Page Segmentation for Manhattan and Non-Manhattan Layout Documents via Selective CRLA

Hung-Ming Sun
Department of Information Management, Kainan University, Taoyuan, Taiwan, R.O.C.
e-mail: sunhm@mail.knu.edu.tw

Abstract

The Constrained Run-Length Algorithm (CRLA) is a well-known technique for page segmentation. The algorithm is fast and can be used to partition documents with Manhattan layouts. It is not, however, suited to deal with pages with layouts beyond the Manhattan format, e.g. irregular halftone images embedded in text paragraphs. A modified version of the CRLA, named selective CRLA, is presented in this paper. The selective CRLA is capable of processing documents with both Manhattan and non-Manhattan layouts. The selective CRLA is performed twice with different sets of parameters on a label image derived from the input document image. After both of its executions, the yielded text regions are extracted. The proposed method has been successfully applied to extraction of text from commercial magazine pages with complicated layouts.

1. Introduction

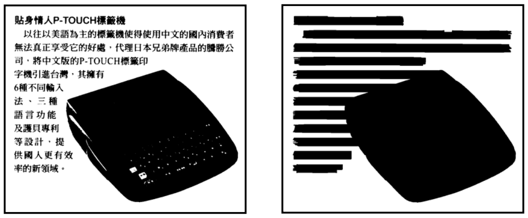
Page segmentation is a necessary step in a document processing system; its objective is to locate different types of contents such as text, graphics, and halftone images from the input document image. The extracted regions can then be processed by a subsequent step according to their types, e.g. OCR for text regions and compression for graphics and halftone images. Finally, the physical structure of the document can be resolved. A number of methods have been reported in the literature for segmenting document pages. These methods have different characteristics and are suitable for different types of documents. The Constraint Run-Length Algorithm (CRLA) [1], also known as the Run-Length Smoothing/Smearing Algorithm (RLSA), and the recursive X-Y cut algorithm [2] are two earlier techniques proposed for segmenting a document image into homogeneous regions. A limitation of both methods is that only Manhattan layout documents (i.e. the text/graphics/halftone-image regions are separable by horizontal and vertical line segments) are applicable.

Some other systems integrate connected component labeling techniques and grouping criteria to cluster foreground components to be text/graphics/halftone-image blocks [3-5]. Such methods can be applied only to documents with character size within a certain range because large characters, such as those in headlines, could be discarded in the geometric filtering step. Some texture-based approaches are also proposed [6, 7]. First, the texture features are calculated for each pixel and then a pixel-level classification scheme is used to group pixels based on their texture signatures and locations. Pixel-level classification, however, is computationally intense for such methods. Works by Lin *et al*. [8] and Etemad *et al*. [9] employ split-and-merge segmentation techniques to treat document images. A set of criteria is designed to check the homogeneity of a region. If the region is considered not homogenous enough, it is further divided into sub-regions. On the contrary, if adjacent regions have similar homogeneity, they are merged. Pavlidis *et al.* [10] present another solution to page segmentation, which searches for long white intervals on the vertical projection profiles to locate small column blocks. These column blocks are merged into larger ones and then grouped according to their alignments. Some other works [11-13] analyze the background structure instead of concerning the foreground objects. The existing methods for document image segmentation can be categorized into top-down, bottom-up, or hybrid approaches based on their processing hierarchies [13, 14].

Although the CRLA is an earlier technique, it is still one of the most popular methods for document analysis systems [15-19] because of its high speed and easy implementation. The CRLA is very efficient in the processing of Manhattan layout documents such as journal articles but it fails to deal with non-Manhattan layout pages (i.e. the text/graphics/halftone-image regions are arbitrarily shaped and in any arrangement), as Fig. 1 illustrates. In the present work, an improved version of the CRLA, named selective CRLA, is

Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)
1520-5263/05 $20.00 © 2005 IEEE

presented to extend its capability to cover non-Manhattan layout documents.



貼身情人P-TOUCH標籤機

以往以美語為主的標籤機使得使用中文的國內消費者無法真正享受它的好處，代理日本兄弟牌產品的騰勝公司，將中文版的P-TOUCH標籤印字機引進台灣，其擁有6種不同輸入法、三種語言功能及護貝專利等設計，提供國人更有效率的新領域。

(a) (b)

Fig. 1. (a) Non-Manhattan layout document; (b) processing result of the CRLA.

2. Page segmentation via selective CRLA

The selective CRLA is performed on a label image, which is derived from the input document image. To generate the label image, a connected component labeling algorithm [20] is used to locate the foreground components in the document image. Then, an additional scan is made to assign certain labels to the pixels of the foreground components according to their size. Three labels are defined in the present system.

(i) If the height of a foreground component is less than 1 cm, its pixels are assigned the label of "1".
(ii) If the height of a foreground component is between 1 cm and 3 cm, its pixels are assigned the label of "2".
(iii) If the height of a foreground component is larger than 3cm, its pixels are assigned the label of "3".

The constant values used above are set in the unit of centimeters to accommodate to the document images scanned at different resolutions. These values can be converted into representation by pixels according to the resolution of the input image.

After the label image is yielded, the selective CRLA is executed on it in the following manner. Consider, for example, a label string as below

11000111000200333000011000000111.

By setting a constraint $C = 5$ to the run length of zeros that are surrounded by label 1 on both sides, if the length of the consecutive zeros is not longer than C, these zeros are replaced by ones. Based on this operation, the preceding string is converted into

11111111000200333000011000000111.

This operation is referred to as *selective-CRLA{1}* where "selective" means it is applied only to those zeros surrounded by the specific label(s) given in the bracket. The selective CRLA is performed with a two-pass processing scheme on the label image to separate text from graphics and halftone-image components.

2.1. Process of the first pass

The following steps are used to accomplish page segmentation in the first pass.

1. Apply *selective-CRLA{1}* row by row to the label image using a constraint C_{hor-1}.
2. Apply *selective-CRLA{1}* column by column to the label image using a constraint C_{ver-1}.
3. Combine the images yielded by steps 1 and 2 using a logical *AND* operation.
4. Apply an additional *selective-CRLA{1}* row by row to the image yielded by step 3 using a constraint C_{sm-1}.

The parameters C_{hor-1}, C_{ver-1}, and C_{sm-1} are chosen as 3 cm, 3 cm, and 0.4 cm, respectively, in the present system. Fig. 2 shows an example after applying the aforementioned procedure to a non-Manhattan layout document. For comparison, Fig. 3 exhibits the execution result of the original CRLA for the same document. It can be seen that the erroneous linking between the text and the graphics is avoided by the new method.

A number of statistical measurements have been proposed for identifying text regions in binary document images [1, 15-17]. Two of them are adopted in the present system: (i) the mean length of horizontal black runs and (ii) the white-black transition count per unit width. These two features can be calculated by scanning a region from left to right in a row-by-row manner. As the scanning proceeds, the horizontal black run-length is accumulated by a counter, *BRL*, and the white-black transition count by another counter, *TC*. After the whole region is scanned, the two features are computed by

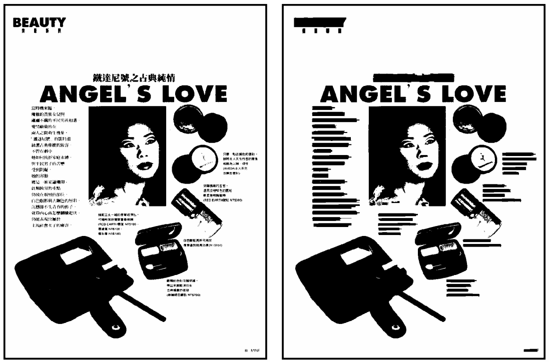
$$\text{mean length of horizontal black runs } MBRL = \frac{BRL}{TC}, \tag{1}$$

$$\text{white-black transition count per unit width } MTC = \frac{TC}{W} \tag{2}$$

where W is the width of the region under consideration. If $MBRL_{\min} \le MBRL \le MBRL_{\max}$ and $MTC_{\min} \le MTC \le MTC_{\max}$, the region is determined to be text. The parameter values used are $MBRL_{min} = 0.01$ cm, $MBRL_{max} = 0.4$ cm, $MTC_{min} = 1.0$, and $MTC_{max} = 3.8$ in the present system for the first pass.

The text regions extracted by the first pass are removed from the label image and the corresponding areas are filled with zeros, i.e. white. Fig. 4(a) shows the text extracted from the document image of Fig.

2(a); Fig. 4(b) shows the updated label image in which black pixels represent where non-zero label values exist.



(a) (b)

Fig. 2. (a) Non-Manhattan layout document; (b) result after applying the selective CRLA in the first pass.

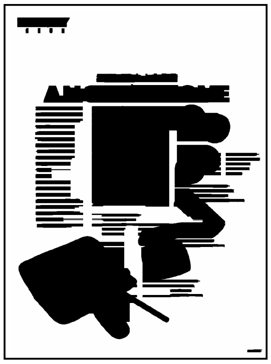
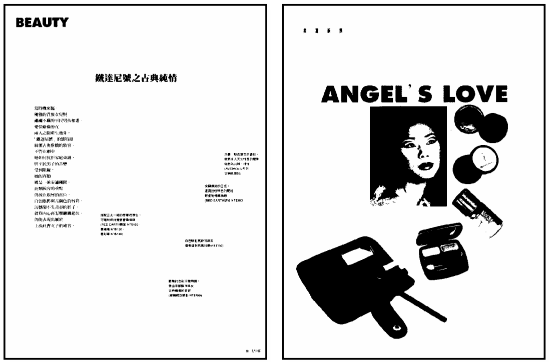


Fig. 3. Result after applying the original CRLA to the document of Fig. 2(a).



(a) (b)

Fig. 4. (a) Text extracted from the document of Fig. 2(a); (b) updated label image.

2.2. Process of the second pass

The algorithm used in the second pass is similar to that used in the first pass, except for some changes in parameter settings. Specifically, the following procedure is applied to the label image output from the first pass.

1. Apply selective-CRLA{1, 2} row by row to the label image using a constraint Chor-2.
2. Apply selective-CRLA{1, 2} column by column to the label image using a constraint Cver-2.
3. Combine the images yielded by steps 1 and 2 using a logical AND operation.
4. Apply an additional selective-CRLA{1, 2} row by row to the image yielded by step 3 using a constraint Csm-2.

Parameters C_{hor-2} = 3 cm , C_{ver-2} = 3 cm, and C_{sm-2} = 1.5 cm are used in the present system. Fig. 5 shows the result after performing the second pass procedure on the label image of Fig. 4(b). As illustrated, the headline characters, which are separated by wider space, are successfully united into an entity in the resulting image.

To identify the text regions, the features and rules used in the first pass is employed again here, but with parameters changed to $MBRL_{min}$ = 0.06 cm, $MBRL_{max}$ = 1.2 cm, MTC_{min} = 1.2, and MTC_{max} = 9.0. Fig. 6 shows the text extracted by the second pass and the left graphics/halftone-image regions.

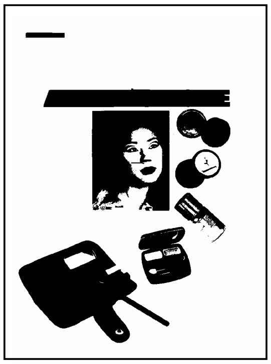


Fig. 5. Result after applying the second pass procedure to the label image of Fig. 4(b).

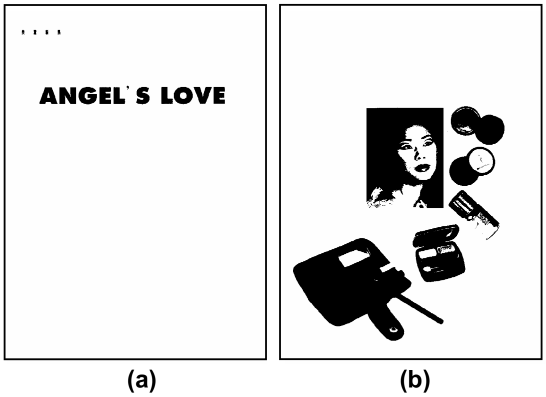


Fig. 6. (a) Text extracted by the second pass procedure; (b) graphics and halftone-image regions left after text extraction.

3. Experimental results

The proposed method is implemented on a Pentium 4/2.6G computer. To test its advantages over the original CRLA, the sample illustrated in Fig. 1 is also tried with the result shown in Fig. 7. As can be seen, this document can also be successfully processed by the new method. Indeed, this new page segmentation algorithm is integrated into a system for extracting text from color-printed documents [21]. That system aims to read commercial magazine pages, which commonly have complicated layouts such as a wide variety of character sizes, irregular graphics embedded in text paragraphs, and sporadically set legends. The following approach is used to deal with them. First, a color edge detection technique is used to find the edges in the document; then a binary edge image is produced. Next, the edge components in the binary edge image are assigned specific labels according to their size and thus a label image is created. Then the proposed selective CRLA is performed twice on the label image to yield regions for text extraction. Fig. 8 illustrates two color pages that have non-Manhattan layout, together with the text extracted from them.

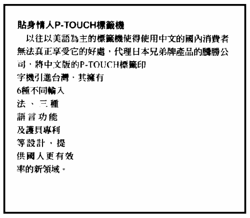

貼身情人P-TOUCH標籤機
以往以美語為主的標籤機使得使用中文的國內消費者
無法真正享受它的好處，代理日本兄弟牌產品的騰勝公
司，將中文版的P-TOUCH標籤印
字機引進台灣，其擁有
6種不同輸入
法、三種
語言功能
及護貝專利
等設計，提
供國人更有效
率的新領域。

Fig. 7. Results after applying our method to the document shown in Fig. 1(a).

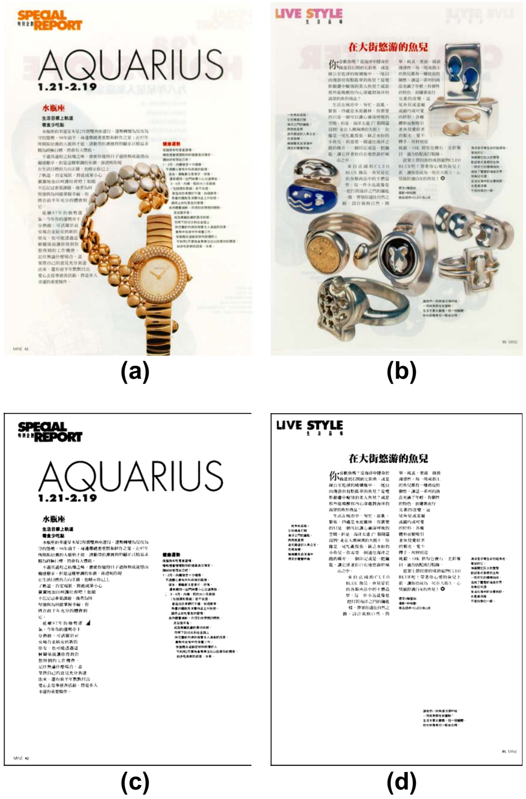


Fig. 8. (a) and (b) Color pages with non-Manhattan layout; (c) and (d) the processing results of our method.

4. Conclusion

A new page segmentation method based on the selective CRLA is proposed. The original CRLA is fast but can treat only Manhattan page layout; the present work extends its capability to cover both Manhattan and non-Manhattan layout types using a two-pass processing scheme. The first pass of the processing aims to extract body text, which usually has a smaller character size. The parameter values used in this pass can prevent erroneous linking of text regions to graphics regions. During the second pass, the parameter settings can join the widely spaced characters, such as those in headlines, to form a complete region for succeeding text identification.

The proposed selective CRLA is still fast in execution. To create the label image, only two sequential scans (one for labeling the foreground components and the other for assigning labels) on the document image are needed. Once this is done, the selective CRLA, which has nearly the same speed as the original CRLA, is applied to the label image. The yielded label image, however, requires additional storage space. Three labels are defined in the present system, so theoretically only two bits are needed to represent the assigned label for each pixel. In consideration of the need for simplicity of implementation and fast execution speed, we use eight bits (i.e. a byte) to represent the assigned label for each pixel; such implementation needs eight times the size of the document image. For instance, if the input image has a dimension of 1000 pixels by 1000 pixels, the storage space for the label image is 1Mbyte.

5. References

[1] F. M. Wahl, K. Y. Wong, and R. G. Casey, “Block Segmentation and Text Extraction in Mixed Text/Image Documents,” *Computer Graphics Image Processing*, vol. 20, 1982, pp. 375-390.

[2] G. Nagy and S. C. Seth, “Hierarchical Representation of Optically Scanned Documents,” Proc. 7th ICPR, 1984, pp. 347-349.

[3] L. A. Fletcher and R. Kasturi, “A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, 1988, pp. 910-918.

[4] L. O’Gorman, “The Document Spectrum for Page Layout Analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, 1993, pp. 1162-1173.

[5] A. Simon, J.-C. Pret, and A. P. Johnson, “A Fast Algorithm for Bottom-Up Document Layout Analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, 1997, pp. 273-277.

[6] A. K. Jain and S. Bhattachariee, “Text Segmentation Using Gabor Filters for Automatic Document Processing,” *Machine Vision and Applications*, vol. 5, 1992, pp. 169-184.

[7] P. S. Williams and M. D. Alder, “Generic texture analysis applied to newspaper segmentation,” Proc. 1996 IEEE Intl. Conf. Neural Networks, Washington DC, 1996, pp. 1664-1669.

[8] J. Lin, Y. Y. Tang, and C. Y. Suen, “Chinese Document Layout Analysis Based on Adaptive Split-and-Merge and Qualitative Spatial Reasoning,” *Pattern Recognition*, vol. 30, no. 8, 1997, pp. 1265-1278.

[9] K. Etemad, D. Doermann, and R. Chellappa, “Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, 1997, pp. 92-96.

[10] T. Pavlidis and J. Zhou, “Page Segmentation and Classification,” *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 6, 1992, pp. 484-496.

[11] H. S. Baird, “Background Structure in Document Images,” *Document Image Analysis*, World Scientific Publishing, 1994, pp. 17-34.

[12] A. Antonacopuolos and R. T. Ritchings, “Flexible Page Segmentation Using the Background,” Proc. 12th ICPR, 1994, pp. 339-344.

[13] Z. Chi, Q. Wang, and W.-C. Siu, “Hierarchical Content Classification and Script Determination for Automatic Document Image Processing,” *Pattern Recognition*, vol. 36, no. 11, 2003, pp. 2483-2500.

[14] G. Nagy, “Twenty Years of Document Image Analysis in PAMI,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, 2000, pp. 38-62.

[15] F. Y. Shih and S.-S. Chen, “Adaptive Document Block Segmentation and Classification,” *IEEE Trans. System Man and Cybernetics-PART B: Cybernetics*, vol. 26, no. 5, 1996, pp. 797-802.

[16] J. L. Fisher, S. C. Hinds, and D. P. D’amato, “A Rule-Based System for Document Image Segmentation,” Proc. 10th ICPR, 1990, pp. 567-572.

[17] F. Y. Shih, S.-S. Chen, D. C. D. Hung, and P. A. Ng, “A Document Segmentation, Classification and Recognition System,” Proc. IEEE Intl. Conf. System Integration, 1992, pp. 258-267.

[18] J. Xi, J. Hu, and L. Wu, “Page Segmentation of Chinese Newspapers,” *Pattern Recognition*, vol. 35, no. 12, 2002, pp. 2695-2704.

[19] K. Hadjar and R. Ingold, “Arabic Newspaper Page Segmentation,” Proc. Seventh ICDAR, Edinburgh, Scotland, 2003, pp. 895-899.

[20] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, 1992.

[21] H. M. Suen and J. F. Wang, “Text string extraction from images of colour-printed documents,” *IEE Proc. Vision, Image and Signal Processing*, vol. 143, no. 4, 1996, pp. 210-216.