Enhanced Constrained Run-Length Algorithm for Complex Layout Document Processing

Hung-Ming Sun*

Department of Information Management, Kainan University,
No. 1 Kainan Road, Luchu, Taoyuan County, 33857, Taiwan, R.O.C.

Abstract: The Constrained Run-Length Algorithm (CRLA) is a well-known technique for page segmentation. The algorithm is very efficient for partitioning documents with Manhattan layouts but not suited to deal with complex layout pages, e.g. irregular graphics embedded in a text paragraph. Its main drawback is to use only local information during the smearing stage, which may lead to erroneous linkage of text and graphics. This paper presents a solution to this problem by adding global information into the process of the CRLA. This enhanced CRLA can be applied to non-Manhattan page layout successfully. It can also extract text surrounded by a box. Both cases cannot be processed by the original CRLA.

Keywords: constrained run-length algorithm; page segmentation; document processing.

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. Earlier studies focused mainly on treating technical articles. The layout of such articles was generally in the Manhattan format (i.e. the text/graphics/halftone-image regions were separable by horizontal and vertical line segments, for example two-column text together with rectangle-alignment graphics and halftone images). The later studies on the document processing field started trying to resolve more complex layout types, for example irregularly shaped graphics embedded in text (i.e. non-Manhattan layout) and mixed horizontal/vertical text lines in one page.

The Constrained 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 are applicable. Some other systems integrate the connected component labeling technique and a set of grouping criteria to cluster foreground components to form text/graphics/halftone-image blocks [3-5]. Such methods can be applied only to documents with character size within a certain range because too large characters, such as

* Corresponding author; e-mail: sunhm@mail.knu.edu.tw

Accepted for Publication: December 31, 2006

those in headlines, could be misclassified as graphics. Some texture-based approaches are also proposed [6, 7]. The texture features are calculated at 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. The work done by Lin et al. [8] employs a split-and-merge segmentation scheme 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. [9] 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 analyze the background structure instead of concerning the foreground objects [10, 11]. The existing methods for document image segmentation can be categorized into top-down, bottom-up, or hybrid approaches based on their processing hierarchies [11, 12].

The CRLA is proposed originally for Manhattan layout document processing. Although there have been many other different methods proposed for dealing with both Manhattan and non-Manhattan layout pages, the CRLA is still one of the most popular techniques for building document analysis systems because of its high execution speed and easy implementation [13-17]. In this paper, an enhanced version of the CRLA is presented to extend its capability to cover non-Manhattan layout documents; this enhanced CRLA is still fast in execution and simple for implementation.

The remaining part of this paper is organized as follows. Section 2 discusses some existing methods for processing non-Manhattan layout documents; some difficult problems which may exist in various documents are also described therein. Section 3 explains the details of the proposed method for complex layout document processing. Experimental results are given in Section 4 and finally discussion is provided in Section 5.

2. Methods for non-Manhattan page segmentation

Both the CRLA and the recursive X-Y cut algorithm involve a scanning stage which checks the document image in horizontal and vertical directions, because they assume that the boundaries between text regions, graphics, and halftone images are either in horizontal or vertical orientation. Such methods have high execution speed as they test only two directions during processing. However, their disadvantage is that only Manhattan layout is applicable. To cope with the arbitrarily oriented boundaries existing in non-Manhattan page layout, more sophisticated techniques are proposed instead of scanning document images along horizontal and vertical directions.

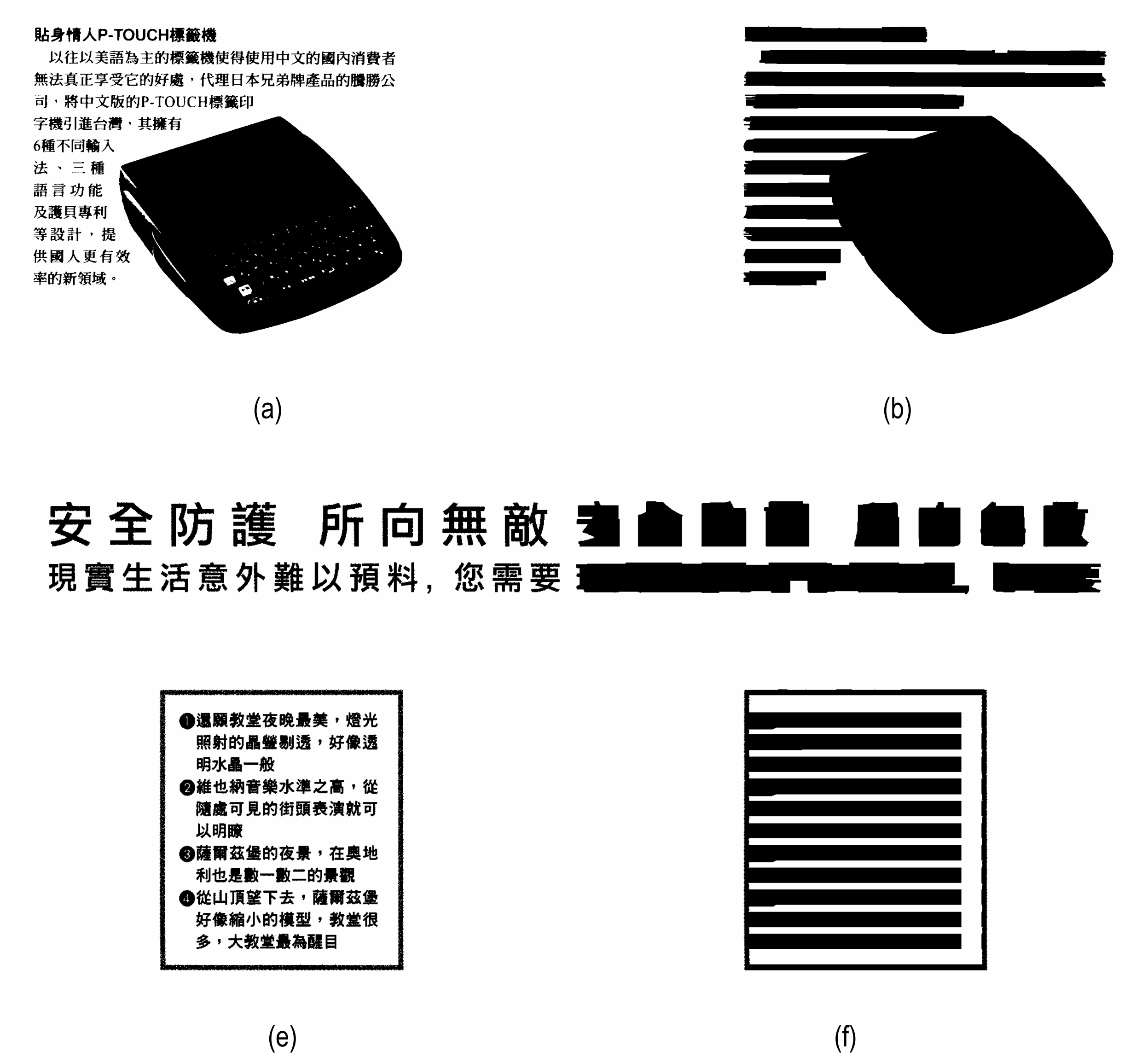
The method proposed by Etemad et al. [18] is able to process non-Manhattan layout documents even if the document image is skewed. Their method is based on texture analysis. A multi-scale wavelet packet is used as a feature vector, and a set of feed-forward multilayer neural networks are utilized in cooperation with a decision integration scheme to segment and identify various document objects. This approach sometimes results in sparse misclassified regions, which need post-processing to remedy. Besides, as mentioned by the authors, its performance depends highly on the right choice of a rich training set because of large intra-class and small inter-class variations in the feature space. In terms of complexity, the authors suggest that one may use other page segmentation algorithms for simple layout documents and their method for complex layout documents. Antonacopoulos [19] introduces another approach to complex layout page segmentation, which analyzes document background instead of concerning the foreground. His method reconstructs the background space by means of white tiles so that the contour of the printed regions can be extracted by tracing the edges of the white tiles appropriately.

Kise et al. [20] use approximated area Voronoi diagrams to resolve non-Manhattan layout pages. Their method is based on connected component analysis. To overcome the component boundaries that cannot be represented by vertical and horizontal line segments, the neighborhood of connected components is expressed with polygons. Their work also includes a detailed comparison with other related methods, which are categorized as either foreground analysis or background analysis. Basically, foreground analysis methods utilize connected components as primitives to construct document objects. Such methods tend to miss the extraction of larger characters because they may be regarded as graphics at a pre-filtering step. By contrast, background analysis is applicable to more complex page layout but it may over-segment a text line if it has wider inter-word spacing than usual. The system of Xiao et al. [21] represents the location of connected components with their centroids and the page structure is described as a set of points in a two-dimensional space. After imposing Delaunay triangulation and triangular features on these points, the text areas can be clustered and identified. Both Kise et al.'s method and this method use geometric computation to conduct page segmentation. However, the work of Xiao et al. shows that their method can avoid erroneous splits for the text lines that have wider inter-word gaps. Chi et al. [22] propose a modified version of the background thinning algorithm. Their method reduces the document image size to one-sixteenth of the original size (i.e. width and height are both reduced by one-forth) to speed up processing. However, reducing image size takes the risk of making regions of different types connected.

As mentioned previously, non-Manhattan page layout, large headline characters, and wide inter-word spacing are some difficulties while dealing with document images. In addition, the box-surrounded text paragraph is also common for magazine articles but it is rarely discussed by the existing page segmentation methods. The original CRLA has problems to process non-Manhattan page layout, text lines with wider inter-word gaps, and box-surrounded text paragraphs, as illustrated in Figure 1. The errors happen due to the smearing stage in which it checks only local pixels without caring for the global components related to the pixels. This paper presents a solution to this weakness by adding global component information into the process of the CRLA and this enhanced CRLA can overcome the failure of the original CRLA while treating these difficult cases.

3. Page segmentation via enhanced CRLA

The proposed new version of CRLA is performed on a label image, which is derived from the input document image and has the same size as the document image. In a binary document image, every pixel carries only foreground/background information, i.e. one means foreground and zero means background or vice versa. In the label image, every pixel can have various values; zero means background and non-zero means not only the foreground but also the component size related to the pixel. To generate the label image, a connected component labeling algorithm [23] is used to locate the foreground components in the document image and calculate their size. Then the label image can be generated by assigning its pixels with certain labels according to the size of the foreground component related to the pixel. Three labels are defined in the present system.



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。

安全防護 所向無敵

現實生活意外難以預料，您需要

❶還願教堂夜晚最美，燈光照射的晶鑾剔透，好像透明水晶一般

❷維也納音樂水準之高，從隨處可見的街頭表演就可以明瞭

❸薩爾茲堡的夜景，在奧地利也是數一數二的景觀

❹從山頂望下去，薩爾茲堡好像縮小的模型，教堂很多，大教堂最為醒目

Figure 1. Difficult documents that the original CRLA fails to handle: (a) non-Manhattan page layout; (b) processing result of (a) where text and image are linked; (c) text lines having wider inter-word gaps; (d) processing result of (c) where the upper text line cannot be united; (e) box-surrounded text paragraph; (f) processing result of (e) where text and box cannot be separated

(1) If the height of the foreground component is less than 1 cm, its pixels in the label image are assigned the label of “1”.
(2) If the height of the foreground component is between 1 cm and 3 cm, its pixels in the label image are assigned the label of “2”.
(3) If the height of the foreground component is larger than 3cm, its pixels in the label image 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 scanning resolution of the input image. The height is used instead of the width when checking the component size because it is assumed that the text on the document is horizontal alignment. As the characters in a text line may be connected in the scanned document image, checking width may lead to misjudgment of the component size.

After the label image is yielded, the new 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 braces. The selective CRLA is performed with a two-pass processing scheme on the label image to separate text from graphics and halftone-image components.

3.1. Process of the first pass

The process of the first pass uses the following steps to accomplish page segmentation.

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 purposes of steps 1 and 2 are to link the foreground components horizontally and vertically. Step 3 is to cut the joined foreground areas according to the space between columns and text lines. Finally, step 4 is to remove the small gaps that may exist between characters in a text line. The parameters C_{hor-1}, C_{ver-1}, and C_{sm-1} are chosen experimentally as 3 cm, 3 cm, and 0.4 cm, respectively, in the present system. Figure 2 shows the results after applying the aforementioned steps to a non-Manhattan layout document. For comparison, Figure 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 happening for the original CRLA is avoided by the new method.

After dividing the document image into homogeneous regions, these regions must be further classified into text or non-text according to their features. A number of statistical measurements have been proposed for distinguishing between text and non-text regions for binary document images [1, 13-15, 24]. 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 mean length of horizontal black runs

$$MBRL = \frac{BRL}{TC} \qquad (1)$$

white-black transition count per unit width

$$MTC = \frac{TC}{W} \qquad (2)$$

where W is the width (in pixel) of the region under consideration. If $MBRL_{\min} \le MBRL \le MBRL_{\max}$ and

$MTC_{\min} \leq MTC \leq MTC_{\max}$, the region is determined to be text. The parameter values are $MBRL_{min}$ = 0.01 cm, $MBRL_{max}$ = 0.4 cm, MTC_{min} = 1.0, and MTC_{max} = 3.8 in the present system. These values are determined based on our tests performed on dozens of document images.

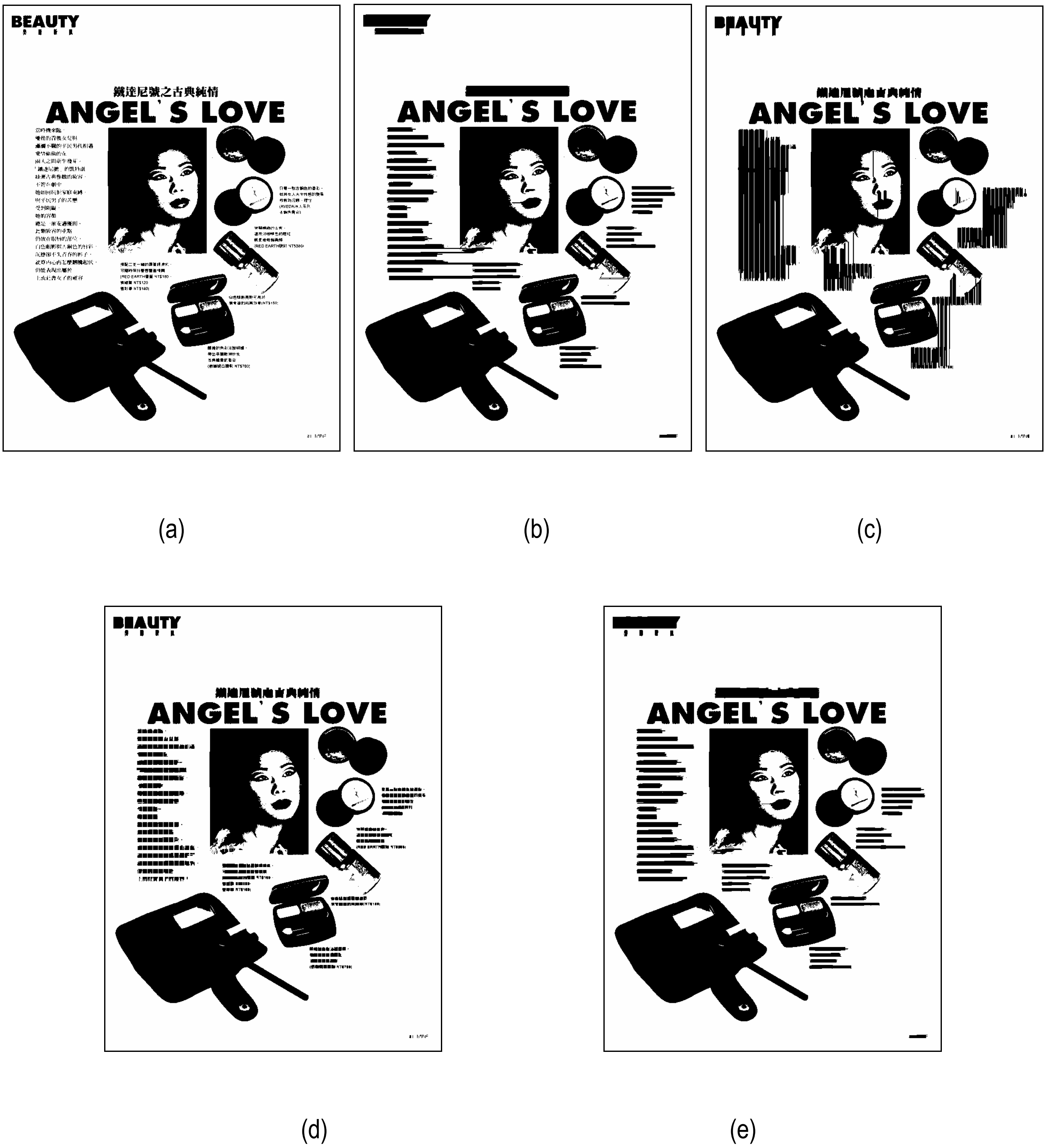


Figure 2. (a) Non-Manhattan layout document; (b) result after applying *selective-CRLA{1}* row by row to (a); (c) result after applying *selective-CRLA{1}* column by column to (a); (d) result after combining (b) and (c) using *AND* logic operation; (e) result after applying *selective-CRLA{1}* row by row to (d)

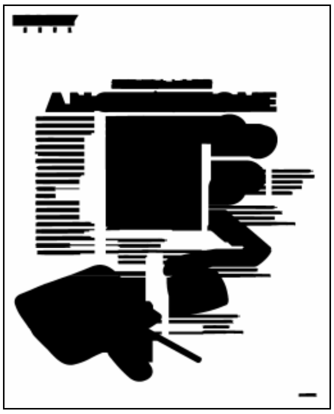


Figure 3. Result after applying the original CRLA to the document of Figure 2(a)

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. Figure 4(a) shows the text extracted from the document image of Figure 2(a); Figure 4(b) shows the updated label image in which black pixels represent non-zero label values.

3.2. 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 C_{hor-2}.
2. Apply selective-CRLA{1, 2} column by column to the label image using a constraint C_{ver-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 C_{sm-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. Figure 5 shows the results after performing the above procedure on the label image of Figure 4(b). As illustrated, the headline characters, which are separated by wider space, are successfully united into an entity in the resulting image.

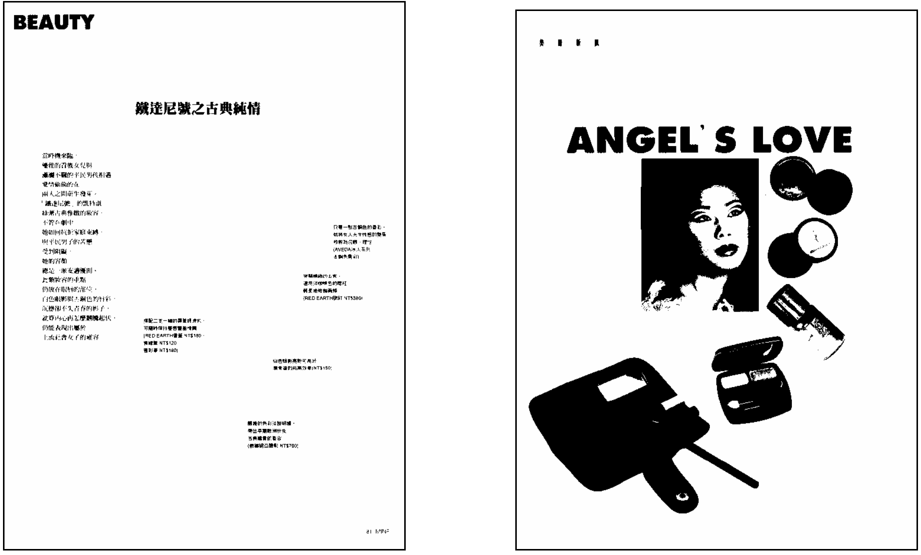
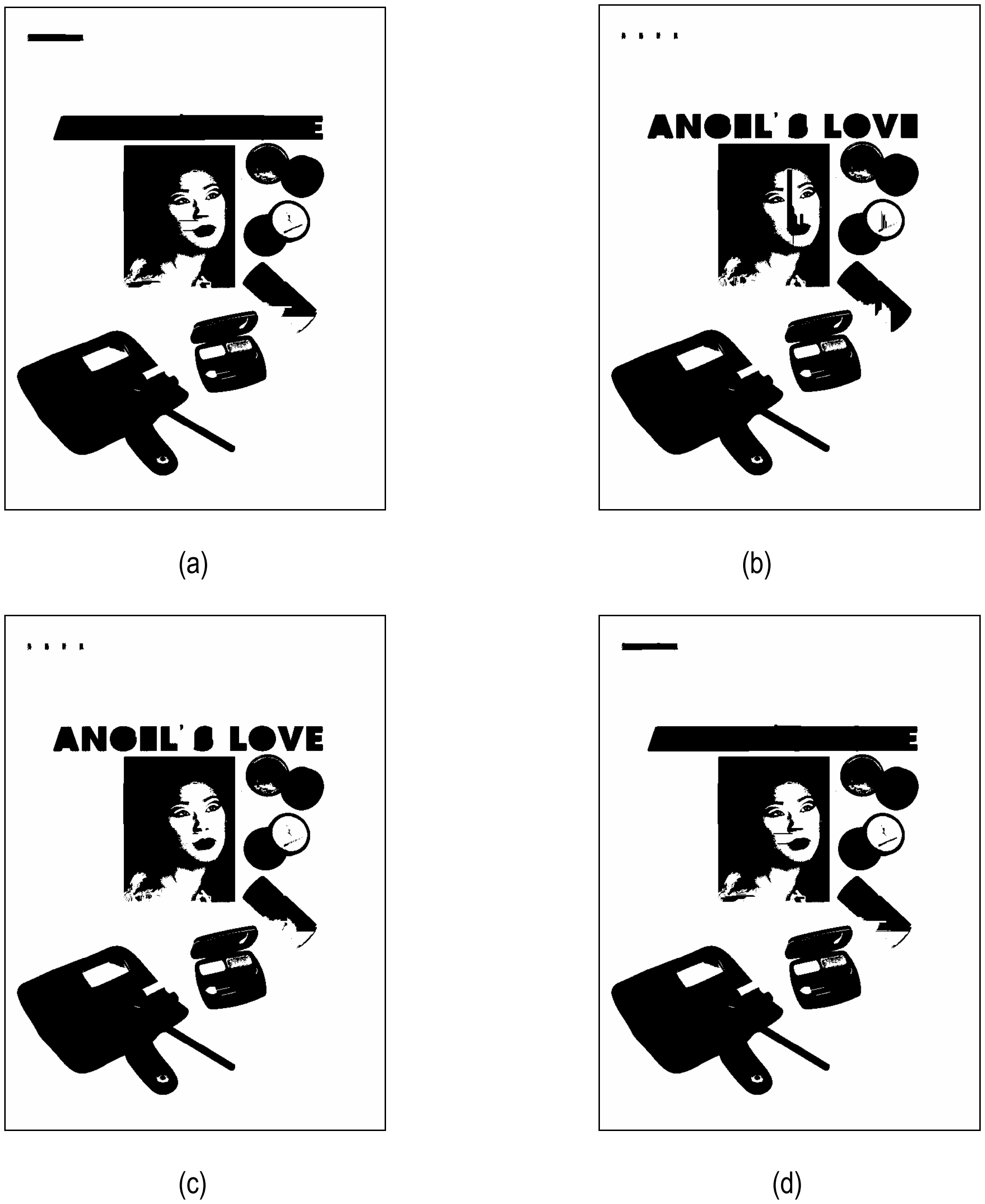


Figure 4. (a) Text extracted from the document of Figure 2(a); (b) updated label image



(a) (b)

(c) (d)

Figure 5. (a) Result after applying *selective-CRLA{1, 2}* row by row to the label image of Figure 4(b); (b) result after applying *selective-CRLA{1, 2}* column by column to the label image of Figure 4(b); (c) result after combining (a) and (b) using *AND* logic operation; (d) result after applying *selective-CRLA{1,2}* row by row to (c)

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. Figure 6 shows the text extracted by the second pass and the remaining graphics/halftone-image regions.

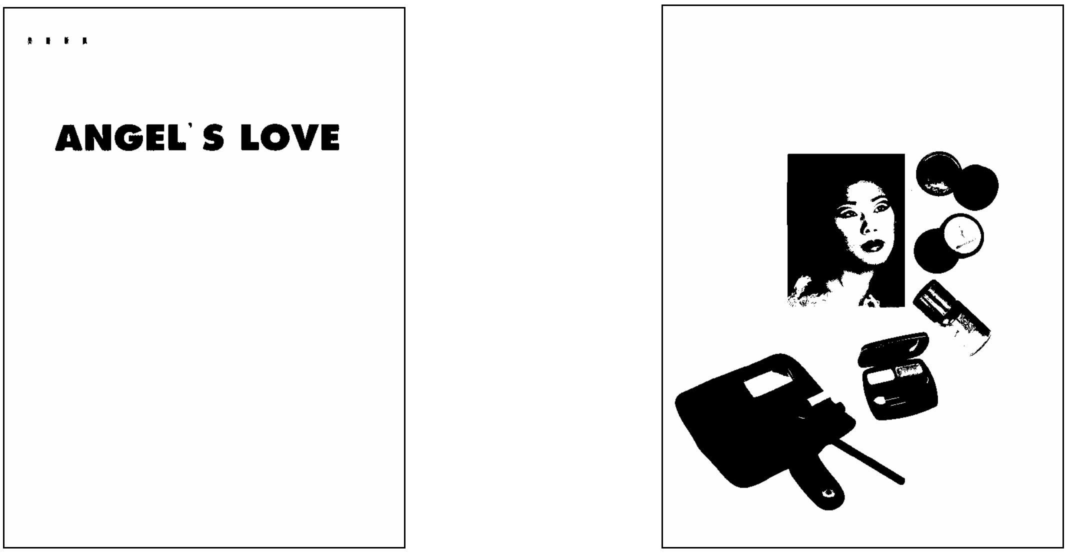


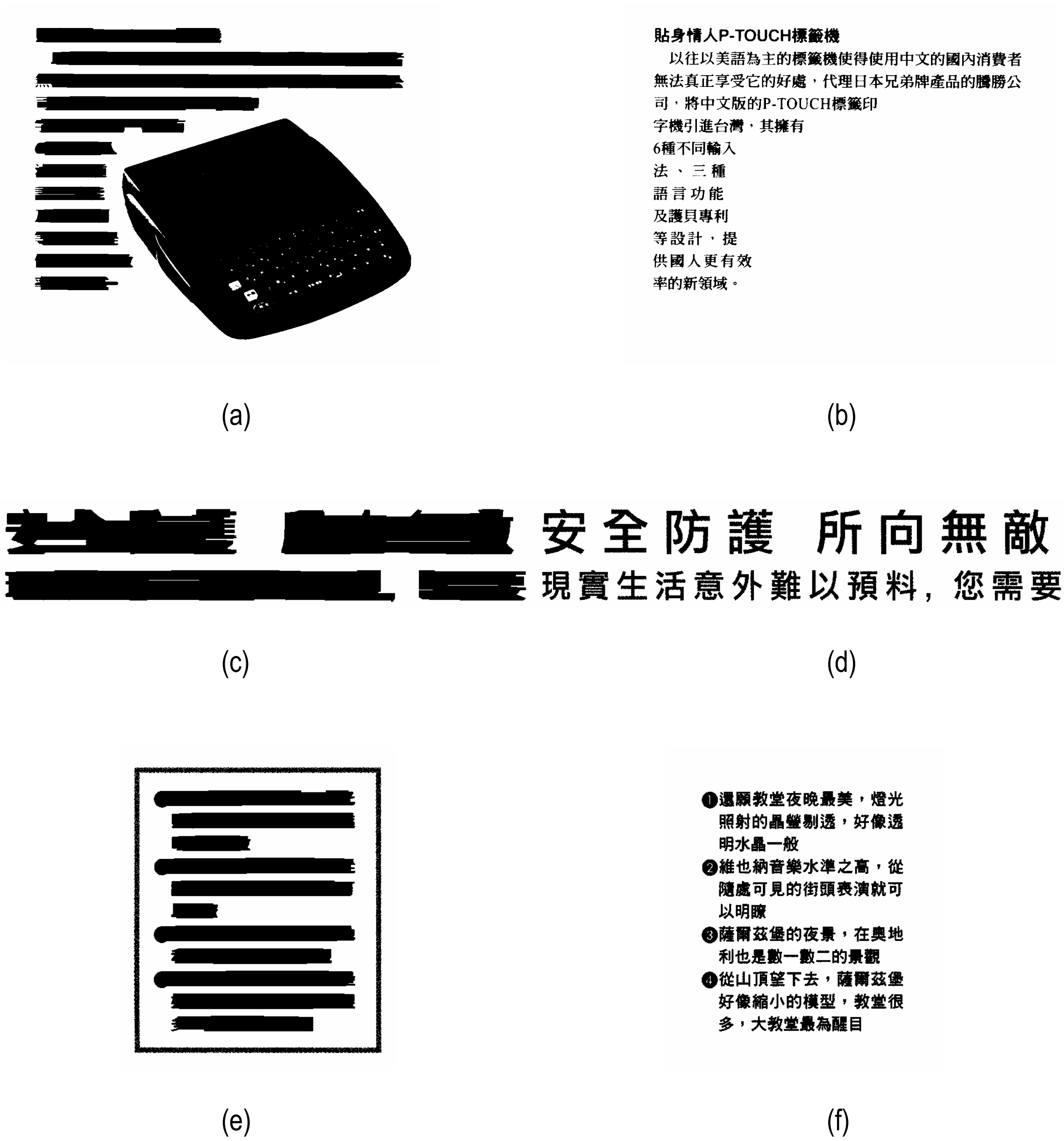
Figure 6. (a) Text extracted by the second pass; (b) graphics and halftone-image regions left after text extraction

4. Experimental results

To test the advantages of the proposed selective CRLA, the difficult documents illustrated in Figure 1 are tried and their results are shown in Figure 7. As can be seen, the new method is able to prevent the mis-linkage of text and non-text regions and meanwhile properly unite the widely spaced headline characters. Also, the box-surrounded text paragraphs can be separated and extracted.

In general, commercial magazine articles have more wide variety in their page layout than technical documents. For instance, the character size of them can change with a wide range; the graphics can be in an irregular shape and embedded in text paragraphs, and a page can even be dominated by graphics with sporadically set text legends. Figure 8 shows two of such examples and their processing results. It can be seen that the proposed method can successfully extract text from these challenging cases. However, some misjudged graphic segments may appear occasionally in the extracted text as shown in Figure 8(d). Such errors are due to the disconnected graphic fractions existing in the original documents, whose pixels are assigned label “1” or “2” because of their small component size. There are two approaches to remedying such mistakes. If the graphic fractions are standing alone without linking to any text after performing the page segmentation procedure, they can be filtered out by improving the text region identification stage (e.g. employing more features to check). On the other hand, if the graphic fractions are linked to text after the page segmentation procedure, an additional post-process is needed to split the erroneous linkage of text and graphics.

It should be noted that the document images included in the experiment are mainly collected from magazine articles, in which the text orientation is horizontal, and they are scanned without significant skew. If the input document image may have severe skew, a skew correction preprocess [12] must be done before applying the proposed page segmentation method. Another issue that should be addressed is the setting of the parameters used in the method. The parameter values of our system have been chosen empirically to suit to the document types in our test database. If different sorts of documents (e.g. technical articles or forms) are to be processed, the parameter values may need to be tuned accordingly. However, finding a set of parameter values which can treat various document types equally well can be one of our further research topics.



(a)

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。

(b)

(c)

安全防護 所向無敵
現實生活意外難以預料，您需要

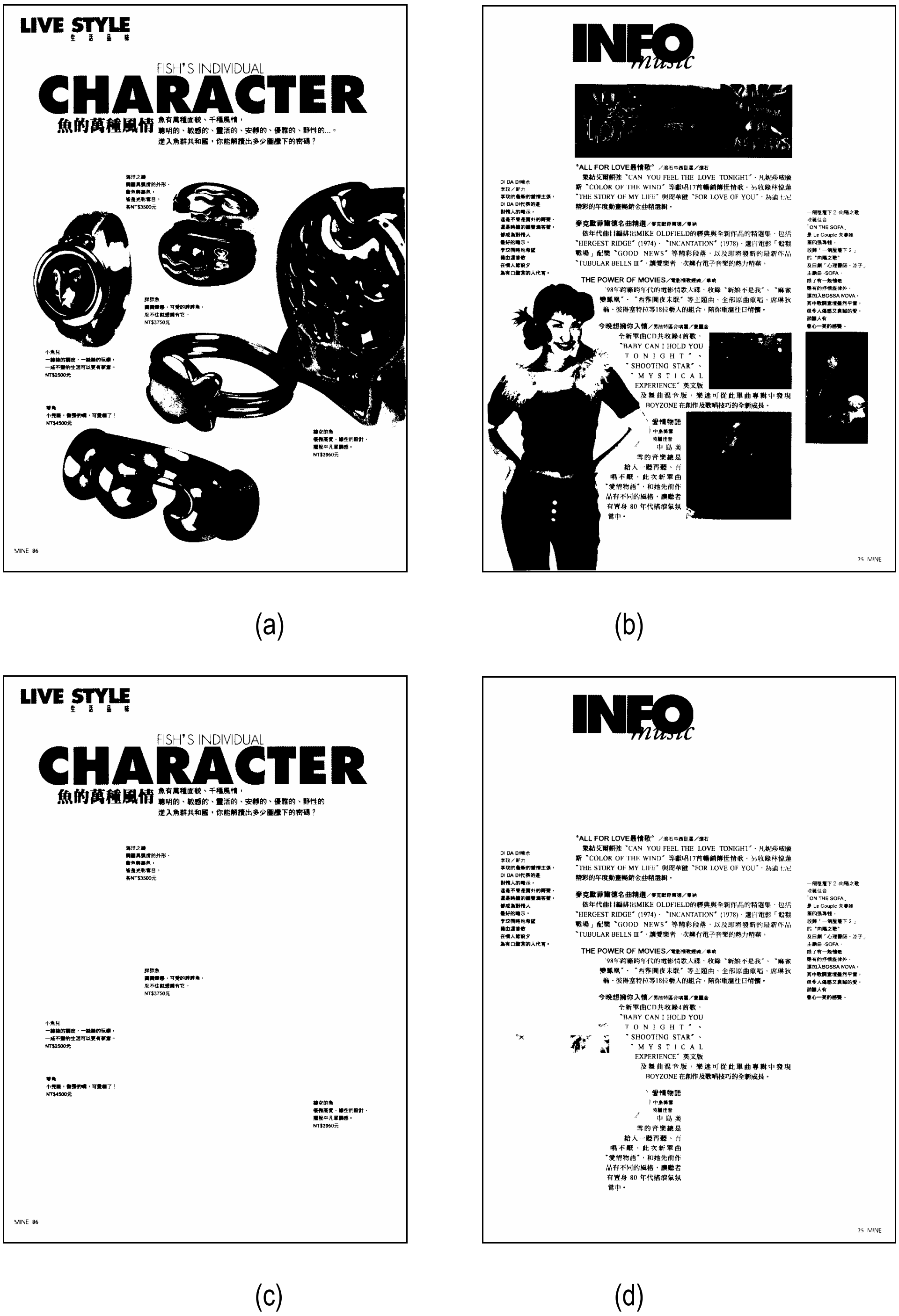
(d)

(e)

❶還願教堂夜晚最美，燈光
照射的晶瑩剔透，好像透
明水晶一般
❷維也納音樂水準之高，從
隨處可見的街頭表演就可
以明瞭
❸薩爾茲堡的夜景，在奧地
利也是數一數二的景觀
❹從山頂望下去，薩爾茲堡
好像縮小的模型，教堂很
多，大教堂最為醒目

(f)

Figure 7. (a) Results after applying selective CRLA to the document of Figure 1(a); (b) text extracted from the document of Figure 1(a); (c) results after applying selective CRLA to the document of Figure 1(c); (d) text extracted from the document of Figure 1(c); (e) results after applying selective CRLA to the document of Figure 1(e); (f) text extracted from the document of Figure 1(e)



(a) (b)

(c) (d)

Figure 8. (a) Magazine page that has wide variety of character size and is graphic-dominant with sporadically set text legends; (b) magazine page with non-Manhattan layout; (c) processing results of (a); (d) processing results of (b)

5. Discussion

As mentioned in Section 1, the CRLA has been chosen to build a number of document processing systems because of its advantages compared with other page segmentation techniques. For example, although the texture-analysis-based approaches are powerful to handle various page layout, they are commonly time-consuming because of the pixel-level classification. The methods based on connected component analysis may have problems to extract large headline characters and thus they suit to deal with documents of certain character size. The background-analysis-based approaches are also robust for complex page layout but they involve geometric computation which is much more complicated than the simple scanning process as the CRLA does. However, the original CRLA can only treat Manhattan page format; the present work extends its capability to cover both Manhattan and non-Manhattan layout and thus expands its application domain significantly.

The strength of the selective CRLA relies on the utilization of the global component information (i.e. the component size), which conducts the run-length smearing process more precisely than the original CRLA. In the first pass, the selective CRLA aims to extract body text, which usually has a smaller character size, and it can prevent erroneous linkage of text and non-text regions. In the second pass, the pixel labels and the parameter settings are relaxed to join the large headline characters and the text lines that have wide inter-character space. Experimental results demonstrate that the new method is capable of handling non-Manhattan layout pages and the difficult documents shown in Figure 1.

The execution speed of the selective CRLA is still fast. To create the label image, only two sequential scans (one for locating 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 implementation simplicity and execution speed, we use eight bits (i.e. one 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 1,000 pixels by 1,000 pixels, the storage space for the label image is 1Mbyte. As this is a very small amount of memory storage, it will not be a problem for implementation.

References

[1] Wahl, F. M., Wong, K. Y., and Casey, R. G. 1982. Block segmentation and text extraction in mixed text/image documents. *Computer Graphics Image Processing*, 20: 375-390.

[2] Nagy, G. and Seth, S. C. 1984. Hierarchical representation of optically scanned documents. *In Proceedings*. *7th ICPR*, Montreal, 347-349.

[3] Fletcher, L. A. and Kasturi, R. 1988. A robust algorithm for text string separation from mixed text/graphics Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10, 6: 910-918.

[4] O’Gorman, L. 1993. The document spectrum for page layout analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15, 11: 1162-1173.

[5] Simon, A., Pret, J.-C., and Johnson, A. P. 1997. A fast algorithm for bottom-up document layout analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19, 3: 273-277.

[6] Jain, A. K. and Bhattachariee, S. 1992. Text segmentation using Gabor filters for automatic document processing. *Machine Vision and Applications*, 5: 169-184.

[7] Williams, P. S. and Alder, M. D. 1996. Generic texture analysis applied to newspaper segmentation. *In Proceedings*. 1996 *IEEE International*. *Conference*. *Neural Networks*, Washington DC, 1664-1669.

[8] Lin, J., Tang, Y. Y., and Suen, C. Y. 1997. Chinese document layout analysis based on adaptive split-and-merge and

qualitative spatial reasoning. *Pattern Recognition*, 30, 8: 1265-1278.

[9] Pavlidis, T. and Zhou, J. 1992. Page segmentation and classification. *CVGIP: Graphical Models and Image Processing*, 54, 6: 484-496.

[10] Baird, H. S. 1994. Background structure in document images. "*Document Image Analysis"*, World Scientific Publishing, 17-34.

[11] Chi, Z., Wang, Q., and Siu, W.-C. 2003. Hierarchical content classification and script determination for automatic document image processing. *Pattern Recognition*, 36, 11: 2483-2500.

[12] Nagy, G. 2000, Twenty years of document image analysis in PAMI. *IEEE Transfusion*. *Pattern Analysis and Machine Intelligence*, 22, 1: 38-62.

[13] Shih, F. Y. and Chen, S. S. 1996. Adaptive document block segmentation and classification. *IEEE Transfusion*. *System Man and Cybernetics-PART B: Cybernetics*, 26, 5: 797-802.

[14] Fisher, J. L., Hinds, S. C., and D'amato, D. P. 1990. A rule-based system for document image segmentation. *In Proceedings*. *10th ICPR*, Atlantic, 567-572.

[15] Shih, F. Y., Chen, S. S., Hung, D. C. D., and Ng, P. A. 1992. A document segmentation, classification and recognition system. *In Proceedings*. *IEEE International*. *Conference*. System Integration, Morristown, NJ, 258-267.

[16] Xi, J., Hu, J., and Wu, L. 2002. Page segmentation of Chinese newspapers. *Pattern Recognition*, 35, 12: 2695-2704.

[17] Hadjar, K. and Ingold, R. 2003. Arabic newspaper page segmentation. *Process. 7th ICDAR*, Edinburgh, Scotland, 895-899.

[18] Etemad, K., Doermann, D., and Chellappa, R. 1997. Multiscale segmentation of unstructured document pages using soft decision integration. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19, 1: 92-96.

[19] Antonacopuolos, A. 1998. Page segmentation using the description of the background. *Computer Vision and Image Understanding*, 70, 3: 350-369.

[20] Kise, K., Sato, A., and Iwata, M. 1998. Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding*, 70, 3: 370-382.

[21] Xiao, Y. and Yan, H. 2003. Text region extraction in a document image based on the Delaunay tessellation. *Pattern Recognition*, 36, 3: 799-809.

[22] Chi, Z., Wang, Q., and Siu, W. C. 2003. Hierarchical content classification and script determination for automatic document image processing. *Pattern Recognition*, 36, 11: 2483-2500.

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

[24] Wang, Y., Phillips, I. T., and Haralick, R. M. 2006. Document zone content classification and its performance evaluation. *Pattern Recognition*, 39: 57-73.