International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 25155-25164

http://www.ripublication.com

Extraction of Touching or Overlapping Lines From Handwritten Document Images Using Normalized Chain Codes And Projection Profiles

N. Shobha Rani [#1], Hemanth K. P. [#2], Hrudya [#3]
[#] Department of Computer Science, Amrita Vishwa Vidyapeetham
Mysore Campus, Mysore-570026, Karnataka, India
[1] n.shoba1985@gmail.com
[2] Hemanthkp123@gmail.com
[3] Hrudya.snair@gmail.com

Abstract

Touching Line extraction from handwritten documents is one of the pertinent problem in the process of optical character recognition, which is generally called as line segmentation of text document images. Identification of the touching or overlapping portions in the document is the major challenge in this regard. In this paper we propose an algorithm for the touching or overlapping line extraction using horizontal projection profiles and a set of predefined chain code sequences that indicates the presence of touching portions between the touching lines. The matched chain code sequence is further normalized to obtain the first difference of the matched chain code sequence. The midpoint of normalized chain code is approximately considered as the segmentation points in the path of line segmentation. The results of proposed algorithm are satisfactory with an accuracy of about 89%.

Keywords: chain code sequence, extraction, horizontal projection profile, optical character recognition, segmentation.

Introduction

Image processing is a method to convert an image into digital form and perform some operations on it, in order to get an enhanced image or to extract some useful information from it. It is a type of signal apportion in which input is image of handwritten document, photograph and output may be image or characteristics associated with that image. Document Image processing aims to develop algorithms and processes through which machines (computers) can automatically read and develop some basic understanding of input documents .Documents may be printed documents – such as memos, letters, technical reports, books or Hand written documents like addresses written on postal mail personal letters, notes in the margins of documents.

Segmentation is one of the most critical as well as important stage in optical character recognition (OCR) system. Especially extraction of touching or overlapping text lines from South Indian scripts has become one of the challenging aspects due to the complexity involved in south Indian scripts. This paper deals with text line extraction of handwritten South Indian scripts with respect to Kannada and Telugu scripts. Segmentation is a very prominent step since the accuracy of the character segmentation and recognition algorithm highly depends on the accurate line segmentation. It is a process of decomposing the text document into lines further to words and then to characters. Inaccurate line segmentation not only decreases the accuracy of the whole OCR but also may lead to some other errors like extract of incorrect features and misclassification results.

Line segmentation is quite easy in printed text as there are equal line spacing in printed text and almost similar writing style, where as in hand written text documents it is not so easy due to the freedom taken by the user at the time of drafting the handwritten documents. The other artifacts that are persistent in the handwritten text documents are line spacing is not equal, also there will be different font size and style which makes the line segmentation a very tough job. In handwritten text, problem of skewness, overlapping and touching of text also makes line segmentation difficult. Because of the overlapping and touching characters the complexity of these documents is high and automatic text line segmentation of handwritten documents technique remains an open research problem in the field of touching line extraction.

The main aim of the proposed experimentation is to provide an efficient touching line segmentation algorithm using horizontal projection profiles and a set of predefined chain code sequences.

Literature Survey

Numerous attempts are reported in the literature in connection with extraction of touching line like documents. Some of the papers reviewed are discussed as below.

Bidyut et.al [1] has proposed an language independent approach of touching line extraction from handwritten documents based on inter line gaps and peaks information obtained through histograms. The curves are drawn between the lines to indicate the path of segmentation between the lines. The methodology has obtained an accuracy of around 92%. L.Pratap Reddy et.al[2], had proposed a method called Touching Syllable Using Split Profile Algorithm. This method is applied on cursive scripts and the performance of the proposed algorithm is compared with two approaches methods viz aspect ratio and syllable width approaches. This method gives good results only with Anupama, Gowthami, and Priyanka font. M. Thungamani et.al[3] have proposed an algorithm namely Horizontal Projection Profile Algorithm using histogram to segment the lines in handwritten text document. The proposed methodology focuses on the white spaces between the text lines to segment the text lines by drawing the histogram of the number of pixels in each row and hence finding the row with no active pixel and by doing false minimum calculation.

Sunanda Dixit, Assistant et.al [4] had proposed a method for segmentation of Handwritten Kannada Text document. The line segmentation problem is computed here through computation of Standard Error and Weighted Bucket Algorithm . The proposed algorithm mainly uses identification and formulation of the composite character syllable and integrating it by computing the standard error of mean to represent a line using weighted bucketing algorithm. A method for segmenting lines from handwritten text document has proposed by Zaidi Razak, et.al [5].In this paper they have used smeared approach where the consecutive black pixels along the horizontal direction are smeared. In this technique, If the distance between the white space is within a predefined threshold, it will be filled with black pixels. Pramod S. Malgi et.al [6] have proposed an algorithm to segment the line from handwritten document of Devanagari script based on the header line detection, base line detection and contour following technique. In this paper they have used two-stripe projection method for segmentation of lines. Alireza Alaei et.al [7] has proposed a technique for text line segmentation of handwritten documents where they have initially the input document is vertically decomposed into parallel pipe structures called stripes. Each row of a stripe is painted by a gray intensity, which is the average intensity value of gray values of all pixels present in that row-stripe. The painted stripes are treated with some smoothing operations. A dilation operation is employed on the foreground portion of the smoothed image to obtain a single component for each text-line. Thinning of the background portion of the dilated image and subsequently some trimming operations are performed to obtain a number of separating lines, called candidate line separators. Overall, 95.32% of text-lines in the handwritten document were accurately segmented using this technique. N Tripathy et.al [8] has proposed The global horizontal projection method where it computes the sum of all black pixels on every row and constructs the corresponding histogram. Based on the peak/valley points of the histogram, individual lines are generally segmented. But this technique does not hold good for unconstrained or overlapping text documents. Satadal Saha et.al [9] has proposed a Hough transform based Technique for text segmentation; the proposed technique is applied not only on the document image dataset but also on dataset for business card reader system and license plate recognition system. And it can also be efficiently utilized in case of various domains of image segmentation which can be subsequently used in various domain specific OCR systems. Mamatha H R et.al [10] has proposed Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document. The proposed technique is a segmentation scheme for segmenting handwritten Kannada scripts into lines, words and characters using morphological operations and projection profiles. Usage of the morphology made extracting text lines efficient by an average extraction rate of 94.5% .Also 82.35% and 73.08% for words and characters respectively.

Most of the existing experimentations has provided appreciable methods and has obtained very good accuracy. The experimentations in the literature are also quite intensive and complicated. The present system provides a simple method using chain code sequences and popular projection profiles. The rest of the paper is organized as follows, the section 3 presents the proposed methodology, section 4 discusses the experimental results and section 5 concludes the work.

Proposed Methodology

The proposed methodology is composed of several stages. In the first the preprocessing performed. The second stage computes the horizontal projection profiles of a given document image., the normalized distinct array is computed in the stage 3. The stage 4 identifies the path of segmentation based on normalized distinct array. Architecture of proposed methodology is depicted in the figure 1 and sample document image is as presented in the figure 2.

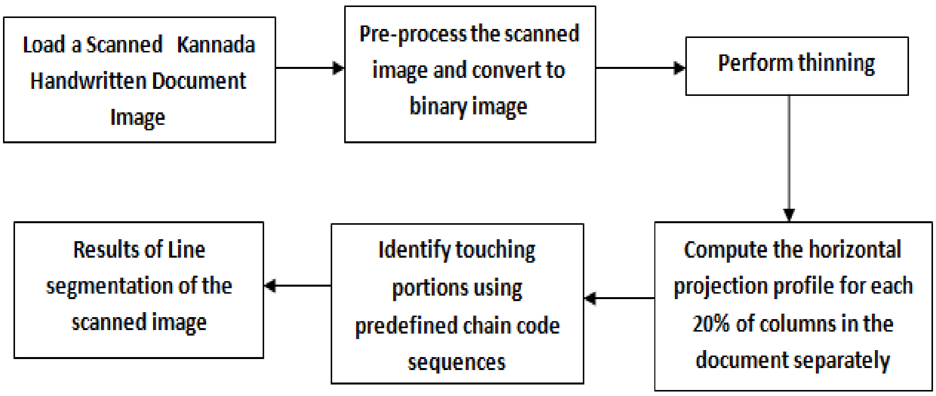

Figure 1: System architecture of proposed system

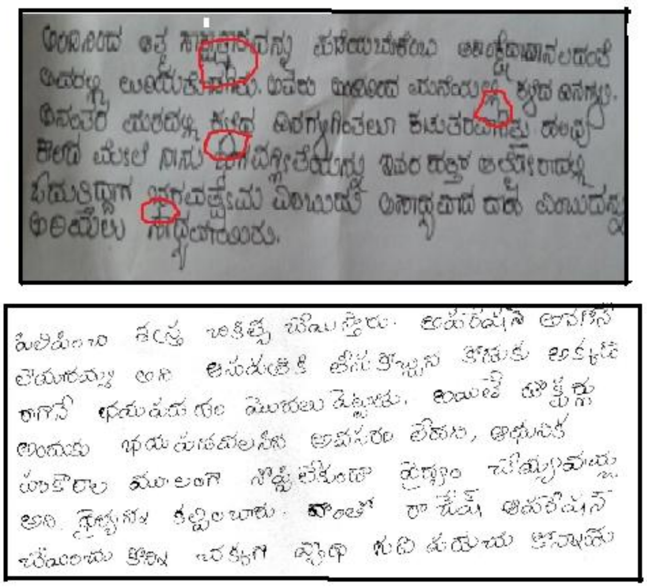

Figure 2: Sample document Images

First we input a document image D which will be pre-processed. The main aim of pre-processing is to enhance the visual appearance of the image; and thus the result of the pre-processing is an enhanced image EI. On the enhanced image EI morphological operations like thinning is performed. Thinning is a process which is used to remove noise from the image. In the present scenario, thinning gives a clear separation between the characters and lines in the image. The algorithm for thinning is as depicted in algorithm 1.

Algorithm 1: Thinning

Text points are assumed to have value 0 & background points are assumed to have value 1.

Consider a pixel representation as

P9	P2	P3
P8	P1	P4
P7	P6	P5

Step 1: Flags a contour point for deletion if the following conditions are satisfied:

(a) $2 \leq N(p1) \leq 6$

(b) $T(p1) = 1$

(c) $p2.p4.p6 = 0$

(d) $p4.p6.p8 = 0$

where N(p1) is the number of nonzero neighbors of p1;

i.e. $N(p1) = p2 + p3 + \ldots\ldots + p8 + p9$

where pi is either 0 or 1.

T(p1) is number of 0-1 transitions in the ordered sequence p2, p3, ….., p8, p9., p2

Step 2: Conditions a & b remain the same, but conditions c & d are changed to

c’) $p2.p4.p8 = 0$

d’) $p2.p6.p8 = 0$

0	0	1
1	p1	0
1	0	1

Step 1 is applied to every border pixel in binary region under consideration

If 1 or more of conditions from a to d are violated, the value of point in question is not changed. The results of thinning algorithm are as shown in figure 3.

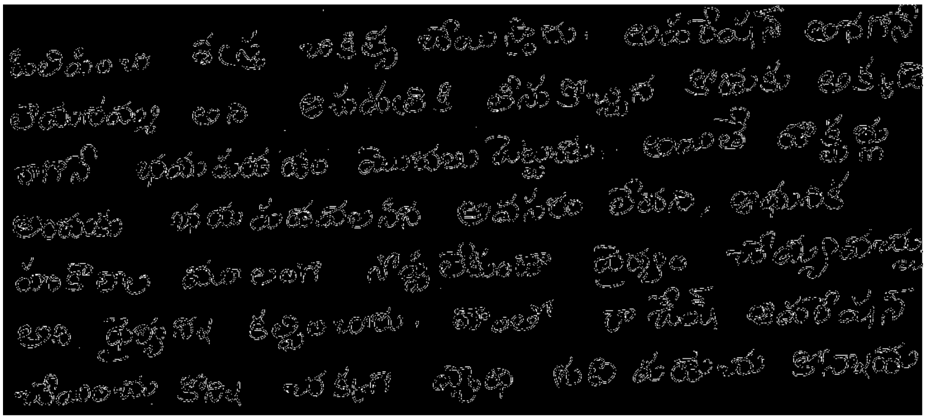

Figure 3: Thinned image

The thinned image BI is partitioned into n number of vertical segments VS_i, which is of size n; where n is the number of columns in each vertically segmented partition. On these vertically segmented partition, Horizontal Projection Profile $H_k(VS_i)$ is applied.

In the Horizontal projection profile, for each of the segmented block, a region of minimum valley is identified and for each of it, a row index is marked and is separately stored in an array (A_s). This is how we obtain a combined array A; where A has all the array's which is separately stored on the basis of marked row indices. A is of the size M x N where M is max number of rows in array A_i and N is number of columns equal to total Number of arrays A_i. For each distinct value in combined array A calculate the count of repetition in the A and obtain a distinct marked row array DA of size M x 2 where column 1conatin each distinct row in all A_i and column 2 contains count of each distinct row in column 1. Then from the DA we obtain a Normalized distinct array NDA.

Based on the background intensity of the document, for each value of V in NDA, a path of traversal for identifying the segmentation curve between the text lines in the document is calculated. And if a black pixel is encountered during the traversal of the path, then 8 directional chain code is calculated from that black pixel until a specified threshold. If chain code sequence belongs to 08888 or 07777 or 03333 or 06666 or 04444 or 02222, then we consider that black pixel as the segmentation path else we move to the next row to find the segmentation path. This is continued until all the rows in the NDA is exhausted. The algorithm for line segmentation using predefined chain code sequences is as presented in the algorithm 2

Algorithm:2

Step 1: Input a document image D

Step 2: Pre-process the document image D to obtain enhanced image EI

Step 3: Thin the image EI

Step 4: Convert the image into binary image BI.

Step 5: partition the image BI into n number of vertical segments VS_i of BI of size n

Where n is the number of columns in each segmented partition

Step 6: Compute the horizontal projection profile $H_k(VS_i)$ of the each vertical segmented block

VS_i where i=1, 2, 3…n and k=1, 2, 3…

Step 7: For each $H_k(VS_i)$ mark the row indices in the regions of minimum valleys and store it in array A_s.

Step 8: Obtain the combined array A, where $A = A_1 \, U \, A_2 \, U \, A_3 \, U.....A_s$ and size of (A)= M x N

Where M is max number of rows in array A_i and N is number of columns equal to total Number of arrays A_i.

Step 9: For each distinct value in combined array A calculate the count of repetition in the A and

Obtain a distinct marked row array DA of size M x 2 where column 1conatin each distinct row in all A_i and column 2 contains count of each distinct row in column 1.

Step 10: Obtained normalized distinct array NDA from DA

Step 11: For each value v in NDA, compute the path of traversal for identifying the segmentation curve between the text lines in the document based on the back ground intensity of document.

Step 12: If any black pixel is encountered in the path, then calculate the 8 directional chain code from that black pixel til a specified threshold.

Step 13: if chain code sequence belongs to 08888 or 07777 or 03333 or 06666 or 04444 or 02222 then consider the encountered black pixel in the path as segmentation path else move to the next row in the NDA.

Step 14: Continue this process til all the rows in the NDA exhausts.

Step 15: Stop.

The chain codes using in proposed methodology are 8-directional chain code sequences. The directional structure for 8-directional chain codes are presented in figure 4.

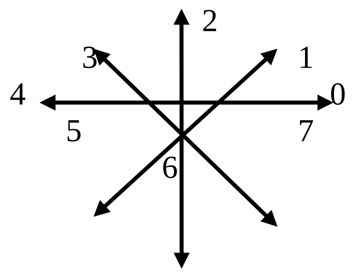

Figure 4: 8- Directional chain code structure

The touching portions between any two lines in the images may follow any of the chain code sequences which are defined as follows. At any particular touching portion between the lines, the subscript may traverse in upward direction or downward direction or any diagonal direction etc. Some of the chain codes defined with respect to various touching portions in the documents are defined in figure 5, figure 6 and figure 7.

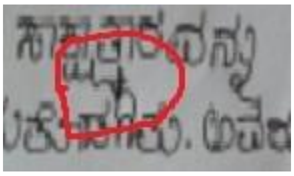

Figure 5: Chain codes to identify these types of touching segments are 02222, 07777, 03333 and 06666. Any of the chain code matching with touching portion may identify the touching portion.

Figure 6: Chain codes to identify these are 01111 or 07777

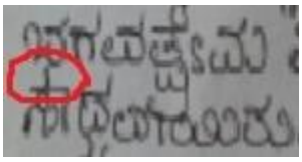

Figure 7: Chain codes to identify these are 02222 or 06666

The normalized chain codes of the predefined chain codes calculated by computing the distance between two adjacent values by moving in anti-clock wise direction from first component to last component. i.e., for 02222 the first difference is 20006. The midpoint is 0 and its corresponding coordinate position is considered as segmentation path. The experimental results of line segmentation are analyzed in the preceding section.

Experimental Analysis

The experimentation of the proposed methodology is performed with around 50 images of varying scripts. The results obtained by the line segmentation process are as presented in figure 8 and extracted lines are depicted in figure 9.

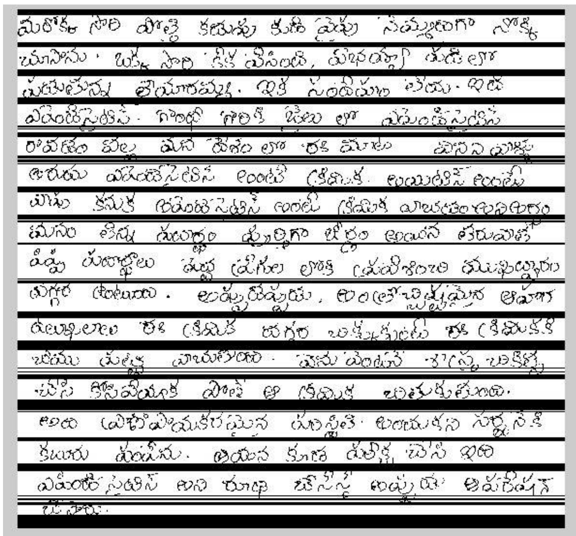

Figure 8: Line segmentation

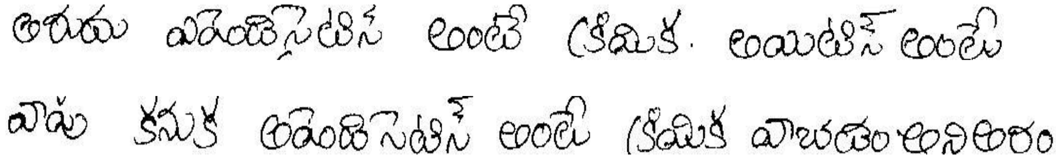

Figure 9: Results of line extraction

The proposed algorithm is able to segment the lines only if there are touching portions between the lines as presented by the chain codes defined, otherwise the algorithm may fail to work on images since the chain codes can be used to represent only limited number of shapes and structures. The investigation on the further improvement of the algorithm using fourier descriptors is under progress.

Conclusion

The horizontal projection profiles and chain codes are simple constructs which are constrained to work only with finite and even very limited number of features in images. The touching portions in the documents are also limited for certain types of cases especially in South Indian scripts, thus in order to simplify the process of line segmentation the chain codes and projection profiles are used as tools to perform the line segmentation. The algorithm is experimented with Telugu and Kannada scripts which are synthetically generated through more than 10 writers for the experimentation.

References

[1] Bidyut B. Chaudhuri, Sumedha Bera, “Handwritten Text Line Identification In Indian Scripts”, 2009 10th International Conference on Document Analysis and Recognition.

[2] L.Pratap Reddy, T.Ranga Babu, N.Venkata Rao and B.raveendra Babu, “Touching Syllable Using Split Profile Algorithm”, IJSC International Journal of Computer Science Issues, Volume 7, Issue 3, No 9, May 2010.

[3] M. Thungamani and P.Ramakhanth Kumar, “A Survey of Methods and Strategies in Handwritten Kannada Character Segmentation”, International Journal of Science Research Volume 01, Issue 01, June 2012, pp. 18-23 .

[4] Sunanda Dixit and C. Dr .H. N. Suresh, “Segmentation of handwritten kannada document through computation of Standard Error and Weighted Bucket algorithm”, International Journal of Advanced Computer Technology, Volume 3, Number 2, .

[5] Zaidi Razak, and Mashkuri Yaacob, “Off-line Handwriting Text Line Segmentation”, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008 .

[6] Pramod S. Malgi & Shailaja Gayakwad, “Line Segmentation of Devanagari Handwritten Documents”, 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, 25-32 .

[7] Alireza Alaei, P. Nagabhushan and Umapada Pal,”A Benchmark Kannada Handwritten Document Dataset and its Segmentation”,2011 International Conference on Document Analysis and Recognition.

[8] N TRIPATHY and U PAL , “ Handwriting segmentation of unconstrained Oriya text”, S¯adhan¯ a Vol. 31, Part 6, December 2006, pp. 755–769.

[9] Satadal Saha, Subhadip Basu, Mita Nasipuri and Dipak Kr. Basu , “A Hough Transform based Technique for Text Segmentation ”,Journal of Computing, volume2, issue 2, february 2010, ISSN 2151-9617 .

[10] Mamatha H R and 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 Volume 4– No.5,October 2012 .