Document Layout Analysis: A Maximum Homogeneous Region Approach

Tuan Anh Tran * †, Khuong Nguyen-An † , Nhat Quang Vo ‡
* School of Business Information Technology, University of Economics Ho Chi Minh City,
59C Nguyen Dinh Chieu, District 3, Ho Chi Minh City, Viet Nam.
† Faculty of Computer Science & Engineering, Ho Chi Minh City-University of Technology (HCMUT),
268 Ly Thuong Kiet, District 10, Ho Chi Minh City, Viet Nam.
‡ Department of Computer Science, Chonnam National University
77 Yongbong-ro, 500-757 South Korea.
trantuananh2006@gmail.com, nakhuong@hcmut.edu.vn, vqnhat@gmail.com

***Abstract*—This paper presents a method for document layout analysis. This method applies the analyzing of whitespace in maximum homogeneous regions. This method focuses on the balance between processing time and performance. It consists of two main stages: classification and segmentation. Firstly, by using the analysis of whitespace analysis on Maximum multi-layer horizontal homogeneous regions, the text and non-text elements are classified. Then, text regions are extracted by using mathematical morphology. Besides, non-text elements are classified into separators, tables, images via a machine learning approach. The proposed method's effectiveness is proved by the tests on UW-III (A1) datasets.**

***Index Terms*—Document layout analysis, whitespace analysis, homogeneous region, page segmentation, OCR.**

I. Introduction

Document layout analysis or page segmentation is the process of identifying and categorizing the regions of interest in the scanned document image. The quality of this process can determine the quality and the feasibility of document image understanding. To deal with this problem, in recent years, many approaches have been proposed in various ways but they are generally divided into three main categories: bottom-up, top-down and hybrid method.

Top-down methods [8]–[11] separate the original document into different regions and then use many heuristic filters to classify each region [18], [20]. These methods only effective when the document has Manhattan layout [4].

In another way, bottom-up methods [3], [5]–[7], [12] start with local information such as words, text lines; and then, merge them into blocks or paragraphs. Bottom-up methods are usually widely applicable to various layouts, especially with the non-Manhattan layout. However, these methods are often required a lot of memory space and time consuming.

In a different approach, people began to pay more attention to the combination of two basic approaches to create hybrid method [19], [21], [27], [4], [29]. This method overcomes some weaknesses of the two classic above methods. However, the results of these methods are still not convincing, such as the non-text identification. Whereas, the method in [29] still requires a large computer memory and processing time.

On the other hand, there are some methods that use the Wavelet Transform and Multi-scale resolution [14], [15] to identify non-text elements. This approach is general can deal with the skew document image. However, by using the multi-scale resolution, the computing time is quite long, and it always creates noise when the resolution of document is changed.

In general, most of these methods are not focusing on classifying text and non-text elements before grouping them although the classification of text and non-text elements of a document plays an important role in the document layout analysis. This leads to the result obtained without high precision.

To deal with this problem, in this paper we propose an efficient hybrid method for page segmentation that includes two main parts: classification, and segmentation. In the first stage, the binary input document is cut vertically to get the vertical homogeneous region. On each vertical region obtained, text and non-text elements are classified by connected components analysis and whitespace analysis on multi-layer horizontal homogeneous regions. The output of this process is the maximum homogeneous text regions and vertical non-text regions. This step not only provides an efficient method to determining non-text elements but also create favorable conditions for segmentation by creating maximum homogeneous regions. Moreover, due to the use of maximum homogeneous region instead of minimum homogeneous region, the processing time of this process is reduced significantly. In the segmentation stage, text regions are extracted by the whitespace analysis and mathematical morphology. Non-text elements are classified into separators, tables, images via a machine learning process.

Like most of page segmentation methods; our method is sensitive with skew document, it may be run well with the documents have the skew less than 5 degrees. Therefore, before implement our method, the skew estimation of document [26] is applied as an optional step to correct the skew angle of document.

An overview of the method and its performance is given next. The proposed method is presented in Section II. The performance and evaluation could be found in Section III. Finally, the paper is concluded in Sections IV.

978-1-5386-4180-4/18/$31.00 ©2018 IEEE

II. Proposed Method

Given the binary document, the proposed method is described as follows. Firstly we cut the input document vertically to get the vertical homogeneous regions (abbreviated to *VR*). Then, on each *VR*, the connected components analysis is performed to get the information of all elements in this region. This process also consider an effective filter to identify and eliminate some strictly non-text elements to get the new vertical homogeneous region.

Actually, *VR* still exists many non-text elements. Therefore, the multi-layer whitespace analysis is performed to identify remain non-text elements. In this step, the horizontal segmentation is performed to extract the horizontal homogeneous regions (abbreviated to *HR*) of *VR*. Then, on each *HR* obtained the whitespaces analysis step is implemented to identify the non-text elements by using statistical method. Similarly, all non-text components are removed from *HR* to get the new HR^*. These regions are reshaped to get the new VR^*.

Generally, the multi-layer whitespace analysis is an iteration method, this process is repeated until VR^* satisfies the convergence condition, see Fig. 1. At this time, VR^* includes only text elements and all HR^* are the maximum homogeneous regions. Then, all VR^* is combined vertically to extract the text document (includes only text elements) and non-text document (includes only non-text elements) is obtained by the xor logical with the binary image. In the next step, text components are group together to get the text regions by extracting text lines and mathematical morphology. Besides, all non-text components are classified in detail to identify the separator, table and image regions.

Finally, all regions are obtained (text regions and non-text regions) on each vertical homogeneous region will be reshaped to get the page layout.

A. Cutting image

Suppose $f(x, y)$ is the binary image with $a \times b$ as the size of it. In this section, we present a method to segment the input document into many homogeneous regions. There are two kinds of homogeneity, horizontal homogeneity and vertical homogeneity. The difference between them is the direction in which we get the projection. For instance, to get the horizontal homogeneous regions, we perform the following steps:

1. Find histogram of horizontal projection.

$$HP = \left\{ p_i \mid p_i = \sum_{y=1}^{b} f(i, y), 1 \leq i \leq a \right\} \quad (1)$$

2. Estimate the large of white lines and black lines.

Convert HP to bi-level value (-1 and 0), $\forall p_i \in HP$, if $p_i > 0$ then $p_i = -1$ else $p_i = 0$. The Run Length Encoding (RLE) is used to find the large of white lines and black lines (the length of -1 and 0 sequence). Let b_i and w_i be the large of i_{th} black line and white line respectively. Put B, W is the set includes all b_i and w_i.

3. Estimate the homogeneity

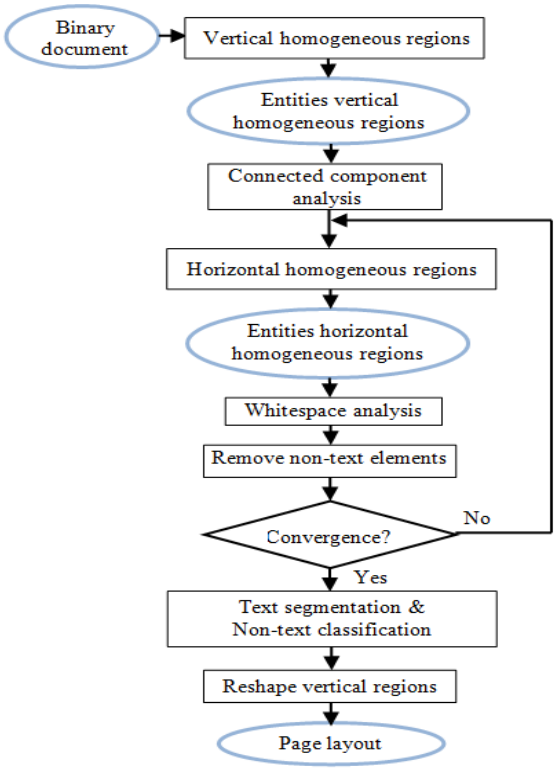


Fig. 1. Block diagram of the proposed method

Let μ_B, μ_W be the mean of black lines in B and white lines in W respectively. The variance of black line and white line is as follow,

$$V_B = \frac{\sum (b_i - \mu_B)^2}{|B|}, \; V_W = \frac{\sum (w_i - \mu_W)^2}{|W|} \quad (2)$$

If the values of V_B and V_W are low, it means the region is homogeneous. Conversely, our region is heterogeneous, and it should be segmented (splitting).

4. Splitting region

The split position is the white line below or above the most distinctive line of black or the largest white line (greater than median). The splitting position is described in [25]. The regions obtained after this process are homogeneous, see Fig. 3b, 3c. Similarly, we can obtain the vertical homogeneous regions by using the vertical projection (1), see Fig. 2b.

B. Connected components analysis

Suppose that we are considering the jth vertical homogeneous region VR_j. Firstly, the connected components are extracted and coordinates are stored. Let CCs be all the connected components in this region, CC_i is the ith connected component and $B(CC_i)$ is the bounding box of it. To improve our performance, a heuristic filter [25] is performed to remove noise and some other tiny connected components. Call CCs'

the set of non-text elements that were found by the above heuristic filter, $CCs = CCs \setminus CCs'$ and $\forall CC_i \in CCs'$

$$VR_j(x,y) = \begin{cases} 0, & \text{if } VR_j(x,y) \in CC_i, \\ VR_j(x,y), & \text{elsewhere} \end{cases} \quad (3)$$

Note that CCs' is stored for the non-text classification process (identify: separator, table, image, etc.).

C. Whitespace analysis on multi-layer horizontal homogeneous region

This is the main process of text and non-text classification. On each VR_j obtained, an algorithm contains three steps (extract the horizontal homogeneous region, whitespace analysis and check the convergence condition) is performed to identify text and non-text elements.

1) Extract the horizontal homogeneous regions: Firstly, on each vertical homogeneous region VR_j (Fig. 3a), again we use the method of cutting image horizontally to get the corresponding horizontal homogeneous regions, HR_k (Fig. 3b). We denote

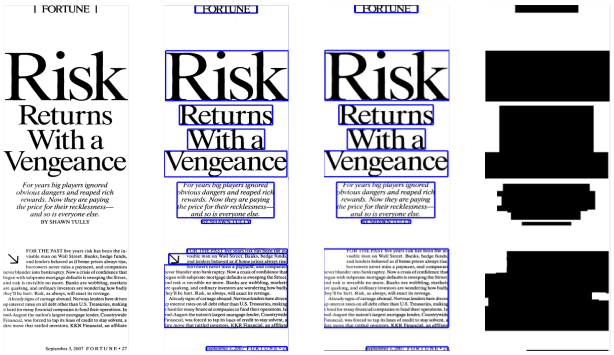
$$VR_j = HR_1 \cup HR_2 \cup \ldots \cup HR_n \quad (4)$$

2) Whitespace analysis: Suppose HR_k is the region being considered, $CCu \subset CCs$ is the set of connected components of HR_k. As we know, text elements in a document language is usually arranged in rows or columns, the difference between these components is always small. Moreover, in a homogeneous region, the text elements often have similar height as well as width. The variance of the neighbor text elements is usually small if the region includes only text elements. Therefore, we can use the recursive filter (based on a statistical approach) which was first proposed Tran et al. [4] to identify the non-text elements in each horizontal homogeneous region. Call CCs' the set of non-text elements that were found by the recursive filter. Similarly, we apply (3) to remove non-text elements and get new horizontal homogeneous regions HR_k^*. Thereby obtaining new vertical homogeneous region VR_j^*.

$$VR_j^* = HR_1^* \cup HR_2^* \cup \ldots \cup HR_n^* \quad (5)$$

(a) (b)

Fig. 2. Example of (a) binary document, (b)vertical homogeneous regions-3 regions in red outlines



(a) (b) (c) (d)

Fig. 3. (a) VR_2 after connected component analysis, (b) horizontal homogeneous regions of VR_2 in blue outlines (c) maximum horizontal homogeneous regions HR_k^* (d) text regions of VR_2

3) The convergence of whitespace analysis: Let S_j, S_j^* be the sum of positive pixels in VR_j, VR_j^* respectively. The whitespace analysis of a vertical homogeneous region is called convergence if the ratio of S_j and S_j^* is 1. In other words, the process of whitespace analysis cannot be find any non-text components and the region obtained after this step include only text elements. Besides, at this time, all horizontal homogeneous region HR_j^* in VR_j^* are the maximum horizontal homogeneous regions, see Fig. 3c.

The text document (f_t) is then created by the combination of VR_j^*

$$f_t = VR_1^* \cup VR_2^* \cup \ldots \cup VR_n^* \quad (6)$$

and the non-text document (f_{nt}) is obtained by the xor logical, $f_{nt} = f \oplus f_t$.

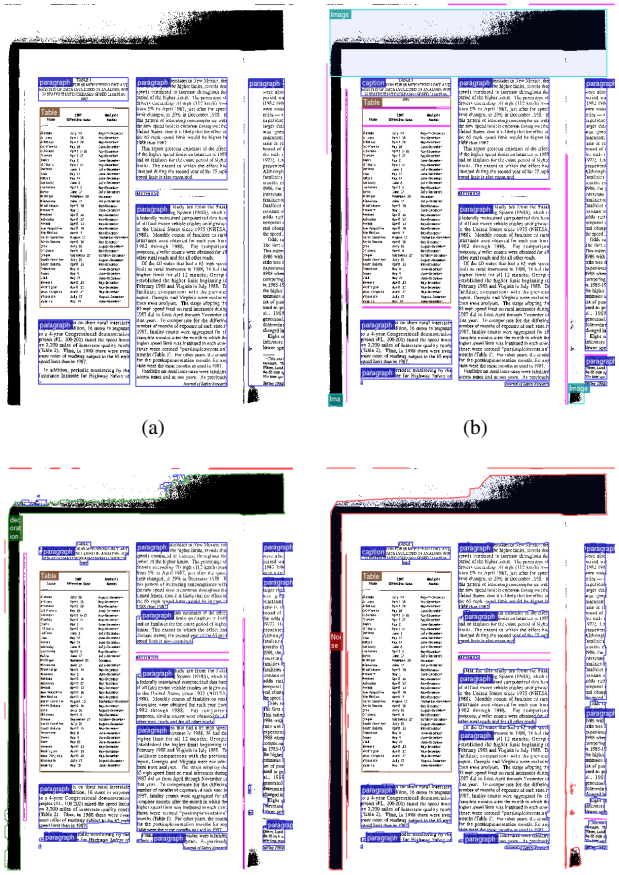
D. Text segmentation and non-text identification

1) Text segmentation: On each HR_j^*, $CCu \subset CCs$ is the set that includes all connected components. Firstly, we extract the bounding box of CC_i. Then, $\forall CC_i, CC_j \in CCu$ if

$$\begin{cases} CC_j \in RNN_i, r_i \leq max(H_i, H_j) \\ \min(H_i, H_j) >= 1/2 \times max(H_i, H_j) \end{cases} \quad (7)$$

CC_i and CC_j are connected together to deduce the text line (RNN_i is the set of right nearest neighbors of CC_i, and r_i is the distance between CC_i and CC_j). After that the mathematical morphology is applied to make the text region. Based on the mean of whitespaces and the large of line spaces in HR_j^* we can compute the size of kernel size in morphology [29].

Moreover, HR_i^* is considered to link with another HR_j^* to extract a new text region if HR_j^* is the below nearest neighbors of HR_i^* [19], the mean height of text line in HR_i^* is approximately in HR_j^*, the large of white line between them less than a half of minimum height of text lines or approximate the mean of white lines in two regions. This method proved to be very effective in clustering text regions; especially in the case where the region is considered have only one text line, (Fig. 3d).



(a) (b) (c) (d)

Fig. 4. Example of the page segmentation results; (a) Fine Reader Professional 12, (b) Tesseract OCR 3.04, (c) The proposed method, (d) Ground truth.

2) Non-text classification: All non-text components obtained are identified to separators, tables, and images in this subsection. The line, table, separator in the non-text document are identified in the order. The remaining components are considered to be images [4]. In our system, a machine learning approach is used for detecting line and table regions [30]. The input of [30] algorithm is the text (f_t) and non-text (f_{nt}) documents that are obtained in the previous step. This method uses Random Forest (RF) classifier to get the probability of each text line (corresponding to its local, contextual features) belong to the table region and SVM with RBF kernel classifier to get the final label for each text line.

Finally, all regions are obtained (include text and non-text regions) will be reshaped (returned to the original coordinate) to get the page layout.

III. Experimental Results

Our algorithm underwent several tests with different datasets and produced relatively positive results, see Fig. 4. The evaluation of document layout analysis is always complicated because they depend heavily on database and ground-truth. In this paper, we use dataset UWIII-A1 for the evaluation. The UWIII-A1 contains 50 images (we are expanding up to 100 images) that are extracted from UWIII (University of Washington) dataset. UWIII contains a large variety of binary document images (more than 1600 scanned images) [31]. The created dataset also contains images with a number of tables that are of a variety of structures. UWIII contains only binary images; therefore, it was possible to evaluate the efficiency of the page segmentation systems without a consideration of the quality of the binarization. The ground-truth of this dataset is made by using Athelia tool which was supplied by PRImA lab [1], the organizer of the page segmentation competition series. The details of this dataset could be found and downloaded in [33].

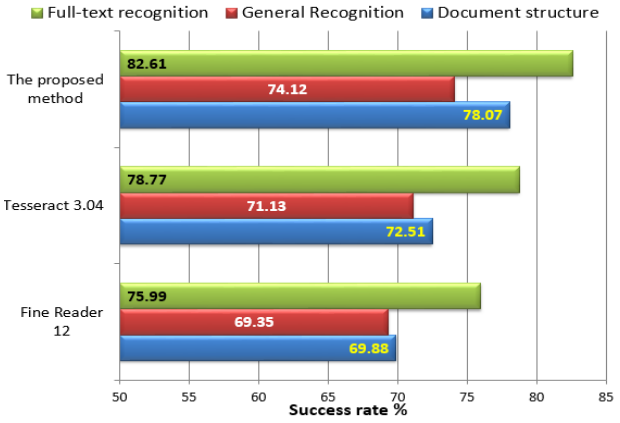


Fig. 5. Result of F-measure evaluation profile

The setting for evaluation follows the instruction of Clausner et al. [32]. In which there are three main criteria are considered: Document structure, General recognition, and Full-text recognition (PRImA measure). These evaluation profiles of our method and two other well-known and commercial systems (Tesseract 3.04 and Fine Reader Professional 12) are presented in Fig. 5. For the computation complexity, our system is faster than other multilevel algorithms. The processing time reduces $1/3$ whereas the accuracy is still guaranteed (compare to [4]).

IV. Conclusion

In this paper, we proposed an efficient hybrid page segmentation method based on whitespace analysis on the maximum horizontal homogeneous region. Our method achieves the balance between accuracy and processing time. It can work well on many document image resolutions. Our method is going to be tested on various public datasets. Experimental results on UWIII-A1 gave high performance and promising (this dataset is also expanding up to 100 images). However, the proposed algorithm still has several limitations. For example, it is sensitive to the skew document and depends on the quality of the input binary image.

Using a statistical approach, our algorithm is designed to be applied for machine learning modeling. This is an important point for the development of document analysis system to adaptive to the variety of document layout. We are going

extract the features in the document image and apply the machine learning to deal with the page segmentation problem.

REFERENCES

[1] C. Clausner, S. Pletschacher and A. Antonacopoulos, Scenario driven indepth performance evaluation of document layout analysis methods, *Proc. 11th ICDAR*, pp. 1516-1520, 2011.

[2] A. Antonacopoulos, S. Pletschacher, D. Bridson, C. Papadopoulos, ICDAR2009 page segmentation competition, *Proc. 10th ICDAR*, pp. 1370-1374, 2009.

[3] K. Kise, A. Sato, M. Iwata, Segmentation of page images using the area Voronoi diagram, *Computer Vision Image Understanding*, vol. 70, no. 3, pp. 370382, 1998.

[4] T. A. Tran, I. S. Na, S. H. Kim, Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology. International Journal on Document Analysis and Recognition, 19(3), 191209, 2016

[5] M. Agrawal, D. Doermann, Voronoi++: A dynamic page segmentation approach based on Voronoi and Docstrum features, *Proc. 10th ICDAR*, pp. 10111015, 2009.

[6] L. OGorman, The document spectrum for page layout analysis, *IEEE TPAMI*, vol. 15, no. 11, pp 11621173, 1993.

[7] A. Simon, J. C. Pret and A. Peter Johnson, A Fast Algorithm for Bottom-Up Document, *IEEE TPAMI*, vol. 19, no. 3, pp. 273-277, 1993.

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

[9] H. Baird, S. J ones and S. Fortune, Image segmentation by shape-directed covers, *Proc. 10th ICPR*, pp. 820-825, 1990.

[10] F. M. Wahl, K. Y. Wong and R. G. Casey, Block segmentation and text extraction in mixed text/image documents, *Graphical Models and Image Processing*, vol. 20, no. 4, pp. 375-390, 1982.

[11] G. Nagy, S. C. Seth and S. D. Stoddard, Document Analysis with an Expert System, *Pattern Recognition in Practice*, vol. 2, pp. 149-159, 1986.

[12] S. Ferilli, T. M. A. Basile and F. Esposito, A histogram based technique for automatic threshold assessment in a run length smoothing based algorithm, *The 9th IAPR International workshop on document analysis system*, 2010.

[13] H. M. Sun, Page segmentation for Manhattan and non-Manhattan layout documents via selective CRLA, *Proc. 8th ICDAR*, pp. 116120, 2005.

[14] S. W. Lee, D. S. Ryu, Parameter - Free Geometric Document Layout Analysis, *IEEE TPAMI*, vol. 23, no. 11, pp. 1240 1256, 2001.

[15] H. Cheng, C. A. Bouman, Multi-scale Bayesian Segmentation Using a Trainable Context Model, *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 511-525, 2001.

[16] T. A. Tran, H. T. Tran, I. S. Na, G. S. Lee, H. J. Yang, S. H. Kim, A mixture model using random rotation bounding box to detect table region in document image. International Journal of Visual Communication and Image Representation, 39, 196208, 2016.

[17] J. Ha, R. M. Haralick and I. T. Phillips, Recursive X-Y Cut Using Bounding Boxes of Connected Components, *Proc. 3rd ICDAR*, pp. 952-955, 1995.

[18] J. Liang , J. Ha, R. M. Haralick and I. T. Phillips, Document Layout Structure Extraction Using Bounding Boxes of Different Entities. *Proc. 3rd IEEE Workshop on Applications of Computer Vision*, pp. 278-283, 1996.

[19] K. Chen, F. Yin and C. L. Liu, Hybrid Page Segmentation with Efficient Whitespace Rectangles Extraction and Grouping, *Proc. 12th ICDAR*, pp. 958-962, 2013.

[20] Y. Pan, Q. Zhao and S. Kamata, Document Layout Analysis and Reading Order Determination for a Reading Robot, *Tencon2010-IEEE Region 10 Conference*, pp. 1607-161, 2010.

[21] R. Smith, Hybrid Page Layout Analysis via Tab-Stop Detection, *Proc. 10th ICDAR*, pp. 241-245, 2009.

[22] F. Chang, C. J. Chen and C. J. Lu, A linear time component labeling algorithm using contour tracing technique, *Computer Vision and Image Understanding*, vol. 93, no. 2, pp. 206-220, 2004.

[23] T. A. Tran, I. S. Na, S. H. Kim, Hybird page segmentation using Multilevel Homogeneity Structure. *In Proceeding of ninth International Conference on Ubiquitous Information Management and Communication*, pp. 78:1- 78:6, ACM, 2015.

[24] F. Shafait, D, Keysers, T.M. Breuel, T.M, Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithm. *IEEE Trans. Pattern Anal. Mach. Intell*,vol. 30, no. 6, pp. 941-954, 2008.

[25] T. A. Tran, I. S. Na, S. H. Kim. Separation of text and non-text in document layout analysis using a recursive filter. *KSII Transaction on internet and information systems*, vol. 9, pp. 4072-4091, 2015.

[26] A. Papamdreou, B. Gatos, A Novel Skew Detection Technique Based on Vertical Projections, *Proc. 11th ICDAR*, pp. 384-388, 2011.

[27] M. Okamoto, M. Takahashi, A hybrid page segmentation method. *Proc. 2nd ICDAR*, pp. 743-746, 1993.

[28] C. Mallows, Another comment on OCinneide. *American Statistician*, vol. 45, no. 3, 256-262, 1991.

[29] T. A. Tran, K. Oh, I. S. Na, G. S. Lee, H. J. Yang, S. H. Kim, A Robust System for Document Layout Analysis using Multilevel Homogeneity Structure,*Expert Systems With Applications*, vol. 85, pp. 99-113, 2017.

[30] T. Huynh-Van, K. T. Le-Ba, T. A. Tran, K. Nguyen-An, H. J. Yang, S. H. Kim, Learning to Detect Tables in Document Images using Line and Text Information, *ACM: Proc. International Conference on Machine Learning and Soft Computing*, pp. 99-113, 2018.

[31] IT Phillips. Users reference manual for the UW english/technical document image database III. *UW-III English/Technical Document Image Database Manual*, 1996.

[32] C. Clausner, A. Antonacopoulos, and S. Pletschacher, ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL2017, *Proc. ICDAR2017*, doi: 10.1109/ICDAR.2017.229, 2017.

[33] T. A. Tran, UWIII-A1 dataset, https://sites.google.com/site/trtanh1988/dataset, 2018.