] T. Sari and M. Sellami, "Overview of Some Algorithms of Off-Line Arabic Handwriting Segmentation", 2007.

[5] B. Gatos, A. Antonacopoulos and N. Stamatopoulos, "ICDAR2009 Handwriting Segmentation Contest", 2009.

[6] C. Zhang and G.S. Lee, "Text Line Segmentation in Chinese Handwritten Text Images", 2011.

[7] R. Kumar and A. Singh, "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text", 2010.

[8] U. Pal and S. Datta, "Segmentation of Bangla unconstrained handwritten text", 2003.

[9] N. Kumar Garg, L Kaur and M. K. Jindal "Segmentation of Handwritten Hindi Text", 2010.

[10] J. D. Gupta and B. Chanda, "A Model Based Text Line Segmentation Method for Off-line handwritten Document", 2010.

[11] Mamatha H and Srikantamurthy K "Skew Detection, Correction and Segmentation of Handwritten Kannada Document", 2012.

[12] J.Venkatesh and C. Sureshkumar "Tamil Handwritten Character Recognition Using Kohonon's Self Organizing Map", 2009.

[13] Mamatha H R and Srikantamurthy K, "Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document", 2012.

[14] Alireza Alaei, P. Nagabhushan and Umapada Pal "A New Dataset of Persian Handwritten Documents and its Segmentation", 2011.