Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000191

OVER-SPLITTED AND MERGED FOR GEOMETRY DOCUMENT LAYOUT ANALYSIS

Ha Dai Ton[1], Nguyen Duc Dung[2], Le Duc Hieu[2]
[1]Ha Long Gifted High School, Address,
[2]Institute of Information Technology, Vietnamese Academy of Science and Technology
hadaiton83@gmail.com, nddung@ioit.ac.vn, ldhieu@ioit.ac.vn

Abstract *- Automatic transformation of paper documents into electronic forms requires geometry document layout analysis at the first stage. However, variations in character font sizes, text-line spacing, and layout structures have made it difficult to design a general purpose method. The use of some parameters has therefore been unavoidable in geometry document layout analysis algorithms. This lead to errors over-segmentation and under-segmentation of previous algorithms. This paper present a new approach to geometry document layout analysis. Our algorithm use a set of whitespace covering document background to reduce candidate zones. Some of them may be considered as over-segmented. The way bottom-up is used to group over-segmentation zones each other based on adaptive parameters. Finally, we proposed context analysis at textline level to segment document images into paragraphs. Experimental results on the ICDAR2009 competition data set shown that the proposed algorithm reduces vast amount of both over-segmentation and under-segmentation errors, thus boost the performance significantly comparing to the state-of-the-art algorithms*

Keywords - Geometry document layout analysis, whitespaces covering document background, text regions over-segmented, parameter adaptive, performance evaluation

I. INTRODUCTION

Document layout analysis is one of the main components of OCR system, automatic data entry, computer vision... The task of structural analysis includes automatically detecting image zones on a document image (analysis physical structure) and classifying them into different zones such as: text, images, tables, header, footer... (analysis logical structure). The results of page segmentation are used as an input to the process of recognition and automatic data entry of image processing systems in general. Compared with the analysis of the logical structure analysis, the physical structure analysis (page segmentation) has attracted more attention of researchers due to the diverse and complex structures of different types of document. Not only the specific types of document (books, newspapers, magazines, reports...) but also the other factors of a page such as editors and font size, layout, alignment constraints... affect detection and segmentation accuracy of the algorithm. Document layout analysis algorithms are primarily divided based on their order of processing into three approaches: bottom-up, top-down and hybrid.

Bottom-up algorithms are both the oldest Wahl and Casey (1982) [17] and more recently published O'Gorman (1993) [14], Chowdhury and Chanda (2007) [4] algorithms. They classify small parts of the image (pixels, groups of pixels, or connected components), and gather those of the same type together to form regions. The key advantage of bottom-up algorithms is that they can handle arbitrarily shaped regions with ease (rectangular or nonrectangular). But the fact that they are really sensitive to the measure used to form higher-level entities is the key disadvantage; this often leads to error of over-segmentation in the page with many changes in font sizes and styles, especial the titles.

Top-down algorithms, e.g. Breuel (2002) [2], Nagy at el (1992) [11] cut the image recursively in vertical and horizontal directions along whitespaces that are expected to be column boundaries or paragraph boundaries. Although top-down algorithms have the advantages that they have low computation complexity and good separation result on images with rectangular layout, they are not really able to handle the variety of formats that occur in many magazine pages, such as non-rectangular regions and cross-column headings that blend seamlessly into the columns below.

A third type of algorithm, such as Smith (2009) [25], is based on bottom-up method to find delimiters such as whitespace, tabstops... This reduce top-down structure. And then, using a combination of bottom-up and top-down methods to detect text regions. So, hybrid algorithms can overcome over-fragmentation error of bottom-up algorithms as well as perform better with non-rectangular regions.

The proposed hybrid algorithms havily depent on delimiter between columns. In general, used delimiters are lines or rectangulars that connect alignment connected components at marginal each other. Therefore, they are limited in case of columns that are not aligned marginal or very close each other. The use of some parameters has been unavoidable in any page segmentation algorithms, such as the used parameters are conditions grouping of bottom-up algorithms, the parameters are condition stopping of top-down algorithms. These parameters are very sensitive to page segmentation results [16]. This lead to a fragile frontier between over-segmentation error and under-segmentation error. While, we try fix over-segmentation error then it is very easy to lead under-segmentation error.

Our approach is to overcome the under-segmentation error, agreeing reluctantly with over-fragmentation which can be controlled. Given a document image, the proposed algorithm (called OverAM) first perform over-splitted zone

hypotheses in a top-down way based on a set of whitespace covering document background. Then, the over-splitted zones are gathered each other based on adaptive parameters.

This paper is organized as follows. In section II we describes in detailed the OverAM algorithm. Section III gives experimental results and analysis on the ICDAR2009 data set. Finally, conclusion and discussion are given in section IV.

II. PAGE SEGMENTATION VIA OVER-SPLITTED AND MERGE

In this section, we present the OverAM algorithm in detail. Our algorithm are divided into main phases. In the first phase, a set of whitespace covering document image background is used to reduce hypotheses zones which may be over-segmented. In the second one, over-segmentation zones are grouped as well as paragraphs are detected based on context analysis at textline level. Figure 1 outlines the main of our document layout analysis algorithm.

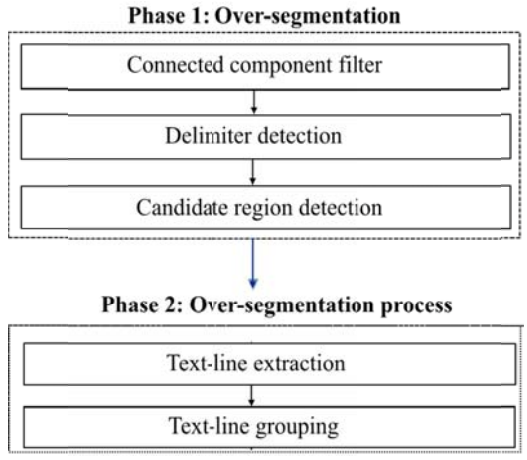


Fig. 1. Our document layout analysis processing chain

A. Phase 1

a) Filtering connected components. The aim of this step is to identify vertical lines, horizontal lines, image regions, and separate the remaining connected components into likely text components and a small number of uncertain types. The connected components (CCs) are filtered by width w, and height h into small, medium, and large size as follow: CCs with $h < 5$ (at 300 dpi) are small, CCs with $w \leq 10$; $h \geq 200$ or $h \leq 10$; $w \geq 200$ are vertical lines, and horizontal lines. The average of height and weight of remainder are computed and denoted by h_a ; w_a. Then, CCs with $w < 0{:}25 * w_a$ or $h < 0{:}25 * h_a$ are small, CCs that satisfy the following conditions are large (likely as non-text) $w/h > 10$; $h/w > 10$; $w > 10 * w_a$ or $h > 10 * h_a$, and the rest are medium (likely as text).

b) Separator detection. When a document is written and laid out by a word processor or a professional publishing system, text regions are usually bounded and thus differentiated from each other by the mean of delimiters. In recently years, the delimiters that used universaly can either be long horizontal/vertical line segments (dubbed solid separator), physical separators (distance of CCs), large elongated empty areas (dubbed whitespaces) or the chain of alignment connected components (tab-stops). In RAST algorithm proposed by Breuel (2002) [3] and The Fraunhofer system which wined within 2009 Page Segmentation Competition [19] uses Whitespace algorithm [2] to find maximal rectangle whitespaces and then retain whitespaces that satisfies certain conditions, e.g. their height must be large enough in relation to the dominant character size or number of connected components on each of sides of whitespaces is large enough. Tab-Stop [25] proposes a bottom-up method based on connected component grouping to detect the tab stops at the margin within a document image and uses the tab stops to deduce its column layout. In general, separators between the columns were used by the state of the art algorithms. They are detected base on the alignment of connected components at the marginal of columns. Thereby, these aprroach are limitted in cases that columns are very close each other or not aligned. In order to overcome this limitation, we use separators to be a set of whitespaces that covering document background. So, they not only allow us solve fully the problems of separator detection but also allow us to determine hypotheses zones. The geometry algorithm of Breuel (2002) [2] is consedered as one of efficient solutions for the problem of full covering the document background. However, the Whitespace algorithm [2] fixed number of rectangular whitespaces to be 300. This is not in favor of our algorithm, when we perform on the document image that has too little connected components. In term of experiments, we found that number of rectangular whitespace that support more better for our algorithm is min(300, nCCs), where nCCs stand for number of CCs.

c) Candidate regions detection.. The first, we initialize a two-dimensional binary array, the size of array are weight and height of the document image. The elements of the array within whitespace are assigned by ”0”, the rest are assigned by ”1”. Since then, distinct areas of the array with value ”1” will be discrimination regions on the image. Determing the bounding box of a homogeneous image regions (Polygon Bounds), we use a recursive algorithm as

follows: the algorithm starts with finding a first element ”1” of the array. After that, the process of boundary detection is began from the element ”1” according to a clockwise direction until you reach the starting point. For more facility, we only save boundary points that liked as coner point. A boundary point ”B” is liked as a coner point, if move from point ”A” to ”B” and move from point ”B” to ”C” then movement direction is changed (A, B and C are not alignment). All of coner points make a simple polygon that is bounding box of each image region. The process of finding the bounding box of a next image region is repeated with the same starting point. But this starting point does not belong to image regions that detected previously. The result of processing of bounding box detection is illustrated at Figue 2.

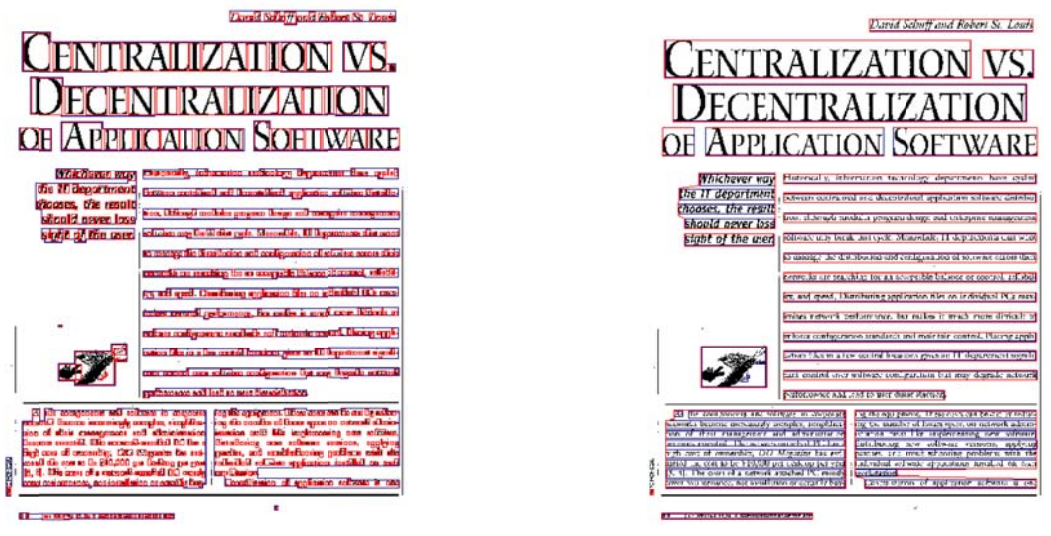
a) b) c)

Fig. 2. a), d) The red rectangular show maximal rectangular whitespaces. b), e) illustrate binary array with blue are asigned ”1”. c), f) The result of finding bounding box.

B. Phase 2

As discussions in phase 1, for successful in dealing with the under-segmentaton error, we admit for the oversegmentation error. However, our algortihm can control the over-segmentation error by the following reason: this error only appears on text regions that distance between CCs within it is larger than other, such as text regions in title, and therefore number of text-lines within over-splitted regions is less. So, the bottom-up way is performed to detect text-lines.

a) Text-line extration. For each image region detected at phase 1, we use RayCasting algorithm to check CCs is in bounding box of the region or not? This way, we obtained distinct regions with a set of connected components within them. Scanning the CCs from left to right and top to bottom, runs of similarly classified CCs are gathered into text-lines, subject to the constraint that no textlines may cross a whitespace with width is enough large, see Figure 3.



a) b)

Fig. 3. a), c) Illustrate CCs belong bounding box of each region image. b), d) Show the result of text-line detection.

b) Text-line over-segmented grouping. A zone is liked as over-segmented if number of text-line within it is less enoungh, see Figure 5a, and text-lines within these regions are dubbed over-textlines. Afterward, using bottomup way to group these text-lines into zones, see Figure 5b. A pair of text-lines of over-textlines satisfies simultaneously the following conditions then they will be grouped into the same region:

$$\begin{cases} (i) & Dist\_Horiz(dong\_i, dong\_j) \le 0.75 * \min(h_i, h_j) \\ (ii) & Dist\_\mathrm{Vert}(dong\_i, dong\_j) \le P_{ij} \end{cases}$$

In the above conditions, Dist_Horiz(., .) and Dist_Vert(., .) are the horizontal and vertical distance of text-lines, h_ is height of line_ . The parameter Per_{ij} is difined as follows, see Fig 4: Set,

$$ratio_{ij} = \frac{|h_i - h_j|}{\min(h_i, h_j)},$$

$$Per_{ij} = \begin{cases} 1.3\min(h_i, h_j) & if \quad ratio \le 0.3 \\ 0.3\min(h_i, h_j) & if \quad ratio \le 0.3 \end{cases}.$$

Fig. 4. The distance parameters per_{ij} vary between pairs of text-lines

a) b)

Fig. 5. a) Show over-textlines, b) the result of grouping over-textlines into text regions, the red polygons are results of over-segmentation processing

III. EXPERIMENT

In this section, we present empirical performance evaluation results of our algorithm compared state-of-the-art algorithms: Docstrum [14], RAST [2, 3], Tab-Stop [25], Whitespace [2], XY Cut [11] and Voronoi [6]. We use toolkit PSET [10] to compare performance of the algorithms on ICDAR2009 competition dataset [20].

A. Data set

We selected the ICDAR2009 competition data set (ICDAR2009 dataset) for the performance evaluation task since they are the datasets used the most popular in document layout analysis. They are the datasets currently available that has textline and paragraphs level ground-truth for each document image. The text-line and paragraphs level ground-truth are represented by non-overlapping polygons. The ICDAR2009 dataset is a subset of PRImA dataset

Antonacopoulos and Pletschacher (2009) and has 305 document image at 300 dpi resolution. It contains a wide variety of different document types, reflecting the various challenges in layout analysis. The layouts of these pages contain a mixture of simple and complex layouts, including many instances of text wrapping tightly around images, varying font sizes and other characteristics which are useful to evaluate layout analysis methods on.

B. Performance metric

Let G be the set of all the ground-truth text-lines in a document image, and |G| denote the cardinality of the set G. Then, three subsets of text-lines are defined as follow, see Fig 6:

a) The set of ground-truth text-lines that are missed (C), i.e. they are not part of any detected text region.

b) The set of ground-truth text-lines whose bounding boxes are split (S), i.e. bounding box of a text-line is not completely within one detected segment.

c) The set of ground-truth text-lines that are horizontally merged (M), i.e. two horizontally overlapping ground-truth

text-lines are part of one detected segment. The possible tex-line errors are showed in Figure 6. Then, the overall performance rate is measured as the percentage of ground-truth text-lines that are identified correctly

$$\rho = \frac{|G| - |C \cup S \cup M|}{|G|}$$

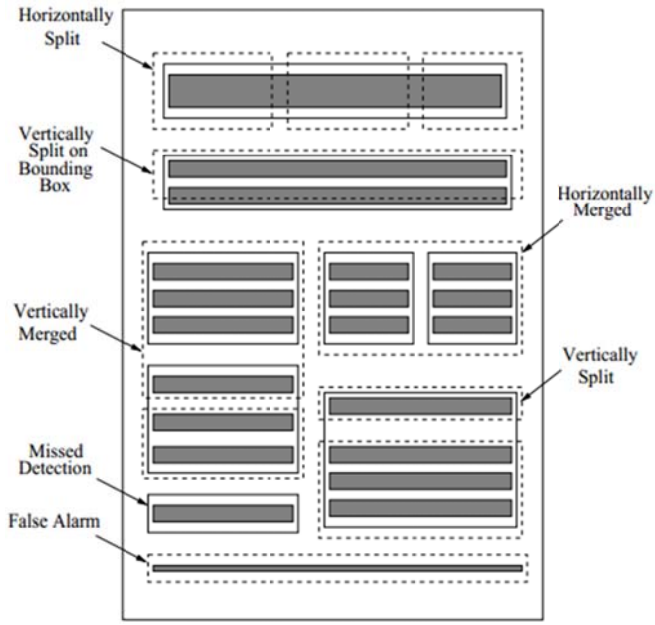


Fig. 6. This figure shows a set of possible text-line errors. Solid-line rectangles denote ground-truth zones, dashed-line rectangles denote OCR segmentation zones, dark bars within ground-truth zones denote groundtruth text-lines, and dark bars outside solid lines are noise blocks. Mao and Kanungo (2002)

C. Result and discussion

Figure 7 shows different types of errors produced by seven page segmentation algorithms. We can see that the OverAM dramatically reduces the merged-line and splitted-line errors when comparing to both top-down and bottom-up algorithms. The OverAM boosts the text-lines performance up to 89:03:04%, compared to 83:16% and 75:26% produced by TabStop and Docstrum, see Figure 8. The large difference between OverAM performance and those of the remain algorithms is partially caused by the difference in algorithm design. OverAM is designed to segment document images into paragraphs, whereas Tab-Stop, Docstrum and others were designed to separate text blocks. There are two ways to make the performance comparison more reasonable. The first one is to extend the algorithms for achieving paragraph-level results. The second way is to modify the ground truth from paragraph-level to text-block level. For the convenience, we select the second solution to modify the ground-truth of the ICDAR2009 data set as illustrated in Figure 9. The experimental results on this modified ICDAR2009 data set is shown in Figure 10. The OverAM still surpasses other algorithms significantly: 91:38% compared to the top 85:61% achieved by the Tab-Stop algorithm.

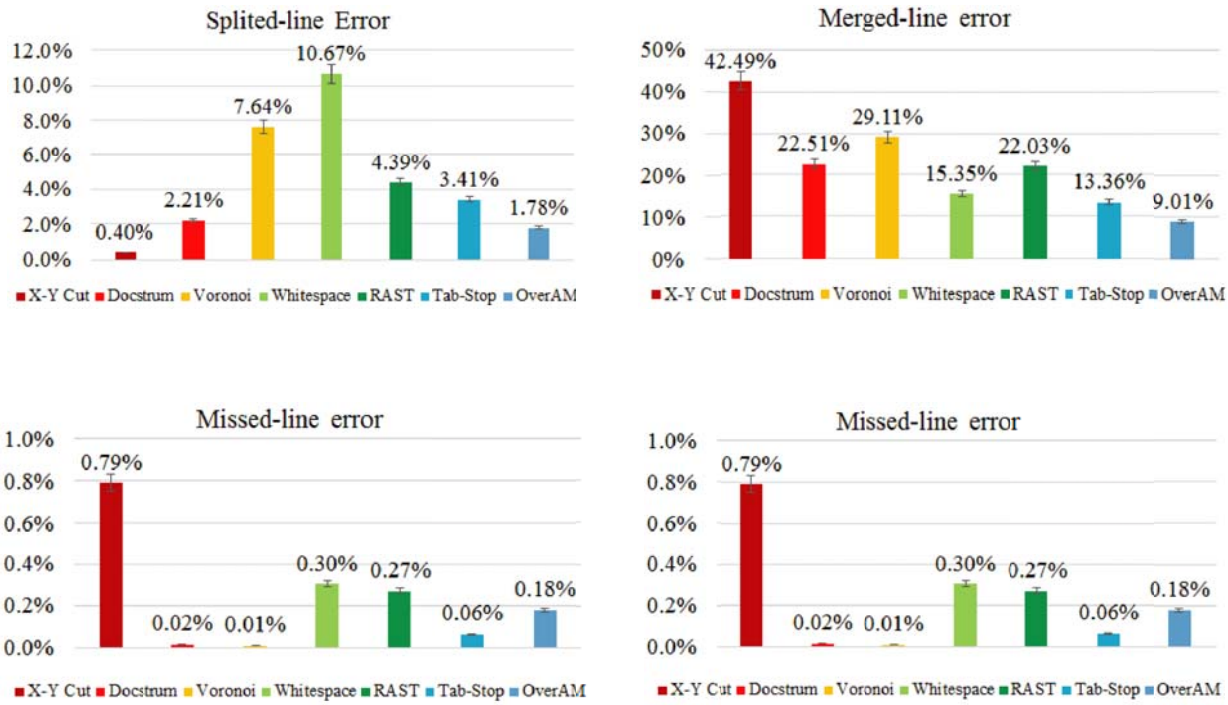


Fig. 7. Errors of different algorithms on the ICDAR2009 dataset.

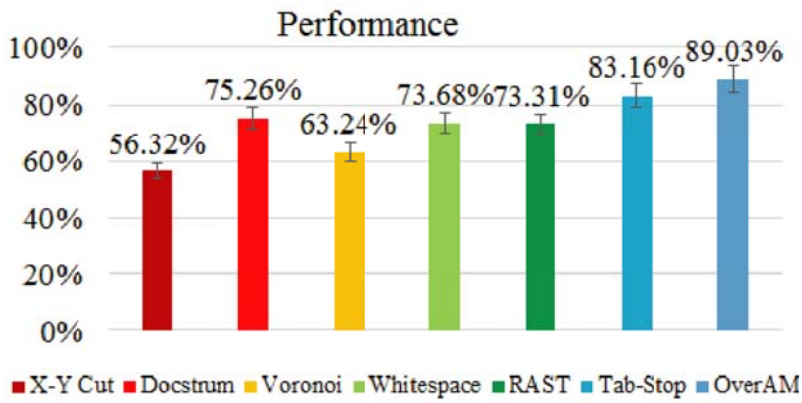


Fig. 8. The experiment result on the ICDAR2009 dataset

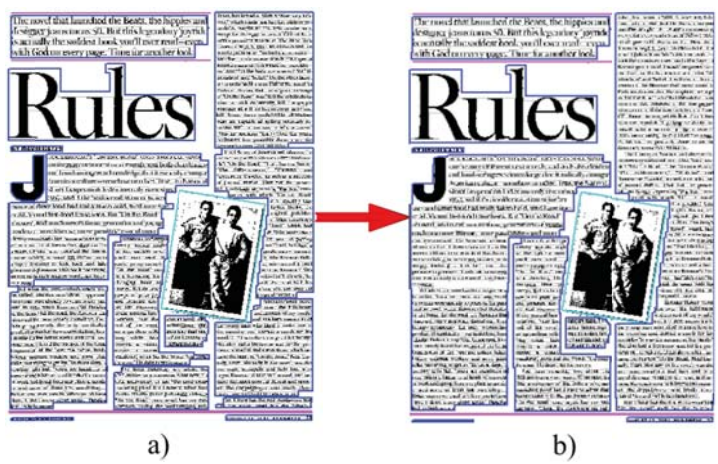


Fig. 9. a) Ground-truth of ICDAR2009, b) ground-truth of ICDAR2009 modified.

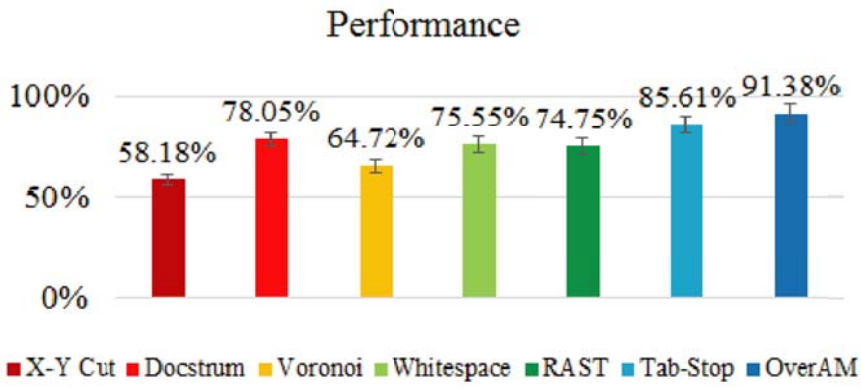


Fig. 10. The experiment result on the ICDAR2009 dataset with modified ground-truth.

IV. CONCLUSION

We have presented an over-splitted and merge approach for page segmentation that segments the document image at level paragraphs. Compared with the sate of the art algorithms relevant to OverAM, the major contribution of proposed algorithm ar summarized as follows: 1) independence on any fixed parameters to segment the document image having various character font size, connected component spacing, and document layout structure, 2) effective detection of paragraphs by local information analysis, and thereby proposed algorithm can handle difficult cases where the purely top-down and bottom-up approaches are not sufficient to separate, 3) so some texture analysis on the ambiguous regions isolated automatically allow us to remove easily small non-texture, 4) the proposed algorithm is easy to implement, because it is inherited from the powerful tools such as: tool of detect whitespace covering the document background, edge detection algorithm, the finding points in the polygon algorithm. The outcomes of experimental show that OverAM has achieved high performance on ICDAR2009 competition data set regardless of changes in character font sizes and complex document layout structure

V. REFERENCE

[1] Antonacopoulos, A., Karatzas, D. and Bridson, D.: Ground truth for layout analysis performance evaluation. in Document Analysis Systems, Nelson, New Zealand, pp. 302–311, Feb. 2006.

[2] Breuel, T. M.: Two geometric algorithms for layout analysis. In Document Analysis Systems, Princeton, NY, pp. 188–199, Aug 2002.

[3] Breuel, T. M.: High performance document layout analysis. In Symposium on Document Image Understanding Technology, Greenbelt, MD, April 2003.

[4] Das, A. K., Saha, S. K. and Chanda, B.: An empirical measure of the performance of a document image segmentation algorithm. International Journal on Document Analysis and Recognition, vol. 4, no. 3, pp. 183–190, 2002.

[5] Jagadish, H. V., O'Gorman, L.: "An object model for image re-cognition," IEEE Comput., vol. 22, no. 12, pp. 3341, Dec 1989.

[6] Kise, K., Sato, A. and Iwata, M.: Segmentation of page images using the area Voronoi diagram. Computer Vision and Image Understanding, vol. 70, no. 3, pp. 370–382, June 1998.

[7] Liang, J., Phillips, I. T. and Haralick, R. M.: Performance evaluation of document structure extraction algorithms. Computer Vision and Image Understanding, vol. 84, pp. 144–159, 2001.

[8] Lotti, F., Heroux, P., Adam, S., Sanchez, G., Valveny, E., Dosch, P. and Llados, J.: Performance analysis and evaluation working group report. In Document Analysis Systems, Florence, Italy, Sep. 2004.

[9] Mao, S. and Kanungo, T.: Empirical performance evaluation methodology and its application to page segmentation algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 242-256, March 2001.

[10] Mao, S. and Kanungo, T.: Software architecture of PSET: a page segmentation evaluation toolkit. International Journal on Document Analysis and Recognition, vol. 4, no. 3, pp. 205–217, July 2001.

[11] Nagy, G., Seth, S. and Viswanathan M.: A prototype document image analysis system for technical journals. Computer, vol. 7, no. 25, pp. 10–22, 1992.

[12] Namboodiri, A. M., Jain, A.K.: Document Structure and Layout Analysis. Digital Document Processing Advances in Pattern Recognition , pp 29-48, 2007.

[13] O'Gorman, L.: "Primitives chain code", in Progress in Computer Vision and Image Processing (A. Rosenfeld and L. G. Shapiro, Eds.). San Diego: Academic, pp. 167-183, 1992.

[14] O'Gorman, L.: The document spectrum for page layout analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1162–1173, Nov. 1993.

[15] Rege, P. P and Chandrakar, C. A.: Text-Image Separation in Document Images Using Boundary/Perimeter Detection, ACEEE Int. J. on Signal & Image Processing, Vol. 03, No. 01, Jan 2012.

[16] Shafait, F., Keysers, D., Breuel, T.: Performance evalution and benchmarking of six page segmentation algorithms. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol 30, pages 941-954, June 2008.

[17] Wong, K. Y., Casey, R. G. and Wahl, F.M.: Document analysis system. IBM Journal of Research and Development, vol. 26, no. 6, pp. 647–656, 1982.

[18] Antonacopoulos, A., Gatos, B. and Bridson, D.: ICDAR 2007: Page segmentation competition. In Proc. 9th Intl. Conf on Document Analysis and Recognition, Curitiba, Paraná, Brazil, September 2007.

[19] Antonacopoulos, A., Pletschacher, S., Bridson, D. and Papadopoulos, C.: ICDAR 2009: Page segmentation competition. in Proc. 10th Intl. Conf. on Document Analysis and Recognition, University of Salford, Manchester, United Kingdom, July 2009.

[20] Antonacopoulos, A., Bridson, D., Papadopoulos, C. and Pletschacher, S.: A Realistic Dataset for Performance Evaluation of Document Layout Analysis. in Proc. 10th Intl. Conf. on Document Analysis and Recognition, University of Salford, Manchester, United Kingdom, July 2009.

[21] Ferilli, S., Biba, M., Esposito, F., Basile, T. M. A.: A Distance-based Technique for non-Manhattan Layout Analysis, In Proceedings of the 10th International Conference on Document Analysis and Recognition, volume I, pages 231–235. IEEE Computer Society, 2009.

[22] O'Gorman, L.: "Image and document processing techniques for the RightPages electronic library system," in Proc. 11 th In(. Conf Patt. Recogn. (ICPR) (The Hague, The Netherlands), pp. 26C263, Aug. 1992.

[23] Shafail, F., Keysers, D. and Breuel, T. M.: Performance comparison of six algorithms for page segmentation. in 7th IAPR Workshop on Document Analysis Systems, Nelson, New Zealand, pp. 368–379, Feb. 2006.

[24] Sung, H. M.: Enhanced Constrained Run-Length Algorithm for Complex Layout Document Processing. International Journal of Applied Science and Engineering, pp. 297-309, 2006.

[25] Smith, R.: Hybrid page layout analysis via tab-stop detection. In Proc. Int. Conf. on Document Analysis and Recognition, pages 241- 245, Barcelona, Spain, July 2009.

[26] Yanikoglu, B. A. and Vincent, L.: Ground-truthing and benchmarking document page segmentation. In Proc. 3rd Intl. Conf. on Document Analysis and Recognition, Montreal, Canada, pp. 601–604, Aug. 1995.

HƯỚNG TIẾP CẬN TÁCH QUÁ VÀ GỘP LẠI CHO BÀI TOÁN PHÂN TÍCH CẤU TRÚC TRANG ẢNH TÀI LIỆU

Hà Đại Tôn, Nguyễn Đức Dũng, Lê Đức Hiếu

Tóm tắt - Phân tích cấu trúc hình học là bước đầu tiên của quá trình chuyển đổi tự động các dữ liệu dạng trang ảnh tài liệu thành dạng file điện tử. Tuy nhiên, sự thay đổi đa dạng của các kiểu font chữ, khoảng cách giữa các dòng chữ, cấu trúc của các trang ảnh đã tạo ra những khó khăn nhất định trong việc xây dựng một thuật toán tổng quát. Bất kì một thuật toán phân đoạn trang nào cũng phải sử dụng đến các tham số (hoặc cố định hoặc tự do), điều này đã dẫn đến các lỗi over-segmentation hoặc under-segmentation. Trong bài báo này, chúng tôi trình bày một hướng tiếp cận mới cho bài toán phân đoạn trang ảnh tài liệu. Thuật toán của chúng tôi dùng một tập các khoảng trắng bao phủ nền của trang ảnh và từ đó suy ra được các vùng ảnh ứng cử viên. Một trong các vùng ứng cử viên có thể có vùng bị phân tách quá nhỏ (over-segmented). Sau đó, cách tiếp cận bottom-up được áp dụng để nhóm các vùng ảnh ứng cử viên này lại với nhau dựa trên các tham số thích nghi. Các kết quả thực nghiệm trên tập dữ liệu của cuộc thi ICDAR2009 đã chỉ ra rằng thuật toán của chúng tôi giảm đáng kể các lỗi over-segmentation và under-segmentation, do đó tăng độ chính xác của thuật toán lên đáng kể so với các thuật toán hàng đầu.

Từ khóa – Phân tích cấu trúc trang ảnh tài liệu, các khoảng trắng bao phủ nền của trang ảnh, các vùng ảnh bị phân tách quá nhỏ, tham số thích nghi, đánh giá độ chính xác.

Lời cảm ơn - Chúng tôi xin chân thành cám ơn những ý kiến đóng góp vô cùng quý giá của các phản biện. Bài báo cũng nhận được sự hỗ trợ từ đề tài "Nghiên cứu phát triển các phương pháp phân tích cấu trúc và nhận dạng văn bản trong bài toán nhập dữ liệu tự động" của Viện Hàn lâm Khoa học và Công nghệ Việt Nam, mã số VAST01.08/15-16.