ISSN:2231-0711

Sukhvir Kaur et al | IJCSET |May 2013 | Vol 3, Issue 5, 149-151

Improved XY cut Page Segmentation Algorithm for Border Noise

Sukhvir Kaur[1], Palvinder Singh Mann[2]

[1]*CT Institute of Engineering Management and Technology, Jalandhar, Punjab,India*

[2]*DAV Institute of Engineering & Technology, J alandhar,Punjab,India*

Abstract-Page segmentation is an important to analyse OCR Systems. In this paper we discussed page segmentation algorithms .We proposed new page segmentation algorithm for border noise .The implementation of the proposed algorithms is done and results are compared with the XY Cut page segmentation algorithm in case of border noise in OCR documents .

Keywords-*OCR, Border, Noise, Preprocessing, Segmentation, Threshold, Skew*

1. Optical Character Recognition

Optical Character Recognition is the automated process of translating an input document image into a symbolic text file. The input document images can be obtained from a large variety of media, such as journals, newspapers, magazines, memos, etc. The format of document image can be digitally created, faxed, scanned, machine printed, or handwritten, etc [33]. The output symbolic text file from an OCR system will include the text content of the input document image but also additional descriptive information, such as page layout, font size and style, document region type, confidence level for the recognized characters, etc.

Optical character recognition is considered as the most successful applications in the field of pattern recognition and artificial intelligence. Many commercial systems are available for performing OCR which exists for number of applications, although the machines are still not able to compete with human reading capabilities.

The main processes that exist in OCR's are

- Optical Scanning
- Preprocessing
- Document Analysis

Optical scanning process can be defined as a scanning process through which a digital image is captured from the original document. In OCR optical scanners are used, which generally consist of a transport mechanism and a Sensing device that converts light intensity into gray-levels [31]. Printed documents usually consist of black print on a white background.

Segmentation is a process that determines the elements of an image. The most important point which is necessary to locate the regions of the document where data is printed and distinguish is from figures and graphics.

The preprocessing stage which includes thresholding, binarizing, filtering, edge detection, gap filling, Segmentation and so on can make the initial image more suitable for later computation. The image resulting from the scanning process may contain a certain amount of noise [30]. Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters.

2. Page Segmentation

Page segmentation is a crucial preprocessing step in an OCR system. It is the process of dividing a document image into homogeneous zones, i.e., those zones only contains one type of information, such as text, a table, a figure, or a halftone image. In many cases, OCR system accuracy heavily depends on the accuracy of the page segmentation algorithm [10].

A page segmentation algorithm is used for particular document because of the preservation of small parts of documents. We can produce smaller blocks of document without horizontal scrolling, further this small block can be sent for further processing instead of the whole document. Straight borders of the documents can be easily preserved by using this method. Layout analysis is also preserved in the OCR so as to maintain reading order. In this way text can be differentiated from images in the OCR systems.

The task of page segmentation is to divide the document image into homogeneous zones, each consisting of only one physical layout structure (text, graphics, pictures etc).Therefore; the performance of optical character recognition (OCR) systems depends heavily on the page segmentation algorithm used. Over the last three decades, several page segmentation algorithms have been proposed.

Page segmentation algorithms can be categorized into three classes:Top-down approaches, Bottom-up approaches, Hybrid approaches.

The top - down approach recursively segment large regions in a document into smaller sub regions. The segmentation process stops when criterion is met and the ranges obtained at that stage constitute the final segmentation results. On the other hand, the bottom-up methods start by grouping pixels of interest and merging them into larger blocks or connected

components, such as characters which are then clustered into words, lines or blocks of text. The hybrid methods are the combination of both top-down and bottom-up strategies.

3. PROPOSED METHOD

XY Cut page segmentation algorithm is used to segment pages in OCR systems. When some noise come in document after scanning a document this mainly leads to segment this border noise into different segments .Whenever border noise come in scanned document this leads to improper segmentation of the document. This new algorithm will improve the Page segmentation process.
Improved XY Cut page segmentation is as follows:

Step 1) Border noise removal

1.1) Get the scanned document.

1.2)Select a pixel(X,Y) from document and get connected pixels corresponding to that for 8-neighbouring pixels get the value of pixels Left(X-1,Y), Right(X+1,Y) , Top(X,Y+1) , Bottom (X,Y-1) and four-Diagonal pixels {(X-1, Y-1) , (X+1,Y-1) , (X-1,Y+1) , (X+1,Y+1) }.

1.3) If all connected pixels are of color black then change all connected pixels to white color and continue this process until whole document is covered otherwise process next pixel and again go to step 1.2.

Step 2) Page Segmentation Process

2.1) Create Horizontal and vertical prefix sum table for OCR document.

2.2) Create histogram for the pixel values at each node.

2.3) Create a threshold value (Tx, Ty) corresponding to each axis i.e X axis and Y- axis compared it with histogram valleys (Vy & Vx) . If Vx > Tx or Vy > Ty split at midpoint so process will continue till condition occurs otherwise stop splitting of the process.

4. RESULTS AND COMPARISON

In process to implement improved XY Cut page segmentation algorithm first of all we scanned original document shown in fig 1.1.

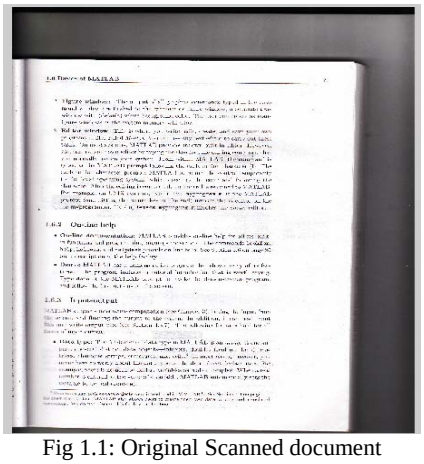

Fig 1.1: Original Scanned document

Then we remove border noise from this document and after removing noise we segment the page into different segment and after implementing the proposed algorithm we see that the results are better and with less noise as shows in fig: 1.2.

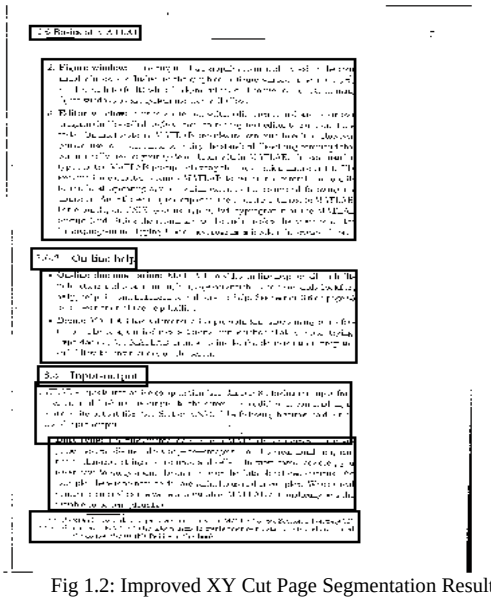

Fig 1.2: Improved XY Cut Page Segmentation Result

We compared existing XY Cut page segmentation with our improved XY Cut page segmentation algorithm using the metric noise ratio.
Noise ratio is defined as

Noise ratio =nfo /nfi (i)

Where nfo is number of the foreground pixels outside the ground-truth page frame and where nfi is the number of the foreground pixels inside the actual page content area of a document image. The noise ratio tells us how much border noise still remains in the document image relative to its actual contents. This measure evaluates how well the algorithm performs in removing the border noise.

Algorithm	Noise Ratio (%)
XY Cut algorithm	2.2527
Improved XY Cut algorithm	0.0678

Table 1.1

Table 1.1 shows that noise ratio in XY Cut algorithm is much more than the improved and high noise ratio in XY Cut algorithm shows that noise is present in document.

5. CONCLUSIONS

In this paper we discussed improved XY Cut Page segmentation algorithm used for OCR systems. In this method first we remove the border noise and then we do the segmentation process. Our results show that proposed algorithm performs better than XY Cut page segmentation algorithm in analyzing Noise ratio.

REFERENCES

[1] G. Nagy, S. Seth, and M. Viswanathan, "A Prototype *Document Image Analysis System for Technical Journals*," Computer, vol. 25, no. 7, pp. 10-22, July 1992.

[2] Fujisawa, Yasuaki Nankano, and Kiyomichi Kurino "*Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis*" Proceedings of The IEEE. Vol. 80. No. 7. July 1992.

[3] L. O'Gorman, "*The Document Spectrum for Page Layout Analysis*," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1162-1173, Nov. 1993.

[4] H.S. Baird, H. Bunke, P. Wang and H.S. Baird "*Background Structure in Document Images*," Document Image Analysis, eds., pp. 17-34, World Scientific, 1994.

[5] Sylwester and S. Seth, "*A Trainable, Single-Pass Algorithm for Column Segmentation*," Proc. Int'l Conf. Document Analysis and Recognition, pp. 615- 618, Aug. 1995.

[6] I. Guyon, R.M. Haralick, J.J. Hull and I.T. Phillips, "*Data Sets for OCR and Document Image Understanding Research*," Handbook of Character Recognition and Document Image Analysis, H. Bunke and P. Wang, eds., pp. 779-799, World Scientific, 1997.

[7] O. Okun, M. Pietikainen, and J. Sauvola, "Robust Skew Estimation on Low-Resolution Document Images," Proc. Fifth Int'l Conf. Document Analysis and Recognition, pp. 621-624, Sept. 1999.

[8] G. Nagy, "*Twenty Years of Document Image Analysis in PAMI*," IEEE Trans. Pattern Analysis and Machine Intelligence,Vol. 22, no. 1, pp. 38-62, Jan. 2000.

[9] Jianbo Shi and Jitendra Malik "*Normalized Cuts and Image Segmentation*" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, Aug 2000.

[10] S. Mao and T. Kanungo, "*Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms*," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23, no. 3, pp. 242-256, Mar. 2001

[11] S. Mao and T. Kanungo, "*Software Architecture of PSET: A Page Segmentation Evaluation Toolkit*," Int'l J. Document Analysis and Recognition, Vol. 4, no. 3, pp. 205-217, 2002.

[12] L. Cinque, S. Levialdi, L. Lombardi and S. Tanimoto, "*Segmentation of Page Images Having Artifacts of Photocopying and Scanning*," Pattern Recognition, Vol. 35, pp. 1167-1177, 2002.

[13] T.M. Breuel, "*High Performance Document Layout Analysis*," Proc. Symp. Document Image Understanding Technology, Apr. 2003.

[14] S. Marinai, E. Marino and G. Soda, "*Layout Based Document Image Retrieval by Means of XY Tree Reduction*," Proc. Eighth Int'l Conf. Document Analysis and Recognition, pp. 432-436, Aug. 2005.

[15] Jean-Luc Meunier " Optimized XY-Cut for Determining a Page Reading Order", Proceedings Eighth International Conference on Document Analysis and Recognition, 2005.

[16] Faisal Shafait, Daniel Keysers and Thomas M. Breuel "*Performance Comparison of Six Algorithms for Page Segmentation*" 7th IAPR Workshop on Document Analysis Systems, DAS'06. , Feb. 2006**.**

[17] S. Mandal, S. Chowdhury, A. Das and B. Chanda, "*A Simple and Effective Table Detection system from Document Images*," Int'l J. Document Analysis and Recognition, vol. 8, nos. 2-3, pp. 172-182, June 2006.

[18] F. Shafait, D. Keyser and T.M. Breuel, "*Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images*," Proc. 18th Int'lConf. Pattern Recognition, pp. 872-875, Aug. 2006.

[19] F. Shafait, J. van Beusekom , D. Keysers, and T.M. Breuel, "*Page Frame Detection for Marginal Noise Removal from Scanned Documents*," Proc. Scandinavian Conf. Image Analysis, pp. 651-660, June 2007

[20] N. Stamatopoulos, B. Gatos and A. Kesidis, "*Automatic Borders Detection of Camera Document Images*," Proc. Second Int'l Workshop Camera-Based Document Analysis and Recognition, pp. 71-78, Sept. 2007.

[21] D. Keysers, F. Shafait and T.M. Breuel, "*Document Image Zone Classification—a Simple High-Performance Approach*," Proc. Second Int'l Conf. Computer Vision Theory and Applications, pp. 44-51, Mar. 2007.

[22] F. Shafait, J. van Beusekom, D. Keysers and T.M. Breuel, "*Document Cleanup Using Page Frame Detection*," Int'l J. Document Analysis and Recognition, vol. 11, no. 2, pp. 81-96, 2008.

[23] T.M. Breuel, "*The OCRopus Open Source OCR System*," Proc. SPIE Document Recognition and Retrieval XV, pp. 0F1-0F15, Jan. 2008.

[24] F. Shafait, D. Keysers and T.M. Breue, "*Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms*," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 6, pp. 941-954, June 2008.

[25] G. Nagy, S.C. Seth, and M. Viswanathan, "*Projection Methods Require Black Border Removal*," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 4, p. 762, Apr. 2009.

[26] F. Shafait and T.M. Breuel, "*A Simple and Effective Approach for Border Noise Removal from Document Images*," Proc. 13th IEEE Int'l Multi-Topic Conf., Dec. 2009.

[27] http://unpaper.berlios.de/, 2010.

[28] Tranos Zuva, Oludayo O. Olugbara, Sunday O. Ojo and Seleman M. Ngwira "*Image segmentation, Available Techniques, Developments and Open Issues*" Canadian Journal on Image Processing and Computer Vision Vol. 2, No. 3, March 2011.

[29] Faisal Shafait and Thomas M. Breuel "*The Effect of Border Noise on the Performance of Projection-Based Page Segmentation Methods*" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 4, April 2011

[30] Sandhya.N , R. Krishnan , D. R. Ramesh Babu " *A language independent Characterization of Document Image Noise in Historical Scripts*" International Journal of Computer Applications (0975 – 8887) Vol. 50 – No.9, July 2012.

[31] Character Recognition is available on URL http://www.intechopen.com/books/character-recognition / preprocessing -techniques-in-character-recognition.

[32] www.ocropus.org

[33]http://www.computerworld.com/s/article/73023/ Optical_Character_Recognition