International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 09 , September 2017
www.ijarse.com

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Run Length Smoothing Algorithm for Segmentation

Mrs. Archana P. Borlepwar[1], Mrs. Sonal R. Borakhade[2]

Ms.Sneha B. Pradhan[3]

CSMSS college of Polytechnic, Aurangabad

ABSTRACT

Segmentation is the operation that seeks to decompose a word image in a sequence of sub images. There are two types of segmentation bottom up segmentation and top down segmentation. The first approach, is a called bottom-up method, which starts by first segments the document into small blocks (marks), and then merges them into bigger blocks. The second approach is called top-down method in that segment the document into large blocks and then analyses them in order to achieve separation of the characters of the text blocks. This paper deal with one of the top down method called Run Length smoothing algorithm used for segmenting the postal document into destination address block. The RLSA used for detecting the non text part of postal document. Using Piecewise projection method DAB is first segmented in lines and then line into words.

Keywords**: — **DAB, Run Length Smoothing Algorithm , Skew detection

I. INTRODUCTION

The Run Length Smoothing Algorithm (RLSA) is a method that can be used for Block segmentation and text discrimination. The method developed for the Document Analysis System consists of two steps. First, a segmentation procedure subdivides the area of a document into regions (blocks), each of which should contain only one type of data (text, graphic, halftone image, etc.). Next, some basic features of these blocks are calculated. The basic RLSA is applied to a binary sequence in which white pixels are represented by 0's and black pixels by 1's. The algorithm transforms a binary sequence **x** into an output sequence **y** according to the following rules:

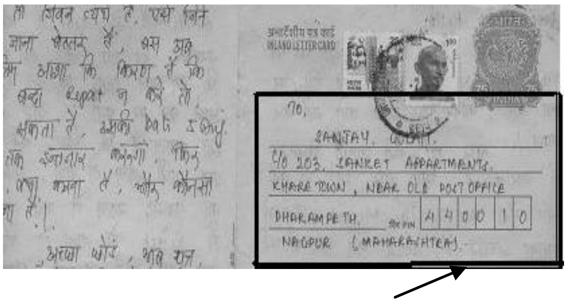
1. 0's in x are changed to 1's in y if the number of adjacent 0's is less than or equal to a predefined limit C.
2. 1's in x are unchanged in y .

For example, with C = 4 the sequence **x** is mapped into **y** as follows:

x : 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0

y : 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1

When applied to pattern arrays, the RLSA has the effect of linking together neighboring black areas that are separated by less than C pixels. With an appropriate choice of C, the linked areas will be regions of a common data type. The RLSA is applied row-by-row as well as column-by-column to a document, yielding two distinct bit-maps. Because spacing of document components tend to differ horizontally and vertically, different values of C are used for row and column processing. The two bit-maps are then combined in a logical AND operation. Additional horizontal smoothing using the RLSA produces the final segmentation result. For implementing RLSA algorithm take the Image of postal document. Postal automation is a topic of research interest for last two decades and many pieces of published article are available towards postal automation of non-Indian languages documents .Several systems are also available for postal automation in USA, UK, France, Canada and Australia. But no work has been done towards the automation of Indian postal system. System development towards postal automation for a country like India is more difficult than that of other countries because of its multi-lingual and multi-script behavior. Some people write the destination address part of a postal document in two or more language scripts. For example, see Fig.1, where the Destination address is written partly in Hindi script and partly in English. In India there is a wide variation in the types of postal documents. Post-card, inland letter, special envelopes are sold from Indian post offices and there is a pin-code box to write pin number and also some commercial envelopes with or without pin-code box. In some documents find partial pin code instead of full pin-code and even no pin-code. In the proposed scheme, at first, the document skew is detected and corrected. Next, using Run Length Smoothing Algorithm(RLSA)[16]



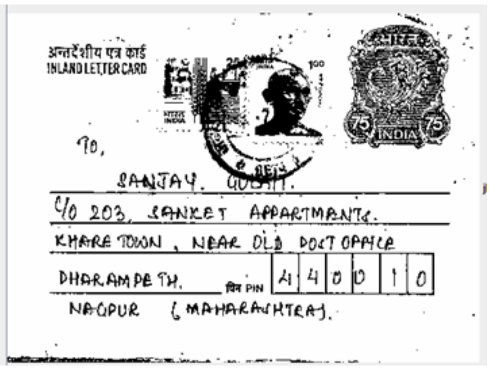
DAB

Fig.1. Example of bi-script postal document DAB

and characteristics of different components of postal document, non-text parts (postal stamp/seal etc.) are detected and removed from the documents, Based on the positional information Destination Address Block (DAB*)* is then located. Using a piece-wise horizontal projection method the DAB is segmented into lines and by vertical histogram the lines into words. For segmentation here use top down approach in that segment the document into large blocks and then analyses them in order to achieve separation of the characters of the text blocks. The top-down approaches belongs the Projection profile method (PPM)[3]and methods based on run-length segmentation algorithm (RLSA)[1**].** PPM is addressed with many disadvantages such as the sensitivity of the document skew. On the other hand, the RLSA is the most powerful procedure for top-down block segmentation. This technique is first introduced by Wong et al. and Wahl et al. and taken up again by Wang and Shihari **.** It is a low complexity technique and can segment documents into rectangular blocks and then classified them into text, graphics or more detailed objects.

II. IMAGE BINARIZATION AND NOISE REDUCTION

Binarization means digitization of image. Binarization of an Image is shown in fig 2. Binarize an image based on the threshold value. Thresholding is an image processing technique for converting a grayscale or color image to a binary image based upon a threshold value. If a pixel in the image has an intensity value less than the threshold value, the corresponding pixel in the resultant image is set to black. Otherwise, if the pixel intensity value is greater than or equal to the threshold intensity, the resulting pixel is set to white. Thus used an image with only 2 colors, black (0) and white (255) .



अन्तर्देशीय पत्र कार्ड
INLAND LETTER CARD

To,
SANJAY. GUPTA.
C/o 203, SANKET APPARTMENTS.
KHARE TOWN , NEAR OLD POST OFFICE
DHARAMPETH. पिन PIN 4 4 0 0 1 0
NAGPUR (MAHARASHTRA).

Fig.2. Binarization of Fig.1.

2.1. Skew Detection

When a document is fed to the optical sensor either mechanically or by a human operator, a few degrees of skew (tilt) is unavoidable. Skew angle is the angle that the text lines in the digital image makes with the horizontal direction. Skew estimation and correction are important preprocessing steps of document layout analysis and OCR approaches. One of the popular skew estimation techniques is based on projection profile of the documents [5]. The horizontal/vertical projection profile is a histogram of the number of black pixels along horizontal/vertical scan lines. For a script with horizontal text lines, the horizontal projection profile will have peaks at text line positions and thorough at positions in between successive text lines. To determine the skew of a document, the projection profile is computed at a number of angles and for each angle, a measure of difference of peak and through height is made. The maximum difference corresponds to the best alignment with the text line direction which, in turn, determines the skew angle.

III. SEGMENTATION

3.1.Postal Stamp Detection and DAB Detection

The binary image is processed to extract the postal stamp and other graphics parts present in the image. There are many techniques for text/graphics separation. For removing the non text part(Postal seal and postal stamp) and other graphical parts used the RLSA(Run Length smoothing algorithm). And From the positional Information DAB is segmented from the document. RLSA algorithm is as follows

RLSA Algorithm

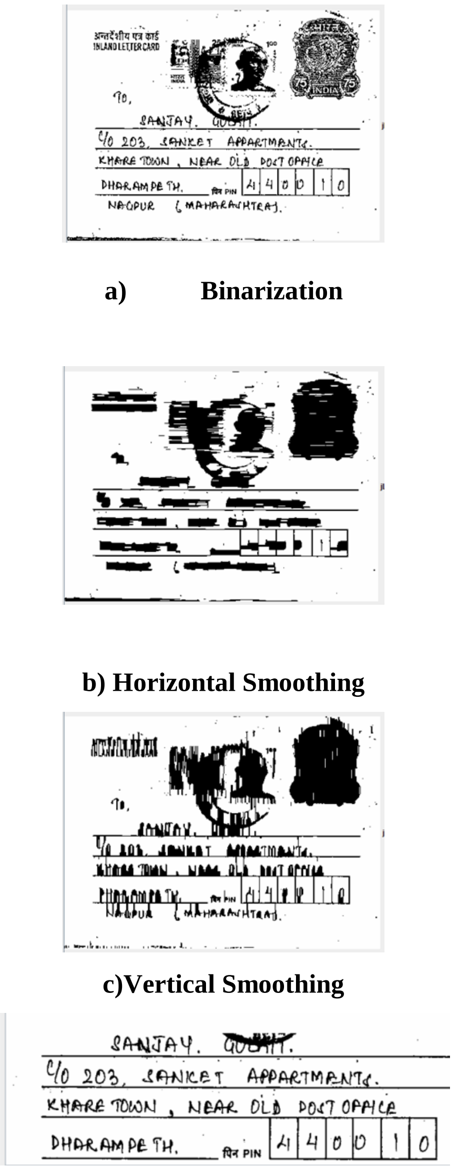
RLSA has three main steps

Step1: In the original image, RLSA is applied horizontally, row-by-row by using hsv and then vertically, column-by-column with vsv.

Step2: After horizontal and vertical smoothing, we have two bit-maps, which are next combined by a logical AND operation to produce a new smoothing image.

Step3: This image has small gaps that interrupt blocks of text lines, therefore, an additional horizontal smoothing operation is performed by using a new suitable smoothing value ahsv.

Graphically it is shown in the fig 3.



a) Binarization

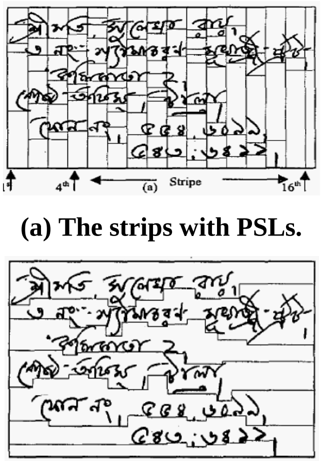
b) Horizontal Smoothing

c)Vertical Smoothing

Fig . 3. Steps for RLSA

3.2. Line Segmentation

For a unconstrained document use piece-wise projection method[1]. In this method divide the text into vertical stripes of width W. Width of the last tripe may differ from W. If the text width is Z and the number of stripe is N then the width of the last stripe is [Z-W*(N-1)]. Next compute Piece-wise Separating Lines (PSL) from each of these stripes. Compute row-wise sum of all black pixels of a stripe. The row where this sum is zero is a PSL. Sometimes get a few consecutive rows where sum of all black pixels is zero. Then the first row of such consecutive rows is the PSL. The PSLs of different stripes of a text are shown by black horizontal lines in Fig4(a). All these PSLs may not be useful for line segmentation. Choose some potential PSLs among these. Compute the normal distances between two consecutive PSLs in a stripe. So if there are n PSLs in a stripe we get n-1 distances. This is done for all stripes. Compute the statistical mode (MPSL) of such distances. If the distance between any two consecutive PSLs of a stripe is less than MPSL remove the upper PSL of these two PSLs. PSLs obtained after this removal are the potential PSLs. The potential PSLs of Fig.4(a) are shown in Fig4(b). Note the left and right co-ordinates of each PSL for future use. By proper joining of these potential PSLs get individual text lines.



(a) The strips with PSLs.

(b) The segmented lines.

Fig.4. Line segmentation

3.3. Word Segmentation

After a text line is segmented, it is scanned vertically. For Word Segmentation Connected Component Labeling algorithm is used . Why labeling , as from the output known that this is PSL, but computer does not understand that that's why given label . After applying labeling getting border of words from this words are separated from text document.

IV. CONCLUSION

There are various methods of segmentation technique but with RLSA getting the better output. Using the RLSA algorithm getting 100% proper segmentation. By using the projection profile method word are separated from the lines and characters are segmented from words.

REFERENCES

[1] "A System for Word-wise Handwritten Script Identification for Indian Postal Automation " , K.-Roy, A. Baner and U. Pal, IEEB INDIA ANNUAL CONFERENCE 2004, INDlCON 2014.

[2] U. Mahadevan, and S. N. Srihari, "Parsing and Recognition of City, State, and ZIP Codes in Handwritten Addresses", In Proc. of 5th Int. Conf. on Document Analysis and Recognition, 1999, pp.

[3] K. Roy, S. Vajda, U. Pal, and B. B. Chaudhuri, "A'System towards Indian Postal Automation", In Proc. of lntemational Workshop on Frontier of Handwriting Reognition-9,2004.

[4] U. Pal and B. B. Chaudhuri, “Script line separation from Indian multi-script documents” IETE Journal of Research, Vol. 49, 325-328. 2003, pp. 3-1 1.

[5] Z. Shi and V. Govindaraju, “Skew Detection for Complex Document Imagedusing Fuzzy Runlength”, In Proc. of 7th Int. Conf. on Document Analysis and Recognition, 2003, pp. 715-719.

[6] F. M. Wahl, K. Y. Wong, R G. Casey, “Block segmentation and text extraction in mixed texthuge documents“, Computer Graphics and Image Processing, Vol. 20, 1982, pp. 375 - 390. 17.1 U. Pat and Sagarika Dana, ”Segmentation of Bangla Unconstrained Handwritten Text”, Proc. 7th Int. Conf. on Document Analysis and Recognition, 2003, pp. 1 128-1 132. 27 I.

[7] .U. Pal and P. P. Roy, ”Multi-oriented and curved text lines extraction from Indian documents“, IEEE Trans. on Systems, Man and Cybernetics- Part B, Vo1.34,2004, pp. 1676-1684.

[8] Andrew W. Senior, Anthony J. Robinson , “An Off-Line Cursive Handwriting Recognition System” . IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 3, march 2015

[9] Steve R.Gunn ,Technical report on ” Support Vector Machines for Classification and Regression “ , university of Southampton .

[10] Tutorial on “ Character recognition system for non experts. “ by Nawwaf N. Kharma & Rabab K. Ward , University of British Columbia.

[11] “An Introduction to Feature Extraction “ , Isabelle Guyon1 and Andr´e Elisseeff2 1 ClopiNet, 955 Creston Rd., Berkeley, CA 94708, USA.

[12] “Devnagari numeral recognition by combining decision of multiple connectionist classifiers “, reena bajaj, lipika dey and santanu chaudhury,Department of Electrical Engineering, Indian Institute of Technology, Department of Mathematics, Indian Institute of Technology, New Delhi 110016,India , Sadhana Vol. 27, Part 1, February 2002, pp. 59–72.

[13] Daniel I. Morariu, Lucian N. Vintan, and Volker Tresp “ Evolutionary Feature Selection for Text Documents using the SVM” International Journal of Applied Mathematics and Computer Sciences Volume 1 Number 1.

[14] “Determination of run-length smoothing values for document segmentation ” , N. Papamarkos, J. Tzortzakis and B. Gatos electric circuits analysis laboratory department of electrical & computer engineering democritus university of thrace 67100 xanthi, Greece.

[15] Manjunath Aradhya , “ Principal Component Analysis and Generalized Regression Neural Networks for Efficient Character Recognition “, ICETET-08,Nagpur,India.

[16] Hung-Ming Sun* , ”Enhanced Constrained Run-Length Algorithm for Complex Layout Document Processing International Journal of Applied Science and Engineering 2016.4, 3: 297-309.