ISSN:0975-9646

Page Segmentation in OCR System- A Review

Sukhvir Kaur[1], P.S.Mann[2], Shivani Khurana[3]

[1,3]*CT Institute of Engineering, Management and Technology, Jalandhar*
[2]*DAV Institute of Engineering & Technology, Jalandhar*

***Abstract*— Optical character recognition is an active field for recognition pattern. In this paper we tried to present how processes work in OCR system, pre-processing in OCR and document analysis. To review the process for analysis pattern from document proper page segmentation should be done. So various categories of page segmentation algorithms are mentioned which are *Top Down, Bottom Up and Hybrid* in this paper.**

***Keywords*— OCR, Skew, Threshold, Segmentation**

I. 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.

Forms containing characters images can be scanned through scanner and then recognition engine of the OCR system interpret the images and turn images of handwritten or printed characters into ASCII data (machine-readable characters).

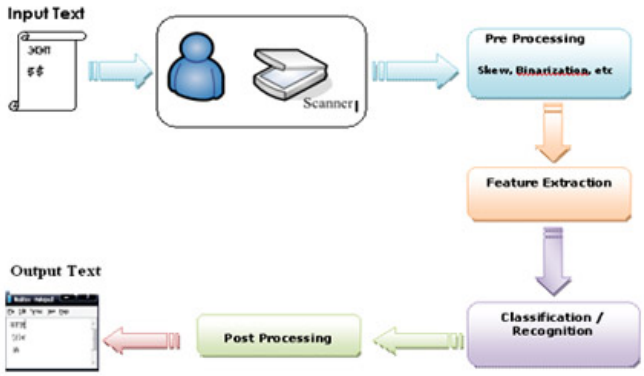

Figure 1.1 - Processes in OCR System.

1.1 Optical Scanning.

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. In performing OCR is considered as common practice to convert the multilevel image into a bi-level image of black and white. Often this process is also known as thresholding and is performed on the scanner to save memory space and computational effort.

The thresholding process is important as it is dependent on the quality of the bi-level image. A fixed threshold is used, this is a technique in which gray-levels below this threshold is said to be black and levels higher to that are white .In process to Obtain Good Results of Scanning paper quality should be good. An appropriate ink, drop out color improve the results.

1.2 Segmentation

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.

Text segmentation is the isolation of characters or words. Many segmentation algorithms in which segment words are used into isolated characters which are recognized individually. This process of segmentation is performed by isolating each connected component. This technique is easy to implement, but problems occurs if characters touch or if characters are fragmented and consist of several parts. The problems in segmentation are divided into various categories: Extraction of touching and fragmented characters, distinguishing noise from text, skewing.

1.3 Preprocessing:

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 .In addition to smoothing, preprocessing usually includes normalization. The

normalization is applied to obtain characters of uniform size, slant and rotation. In order to specify for rotation, the angle of rotation must be found. For rotated pages and lines of text, the variants of Hough transform are commonly used for detecting skewing. However, to find the rotation angle of a single symbol is not possible until after the symbol has been recognized.

II. DOCUMENT ANALYSIS

Document analysis is process of extracting text from the document. Reliable character segmentation and recognition depend upon both original document quality and registered image quality. Processes that attempt to compensate for poor quality originals and/or poor quality scanning include image enhancement, underline removal, and noise removal. Image enhancement methods emphasize character versus non-character discrimination. Underline removal erases printed guidelines and other lines which may touch characters and interfere with character recognition and noise removal erases portions of the image that are not part of the characters. In document analysis, it is necessary to isolate individual characters from the text image. Many OCR systems use connected components for this process.

III. 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.

The task of a geometric layout analysis system is to segment the document image into homogeneous zones, each consisting of only one physical layout structure, and to identify their spatial relationship (e.g. reading order). Therefore, the performance of layout analysis methods depends heavily on the page segmentation algorithm used [16].

The overall goal of layout analysis is to take the raw input image and divide it into non-text regions and "text lines"–sub images of the original page image that each contains a linear arrangement of symbols in the target language. Text lines need not be horizontal and left-to-right. The system must impose no constraints on the shape or direction of the text lines generated by the layout analysis. However, layout analysis must indicate the correct reading order for the collection of text lines. The text line recognizer needs to cope with the direction and the nature of text lines returned by page layout analysis.

Column segmentation is an essential part of any document analysis process. The XY CUT algorithm, attempts on column segmentation of page in which images are guided by a single parameter based on a nearly universal feature of page layout design, Here the column gaps are wider than word and character gaps. Further, this single parameter is expressed in relative terms and is therefore independent of the scanning resolution.

3.1 XY Cut Algorithms

The XY cut segmentation algorithm, also referred to as the recursive XY cuts (RXYC) algorithm and it is referred as tree-based top down algorithm [1]. The root of the tree represents the entire document page. All of the leaf nodes together represent the final segmentation. The RXYC algorithm recursively splits the document into two or more smaller rectangular zones, which represent the nodes of the tree. At each step of the recursion, the horizontal and vertical projection profiles of each node are computed. To compute the valleys in the projection profile histograms, noise removal thresholds tnx and tny are used. First, the thresholds tnx and tny are scaled linearly based on the current zone's width and height. Then, all bins of the histograms that contain values less than the scaled thresholds are set to zero. The valleys along the horizontal and vertical directions, vx and vy, are then compared to the corresponding predefined thresholds tx and ty. If the valley is larger than the threshold, the node is split at the midpoint of the wider of vx and vy into two children nodes. The process continues until no leaf node can be split further. [24]

When we apply page segmentation on XY algorithm it is very important to choose the thresholds. The control on the tolerance level along the horizontal and vertical directions should be such that there is less differences in overlap than the threshold value in that particular direction (i.e. horizontal and vertical). Otherwise it will lead to over segment and under segment.

IV. CONCLUSIONS

In this paper we discussed Page segmentation process, algorithm used for OCR systems. XY Cut algorithm is used for column segmentation. Preprocessing is crucial step for OCR systems and Document analysis which should be done properly to generate proper character sequences for pattern recognition process.

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