

Fast Document Segmentation Using Contour and X-Y Cut Technique

Boontee Kruatrachue, Narongchai Moongfangklang, Kritawan Siriboon

Department of Computer Engineering, Faculty of Engineering,
King Mongkut's Institute of Technology Ladkrabang Bangkok, Thailand 10520.
kkboonte@kmitl.ac.th, nanfa_blue@yahoo.com, kritawan@diamond.ce.kmitl.ac.th

ABSTRACT

This paper describes fast and efficient method for page segmentation of document containing nonrectangular block. The segmentation is based on edge following algorithm using small window of 16 by 32 pixels. This segmentation is very fast since only border pixels of paragraph are used without scanning the whole page. Still, the segmentation may contain error if the space between them is smaller than the window used in edge following. Consequently, this paper reduce this error by first identify the missed segmentation point using direction information in edge following then, using X-Y cut at the missed segmentation point to separate the connected columns. The advantage of the proposed method is the fast identification of missed segmentation point. This methodology is faster with fewer overheads than other algorithms that need to access much more pixel of a document.

Keywords: Contour Direction Technique, Missed Segmentation Points, Page Segmentation, Recursive X-Y Cut Technique

1. INTRODUCTION

Page segmentation [1] is the process to identify the areas of interest in the image of a document page. For a conventional document page with material printed in dark ink on a light colored paper, the areas of interest in the (binary) image will be neighborhoods of black pixels. Page segmentation produces a description of the geometrical aspects of the areas of interest. The most common aspects are spatial extent and position on the page. Page segmentation can be thought of as a mapping from the pixel-based image data to a description of the areas of interest.

Several methods have been explored to solve the problem of automated document segmentation. There are three main methods for automatically document segmentation, top-down method [2], [4], [5], bottom-up method [6], [1] and mixed method [3]. A successful method must cope with as many variations (e.g., shapes of regions, skew) in the document as possible. In doing so, it must also not compromise in processing speed. This paper tries to improve the segmentation speed by using contour following technique. This technique is very fast since only border pixels information is used in segmentation. However, the segmentation correctness depends on the window size used in edge following a paragraph. If the window is too small it

will extract letters or words but not paragraph (over segment). On the other hand, if the window is too big it will segment multiple structures (columns) together (under segment). This paper proposed a fast connected point identification to handle the under segment problem. Once connected point is identified, X-Y cut at the point will separate the connected blocks.

2. PREVIOUS WORKS

Several methods and their variants have been employed in page segmentation approaches. Some of them include our previous work are briefly reviewed below.

The well known top-down page segmentation technique is the recursive X-Y cut [2]. This approach decomposes a document image by projection profile cuts recursively in to a set of rectangular block. The disadvantage of this method is the rectangular structure of a block.

The hybrid (mixed) approach was presented in [3] using contour following technique. This technique used the square window 32 by 32 pixel followed the rim of paragraph, to specified the boundaries of paragraph in image document. To separate the potential connected blocks, vertical cuts on space between characters are used to separate connected blocks. The advantages of method are the fast computation time, and nonrectangular format block in the contour segmentation. Due to vertical cuts, it can not separate connected columns with nonrectangular format.

3. FAST DOCUMENTATION USING CONTOUR AND X-Y CUT TECHNIQUE

Even though many researches have been studied the automated document segmentation, there is no appropriate and complete method for every document. Our approach focus on the columns with narrow gab between them .

The proposed page segmentation consists of three main processes as shown in Fig.1. Firstly; we used the modified contour following technique in clockwise direction [3]. This technique is very fast since only border pixels information is used in segmentation. Then, we identify the missed segmentation point by looking at the turning point of window between columns. The missed segmentation point is the point where the direction of window is counter clockwise.

Finally, we apply X-Y Cut Technique at the missed segment points to separate the connected columns.

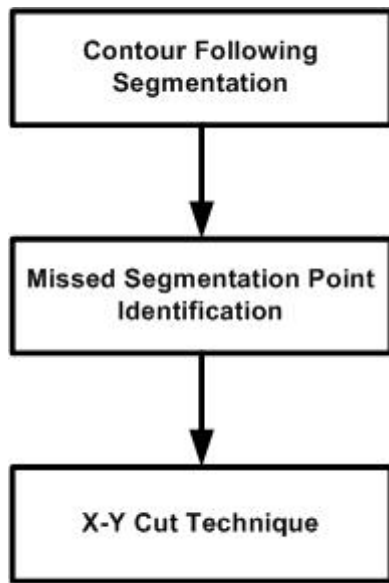


Fig.1: the overview of document segmentation

3.1 Contour Following Segmentation

The first stage page segmentation is a top-down approach using modified contour following algorithm. The algorithm starts with the searching of the first black pixel. It was used as the starting point of the 16 by 32 window. Then normal edge following algorithm was performed, where the window was considered as located on the black pixels if 10 black pixels were in the window. In spite of its fast execution time, the segmented block may contain multiple columns or paragraphs. In order to reduce the multiple connected columns, the window use in contour following is rectangular of 16 by 32 pixel instead of square 32 by 32 pixel as in our previous paper [3]. The main advantage of contour following segmentation is the shape of block can be in any form (non rectangular) and its execution speed.

3.2 Missed segmentation point identification

In order to make use of the fast segmentation in the previous stage, we must be able to also identify the connected column fast. Since the block segmentation is performed by contour following in the clockwise direction. The single column will contain only windows in clockwise direction only. Hence, the potential missed segmentation point is at the window that has counter clockwise direction.

Some window that has counter clockwise direction may not be the connected point as shown below by the dash circled in Fig2. In order to identify the real connected point, the height or depth of the connected point has to be greater than some threshold.

The connection between top and bottom (row) paragraph is allowed. Since a paragraph may not has the justified format (center, align right , align left), it is hard to separate the connected top and bottom paragraph.

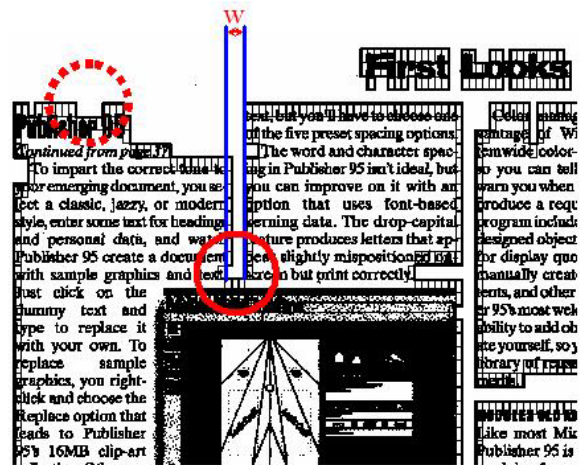


Fig. 2: show the candidate (dash circle) and real (circle) missed segmentation point.

3.3 X-Y Cut Technique

Once we find the real missed segmentation point, the next step is to separate the connected structures. There are two types of separation, horizontally and vertically. If the space (w) between structures at the connected point is wider than the window width (16 - pixels), the separation starts horizontally, as shown in Fig. 2. (by bold line). On the other hands, if the width is smaller or equal to the width, the separation starts vertically.

Vertical separation can be done by projection cut vertically direction in top or bottom until facing the border window of structure. If the projection hits black pixel before border windows, the projection stops and turns horizontally to both left and right directions. This projection continues until it hits the black or border of window. If it hits black pixel, it changes to perpendicular direction. If it hits the border window, the projections stop. Horizontals separation also performs in the same manner.

4. EXPERIMENT AND RESULT

Hundred of documents have been used in segmentation test. Some of the difficult segmentation samples are shown in Fig. 3. The boundaries result of Fig. 3 show in Fig 4. Since the connected column separation use X-Y cut, the algorithm still has problem with non rectangular shaped of connected structures, as shown in Fig 5.

5. CONCLUSION AND RESULT

Our paper discusses an improved version of the contour following technique using rectangular window, fast missed segmentation identification and correct separation of connected block. This technique is very fast since only border pixels are accessed. We improved the speed of this technique with fast automatically identification of the connected point using window directions. Then apply an X-Y cut

technique to split the connected structure at the connected point. If the connected structures are nonrectangular shape, they still can not be separated due to the X-Y cut technique.

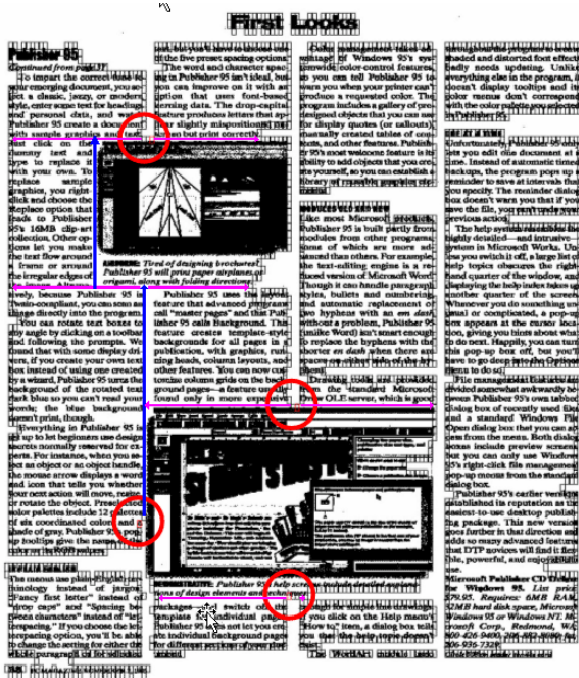


Fig. 3: Missed segmentations point and trace of projection X-Y cut technique

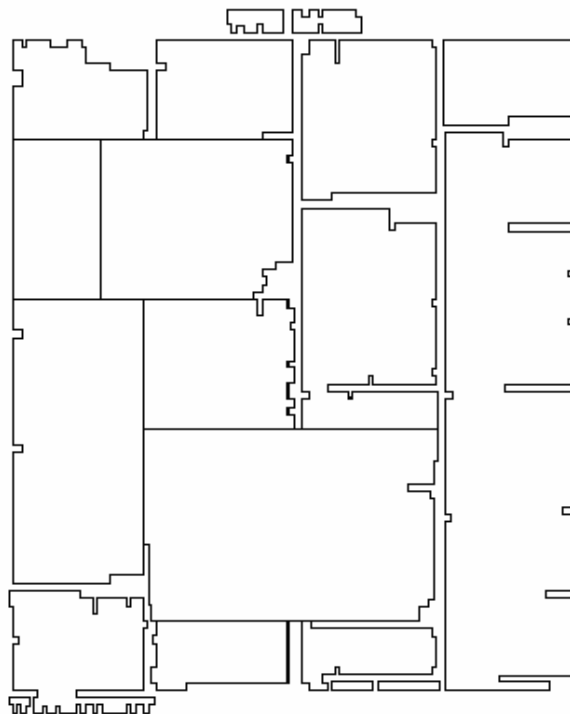


Fig. 4: Show the result region of structure of the original image in Fig 3, a space between each column is very small due to noise and layout of document. In this figure, the contour direction only is not enough to correct in the part of column



Capital Equipment

Tariffs, standards and IP theft—oh my! Business in China is proving problematic

Tariffs and chip process, test and assembly equipment has been going for years. But it was only in the last few, as the industry struggled with the downturn, that China became not only an important market but also increasingly a competitor. Both the industry and the U.S. government have ardently courted China, working to get the Communist giant into the World Trade Organization (WTO), the industry covers not only China's growing market but also its rapidly rising manufacturing and labor force.

But as an old Chinese proverb states, there said that it would honor the U.S. request; the two sides have 60 days from out the issue before the WTO gets involved.

Another headache for foreign chip makers is the implementation of proper IP standards, ones that naturally favor domestic manufacturers. Although the Communist nation may be making overtures to free market entry, old habits apparently die hard. The biggest example of this is China's favoring Authentication and Privacy Infrastructure (WAPI) standard. The industry announced in November of last year that it was adopting the national wireless LAN standard, and not only is it incompatible with the existing 802.11 wireless standard, but China is also prohibiting the export of leading-edge process tools and their European and Japanese counterparts.

But the entire industry is protesting at the idea. The Semiconductor Industry Association has called on China to drop the WAPI standard, and no less giant than Intel essentially washed its hands of the matter in March. Intel, which has made a big push over the past couple of years into mobile chip technology, most notably with its Centrino chip set, said in early March that it wouldn't be able to produce chips that achieve its own standards of quality and meet China's June 1 deadline for implementing the WAPI standard. More notably and perhaps more telling, Intel said that it had no scheduling problems with licensing such chips.

Even while the WAPI standard is being the stakes may be higher in capital equipment.

Fig.5: Show the error segmentation result from missed segmentations of the skew object inside paragraph, missed segmentation points are circled.

6. REFERENCES

- [1] A. Antonopoulos, "Page Segmentation using the Description of the Background Computer Vision and Image Understanding, Vol. 70, (1998) 350-369.
- [2] G. Nagy and S. Seth, "Hierarchical representation of optically scanned documents," *Proc. of ICPR*, (1984) 347-349.
- [3] B. Kruatrachue, P. Suthaphan, "A fast and efficient method for document segmentation for OCR", *Electrical and Electronic Technology*, 2001. *Proceeding of IEEE Region 10 International conference on*, Volume: 1, 19-22 Aug. (2001) 381-383 vol.1
- [4] Jaekyu Ha, R.M. Haralick, I.T. Phillips, "Recursive X-Y Cut using Bounding Boxes of connected components", *Proceedings of the Third International Conference on Document Analysis and Recognition*, Volume:2, 14-15 Aug. (1995) 952-954.
- [5] Jaekyu Ha, R.M. Haralick, I.T. Phillips, "Document Page Decomposition by the Bounding-Box Projection Technique", *Proceedings of the Third International Conference on Document Analysis and Recognition*, Volume:2, 14-15 Aug. (1995) 1119-1122.
- [6] T. Saitoh, T. Pavlidis, "Page Segmentation without Rectangle Assumption", *Pattern Recognition Methodology and Systems, Proceedings, 11th IAPR International Conference on*, 30 Aug.-3 Sept. (1992) 277-280.