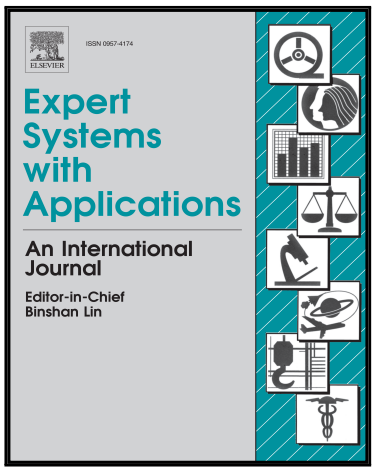


A Robust System for Document Layout Analysis Using Multilevel Homogeneity Structure

Tuan Anh Tran, Kanghan Oh, In-Seop Na, Guee-Sang Lee, Hyung-Jeong Yang, Soo-Hyung Kim

PII: S0957-4174(17)30346-9
DOI: 10.1016/j.eswa.2017.05.030
Reference: ESWA 11322

To appear in: *Expert Systems With Applications*

Received date: 9 January 2017
Revised date: 10 May 2017
Accepted date: 11 May 2017

Please cite this article as: Tuan Anh Tran, Kanghan Oh, In-Seop Na, Guee-Sang Lee, Hyung-Jeong Yang, Soo-Hyung Kim, A Robust System for Document Layout Analysis Using Multilevel Homogeneity Structure, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.05.030

Highlights

- This paper presents a robust system for the document layout analysis.
- The proposed system is based on multilevel homogeneity structure (MHS).
- The proposed system is designed to work with many different document languages.
- Our system is tested on four published datasets with different document languages.
- The proposed system (MHS) won the RDCL-2015 competition (ICDAR2015).

A Robust System for Document Layout Analysis Using Multilevel Homogeneity Structure

Tuan Anh Tran[a], Kanghan Oh[b], In-Seop Na[b], Guee-Sang Lee[b], Hyung-Jeong Yang[b], Soo-Hyung Kim[b,c]

[a]*Faculty of Computer Science and Engineering, HoChiMinh City University of Technology, 268 Ly Thuong Kiet, District 10, Ho Chi Minh City, Viet Nam.*
[b]*School of Electronics and Computer Engineering, Chonnam National University, 77 Yongbong-ro, Gwangju 500-757, South Korea.*
Email: trtanh@hcmus.edu.vn, blastps@naver.com, ypencil@hanmail.net, gslee@jnu.ac.kr, hyungjeong@gmail.com, shkim@jnu.ac.kr
[c]*Corresponding author: Soo-Hyung Kim*

Abstract

One of the difficulties in the understanding of document images is document layout analysis, which is the first step in document image modeling. In this paper, a robust system for which a multilevel-homogeneity structure is used in accordance with a hybrid methodology is proposed to deal with this problem. Our system consists of the following three main stages: classification, segmentation, and refinement and labeling. Different from other page segmentation methods, the proposed system includes an efficient algorithm to detect table regions in document images. Besides, to create an effective application, the proposed system is designed to work with a variety of document languages. The proposed method was tested with the ICDAR2015 competition (RDCL-2015) and three other published datasets in different languages. The results of these tests show that the accuracy of proposed system is superior to the previous methods.

Keywords: Document layout analysis, multilevel homogeneity structure (MHS), page segmentation, document image processing, OCR.

1. Introduction

Document layout analysis is the process of identifying and categorizing the regions of interest in a document image. This process involves a separation of the document into zones, and a subsequent classification of individual

zones into one of the categories of texts, tables, images, or lines. OCR (Optical Character Recognition) systems require the segmentation of text regions from non-text regions and the arrangement of the former in their correct reading order. Obviously, the quality of the layout analysis can determine the quality of the whole document processing activity.

In this paper, a robust document layout analysis system for which the homogeneity structure is used as a basis structure is presented. For the proposed system, the following three stages are implemented: First, the text and non-text elements of the document image are classified by the use of multilevel and multi-layer homogeneity structure. Second, the text elements in the text document are grouped and segmented based on the combination of text line extraction, paragraph segmentation, and homogeneity. Besides, the non-text elements are further identified to negative-text regions, line, table, separator, and image regions. Third, all of the regions are refined to remove the noise and enhance the quality of each region. Then, all of the text regions are labeled based on their position and textual characteristics.

Different from the other systems, the proposed system involves an efficient method for table detection in the document image through the use of "Random Rotation Bounding Box" (abbreviated to "RB"). All of these processes are designed for consistency to improve the effectiveness of the proposed system.

The proposed system is also designed to work with not only the Latin languages but also the non-Latin languages, thereby making the system more efficient in terms of real applications. Like other page segmentation methods, the input document is assumed non-skew; however, if it is necessary, a process for the detection and correction of a skew or warp may be invoked as an optional step.

This paper is organized as follows: Section 2 presents the review of well-known and recent document layout analysis techniques; Section 3 presents an overview of our document layout analysis algorithm; the experiment results are covered in Section 4, and Section 5 presents the conclusion and future research.

2. Related works

A number of document layout analysis systems have been proposed, and they can be divided into the following four categories: bottom-up, top-down, hybrid, and multi-scale resolution.

One of the earliest approaches is the top-down (Nagy et al., 1992) (Baird et al., 1990) (Wahl et al., 1982) (Sun, 2005) (Ha et al., 1995). For this method, the original document is typically separated into many different regions, followed by the classification of each region by the use of many heuristic filters. The top-down method is fast and efficient for Manhattan layout document; however, the recognition of non-Manhattan layout becomes more difficult, as the effectiveness of the top-down algorithm is annulled.

In terms of the bottom-up methods (Kise et al., 1998) (Agrawal & Doermann, 2009) (O'Gorman, 1993) (Simon et al., 1997), the local information is first employed to determine the words that are then merged into the text lines, and text blocks (or paragraphs). Bottom-up methods are widely applicable to a variety of document layouts, but at the least, their complexity is quadratic in time and space. A number of improvements have been proposed for this approach, such as (Simon et al., 1997), (Ferilli et al., 2010), and (Caponetti et al., 2008); however, there still remains many problems regarding the processing time, selection thresholds, requirement for a large training dataset, and the identification of the non-text regions. A detailed comparison and benchmarking of the well-known top-down and bottom-up methods can be found in (Shafait et al., 2008).

Another approach for which the multi-resolution method is used was proposed in the late 1990s and early 2000s (Cinque et al., 1998) (Lee et al., 2001) (Cheng & Bouman, 2001); this approach is based on the analysis of a set of feature maps at different resolution levels. The methods which use this approach yielded relatively positive results for English (Latin) documents, but the use of multi-scale resolution resulted in a quite long computing time. Moreover, due to the use of wavelet, the difficulties arise when the size of the text elements in a document is similar to that of the non-text elements (significant drop-capital or title of the paper). Besides, a detailed classification of the non-text elements has not received much attention.

During the last several years, a number of document layout analysis systems have been proposed in relation to the bottom-up approach such as ISPL, PAL, and AOSM. For ISPL (Koo & Kim, 2013) method, several techniques such as scene text detection, salient object detection, connected component analysis, and line segment detector have been combined (Gioi et al., 2010). For PAL method (Chen et al., 2013), the edge boxes of the text proposals (Zitnick & Dollr, 2014) and Otsu's algorithm are employed to binarize the images; here, an SVM is used with the features extracted from individual elements (e.g., skeleton, stroke width, color) to classify the text and non-text

components (Antonacopoulos et al., 2015). Then, a white-space analysis is used to extract the text blocks. The primary focus of AOSM method (Ha et al., 2016) is the segmentation of the text through the utilization of adaptive over-split and merge algorithm, while the classification of text and non-text components is based on a filter which uses the morphological process (Bukhari et al., 2011). The performances of these methods are positive; however, they are relatively complex and inconsistent due to the combination of a variety of techniques. Their performances regarding binary images differ from the expectations due to the use of a color-depth feature or a simple filter for the classification process. Besides, these methods are not good for non-English documents, in which classification errors occur frequently for large characters, and the consideration of the non-text elements is not robust.

By using a combination of the top-down and bottom-up methods, a hybrid approach has been proposed to overcome the common weak points. This method was first proposed during the 1990s (Okamoto & Takahashi, 1993) (Jain & Yu, 1998), but its usage has been continued (Chen et al., 2013) (Smith, 2009). These algorithms focus on an analysis of the connected components and the white spaces between them; however, this analysis is based on the first level of the binary document, so the corresponding results are unsatisfactory, especially the classification of the non-text regions. In 2016, Tran et al. (2016b) proposed a hybrid method using multilevel homogeneity structure to analyze document layout. Even though this system proves the effectiveness in the classification of text and non-text on the benchmark datasets (Sebastien et al., 2017), their Minimum Homogeneity Algorithm still has problems with the heuristic filter and multilevel classification in small homogeneous regions. Besides, the layout text segmentation process also has difficulties with the text clustering where the simple projection is used in text segmentation (see Figure 6).

Although most of the page segmentation methods follow the rule-based approach, we should mention the learning based methods in document layout analysis. Up to now, there is no full learning-based system (image-to-layout). Instead, several partial-learning-based approaches are proposed in the recent year such as PAL, ISPL (Antonacopoulos et al., 2015), Bukhari (Bukhari et al., 2010). Bukhari's method applies a multi-layer perceptron (MLP) classifier with shape and context features to classify text and non-text components. PAL uses SVM with features extracted from the skeleton, stroke width, and textual color, whereas ISPL applies a scene text detection technique. These methods use the learning-based method for the classification purpose and

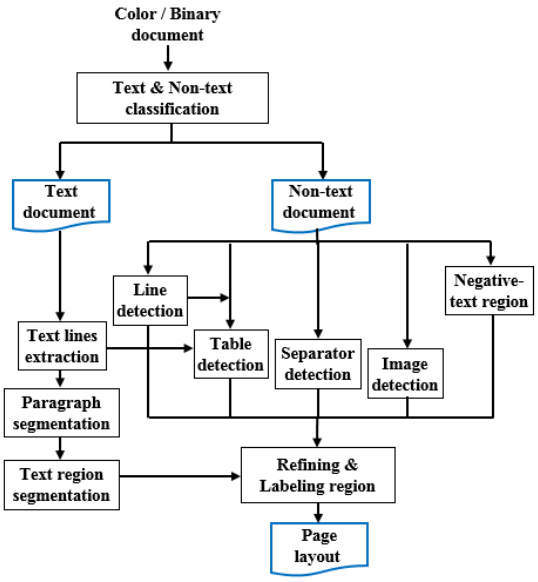


Figure 1: Block diagram of the proposed system.

rule-based one for the remaining parts. However, the result of classification is unexpected especially for the binary image or non-Latin language document. These methods often cause the detection errors with big text characters or can be confused with the regions of different types with a similar texture. The learning-based approach still requires more improvement in the future because the difficulty in extracting feature vectors as well as the limitation of datasets.

Summarizing the state of the art in document layout analysis, we can see the limitation of existing methods. These are usually designed for one specific document language (English) and do not work well (or cannot be extended) for other document languages. These methods are not stable for real document image datasets. The performance is unsatisfactory for the document having a complex layout, where the text and non-text classification is restricted. Besides, these methods usually ignored the table detection, although it is an important part of document layout analysis.

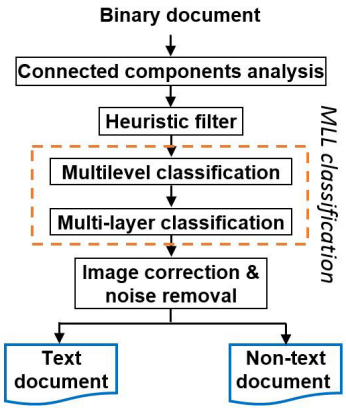


Figure 2: Flowchart for the classification process (MHA) of the MHS system.

3. System Overview

The input of the proposed system could be the color or binary image that is obtained via scanning process. The output of the proposed system is a set of separated zones with their labels in XML format (see Figure 12(b)). If f is the binary image after a binarization step, the analysis of document layout proceeds as follows (Figure 1):

- *Step 1.* Text and non-text classification with the minimum homogeneity algorithm (Sect. 3.1).
- *Step 2.* Text segmentation and non-text identification (Sect. 3.2).
- *Step 3.* Refinement and labeling of individual regions (Sect. 3.3).

3.1. Text and non-text classification

In this section, a method for the classification of text and non-text elements in a binary document image is presented. It is based on the Minimum Homogeneity Algorithm (MHA) which was first proposed by the authors in 2016 (Tran et al., 2016b). In general terms, the classification of text and non-text elements by the multilevel homogeneity structure (MHS, an updated version of MHA) proceeds as follows (Figure 2):

1. First, a connected components analysis is performed to extract the set of connected components (CCs) and their properties.

2. Second, based on the connected component properties, a heuristic filter is applied to remove the obvious non-text components and noise. The following five text features are used in this filter: area, density (λ_i), number of contained elements (Inc_i), and ratio between width and height $\left(\gamma_i = \frac{\min(H_i,W_i)}{\max(H_i,W_i)}\right)$ of the i^{th} connected component (CC_i).

3. Third, a multilevel/multi-layer classification (MLL) is performed to classify all of the text and non-text elements (Section 3.1.2). In this process, we utilize a combination of the multilevel and multi-layer homogeneous regions and white-space analysis in an iterative process.

4. At the end of this process, a simple technique using the bounding box of the text elements is performed to remove noise and to correct the text and non-text components.

To make the system work with many different document languages and improve the performance, two main upgrades have been applied upon MHA.

3.1.1. Heuristic filter

In the proposed system, the heuristic filter for the classification purpose is reconsidered to enable it to work efficiently with many different document languages, including Latin and non-Latin (e.g., Korean) languages. Compared to the old version, there are two changes in this filter as follows:

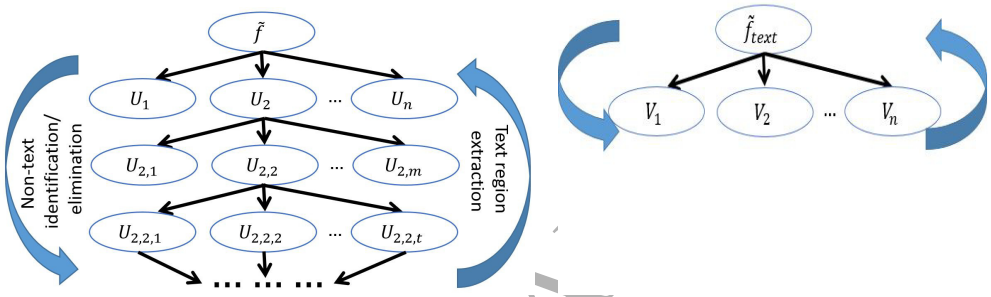
- $Inc_i > T^{inside} = 4$; that is, if the bounding box of CC_i contains more than four other components, it can be regarded as a non-text element. This value is optimized by an analysis on the letters/words of different languages such as English, Korean, Chinese, and Arabic. Let σ be the parameter of the input document language, $\sigma \in \{English, France, Vietnamese, Korean, etc.\}$. As we can see in Figure 3,

$$\begin{cases} \forall i \in \mathbb{N}, Inc_i \in [0,3] \text{ if } \sigma \in \{\text{Latin languages}\} \\ \forall i \in \mathbb{N}, Inc_i \in [0,5) \text{ if } \sigma \in \{\text{Non-Latin languages}\} \end{cases} \quad (1)$$

  Based on this observation, the threshold T^{inside} is set to 4 (for all type of document languages) to ensure the accuracy of this filter, especially for the special language documents (e.g., Arabic, Vietnamese) or documents having a lot of noise (see Figure 3(b)).

(a) English, $Inc_1 = 0; \{\emptyset\}$ (b) Korean, $Inc_1 = 2; \{2, 3\}$ (c) Chinese, $Inc_1 = 3; \{2, 3, 4\}$ (d) Arabic, $Inc_1 = 4; \{2, 3, 4, 5\}$

Figure 3: Example numbers of contained elements (Inc_i) on different languages



(a) Multilevel classification (b) Multi-layer classification

Figure 4: Illustration of multilevel classification and multi-layer classification.

- The condition $(\gamma_i < 0.06) \wedge (H_i < W_i)$ is removed because it may cause an error with respect to the Korean language containing the component "_".

These changes do not decrease the accuracy of the proposed system; this is because the classification process of the proposed system is not primarily based on the heuristic filter. This filter serves to find the distinctive non-text elements. Therefore, it may not detect some non-text elements, but it does not cause any false positive error or rejected text characters. The main idea of the text and non-text identification is the MLL classification (Section 3.1.2).

3.1.2. MLL classification

This is the primary process of text and non-text classification. The MLL classification that is a combination of Multilevel and Multi-layer classification (see Figure 2). Both of them are based on a recursive filter. This is the core and a major improvement over the previous algorithm. This upgrade

helps us to overcome the errors in the diacritic document languages such as Vietnamese, France, Arabic, etc.

Let $\tilde{f}$ be a binary document obtained after the heuristic filter. First, the multilevel classification is applied.

a) Multilevel classification

In this process, $\tilde{f}$ is first segmented to a set of homogeneous regions $U_j, j = \overline{1,n}$, $l = 1$

$$\tilde{f} = U_1^l \cup U_2^l \cup \ldots \cup U_n^l \tag{2}$$

Then, for each U_j^l, a recursive filter is applied to identify and eliminate the non-text elements (in an iterative way, $l = l + 1$) to extract the text homogeneous region U_j^{text} (see Figure, 4(a)). Finally, a text document image $\tilde{f}_{text}$ is conducted by a combination of U_j^{text}.

$$\tilde{f}_{text} = U_1^{text} \cup U_2^{text} \cup \ldots \cup U_n^{text} \tag{3}$$

As we can see in the above multilevel classification, the quality of classification depends on the quality of the first level homogeneous regions U_j^1. Therefore, in some cases, especially with non-Latin languages (e.g., Korean, Chinese) or diacritic languages (e.g., Vietnamese, France), the multilevel classification still does not provide satisfactory results. When the binary document image has a lot of noise, the MHA cannot identify the small components or the components located on (or nearby) the border of the homogeneous region.

b) Multi-layer classification

To overcome the above problems as well as to improve the performance of MHA algorithm, we propose another classification step, called multi-layer classification. This process is applied just after the multilevel classification. First, the $\tilde{f}_{text}$ is re-segmented into homogeneous regions $V_j, j = \overline{1,m}$

$$\tilde{f}_{text} = V_1 \cup V_2 \cup \ldots \cup V_m \tag{4}$$

For each V_j, a recursive filter (Tran et al., 2016b) is applied only one time to extract the V_j^{text}. Then, all of V_j^{text} are combined to extract the next layer text document $\tilde{f}_{text}^k, k = 1, 2, \ldots$.

$$\tilde{f}_{text}^k = V_1^{text} \cup V_2^{text} \cup \ldots \cup V_m^{text}, k = k + 1 \tag{5}$$

This process is performed repeatedly until we cannot find any non-text element in $\tilde{f}_{text}^{k}$ (or satisfy the convergence condition $\Psi = \sum \tilde{f}_{text}^{k+1} / \sum \tilde{f}_{text}^{k} \approx 1$). Then, a final text document f_t is extracted. In the process of multi-layer classification (Figure 4(b)), we do not try to segment the homogeneous region into smaller (higher level) regions. Instead, many layers of $\tilde{f}_{text}$ are considered with the recursive filter to identify the non-text elements and to extract f_t.

The multi-layer classification process can identify small non-text elements better than the multilevel classification step. In the experiment, the multi-layer classification often satisfies the convergence condition after the first or second layer. Therefore, this process helps us to extract a better result in the classification process, but it does not increase much processing time. According to many experiments in which different document images were used, this filter always provides a superior performance with respect to all types of documents, especially for low-resolution images.

In document layout analysis, most of the previous methods often depend heavily on the language of the input document because the models of these methods are designed for a specific language. Different from other systems, our MHA algorithm is language independent. Our model can be applied to any type of document language. As we can see in Figure 5, although the MHS system is kept in default mode (English), the results for Arabic or Vietnamese document language are very encouraging.

As shown in Figures 8(a), 8(b), 8(c), the output of the classification process consists of the following two documents: text documents f_t (containing only the text elements) and non-text documents f_{nt} (containing only the non-text elements).

Let $CCs^T, CCs^N \in CCs$ be the sets containing all of the connected components of f_t and f_{nt}, respectively. In this paper, $(Xl_i, Yl_i), (Xr_i, Yr_i), H_i, W_i$, $\mathcal{A}_i, \mathcal{A}_i^B$ and λ_i, γ_i are denoted as the top-left coordinates, bottom-right coordinates, height, width, area, bounding-box area, density, and ratio of the width and height of CC_i, respectively.

3.2. Text segmentation and non-text identification

The text elements in f_t should be grouped together for the deduction of text regions, while the non-text elements in f_{nt} should be identified as negative-text regions, separator regions, table regions, or image regions.

(a) Original

(b) Text document

(c) Original

(d) Text document

Figure 5: Example results of the proposed system (default mode) on different document languages, (a,b) Arabic language, (c,d) Vietnamese language.

3.2.1. Layout text segmentation

In this section, a method for the grouping of text elements into a set of text regions is presented. First, all of the text elements in f_t are grouped row by row to extract the text lines based on the white-space analysis. Then, the homogeneous regions are extracted, and the paragraph segmentation is performed to separate the text document into individual paragraphs or heading

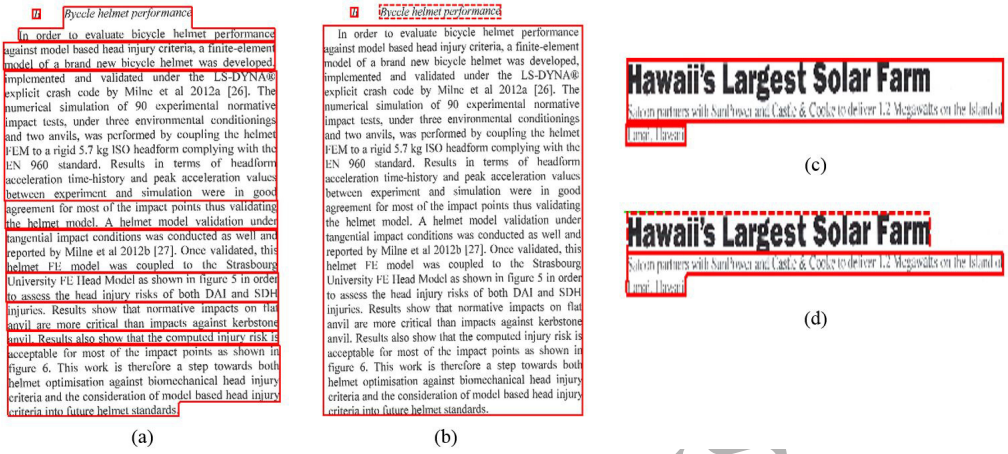


Figure 6: Examples of text segmentation in which a projection and a white-space analysis are used to find the homogeneous regions (in red outlines): (a, c) horizontal projection and (b, d) white-space analysis.

sentences.

Instead of using horizontal and vertical projections as in Tran et al. (2016b), in this system, the text lines created by the white-space analysis are used to find the homogeneous regions. This will eliminate the mistakes caused by the distortion due to the touching characters in adjacent text lines (Figure 6), or the skew document (Figure 5(c)).

a) Text line extraction

In this step, the text elements in f_t are grouped together to extract the text lines (TL) and then the text line document f_t^{TL} (see Figure 8(d)). Following the extraction, $\forall CC_i, CC_j \in CCs^T$, CC_i and CC_j are connected if

$$\begin{cases} \max(Yl_i, Yl_j) - \min(Yr_i, Yr_j) < 0 \\ |Xl_i - Xr_j| \leq \theta \times \max(H_i, H_j) \\ \max(H_i, H_j) \leq 2 \times \min(H_i, H_j) \end{cases} \quad (6)$$

The text line document f_t^{TL} is used not only for the extraction of text regions but also for the table detection. Due to the differences between the structures of the characters in each language, the value of θ depends on the language in the document. For instance, in Korean or Chinese documents, the text lines having the same font letters will be of the same height, whereas

Figure 7: Example of paragraph segmentation.

this alignment cannot be guaranteed for the English (Latin) documents. This is because it depends on the upper-line ("t, h, l, etc.") or lower-line ("q, p, g, etc.") letters in a text line. In usual, the distance between two words in Latin language (English, Vietnamese) is less than the maximum height of them. Besides, the height of two adjacent elements of the same line cannot be doubled. Therefore, the value of θ is set as follow:

$$\begin{cases} \theta = 1.0 \text{ if } \sigma \in \{\text{Latin languages}\} \\ \theta = 1.3 \text{ if } \sigma \in \{\text{Non-Latin languages}\} \end{cases} \quad (7)$$

Actually, the MHS system is insensitive to the value of θ. The page segmentation is almost independent of it. As mentioned above, the main task of the text line extraction is to increase accuracy in estimating the height of the text line. The errors will be eliminated by the combination of homogeneous region and mathematical morphology in the text region segmentation. However, the table detection and table decomposition processes are the opposite. Our primary experiment indicates that the suitable value for θ is in the range $[1.0, 1.3]$ for English and Vietnamese (Latin) language, and θ is in the range $[1.1, 1.5]$ for the Korean (non-Latin) language.

b) Paragraph segmentation

First, the text lines in table regions are eliminated. Based on the remaining f_t^{TL}, all the homogeneous regions $HR_k, k = \overline{1,n}$ are extracted by the use of a combination of homogeneity structure and text line merging (Chen et al., 2013), as shown in Figure 8(d). On each HR_k, the paragraph segmentation is performed as follows.

The text line $(tl_i, i = \overline{1, num(TL)})$ in HR_k is usually aligned vertically (left, right, or both of them); therefore, the left indent (right indent) of the first text line (last text line) of each paragraph in HR_k is used to segment the region into paragraphs, as shown in Figure 8(e). For the proposed method,

a filter that scans three adjacent text lines $(tl_{i-1}, tl_i, tl_{i+1})$ simultaneously is used to obtain the segmentation position, as shown in Figure 7.

Note that the paragraph segmentation process is applied in the regions that have more than three text lines and are of a sufficient width.

c) Text region segmentation

After the paragraph segmentation, all of the text lines in each paragraph are grouped together to deduce the corresponding text regions, as shown in Figure 8(f). Let $HR_j^*, j = \overline{1,m}$ $(m \geq n)$ be a homogeneous region obtained after the paragraph segmentation. First, all of the text lines in each HR_j^* are merged if their heights and vertical distances are similar. Then, an adaptive morphological closing with a rectangle kernel is applied to extract the corresponding text region. To compute the kernel size of each HR_j^* (after the merging of the text lines), the following equation is used:

$$\begin{cases} height_{kernel} = 2 \times Q_3\{w\} \\ width_{kernel} = 2 \times Q_3\{WS\} \end{cases} \tag{8}$$

where Q_3 is the upper quartile of the corresponding set, w is the set containing all of the heights of the white lines w_i (vertical distance between the two adjacent text lines), and WS is the set of distances between the two nearest text elements.

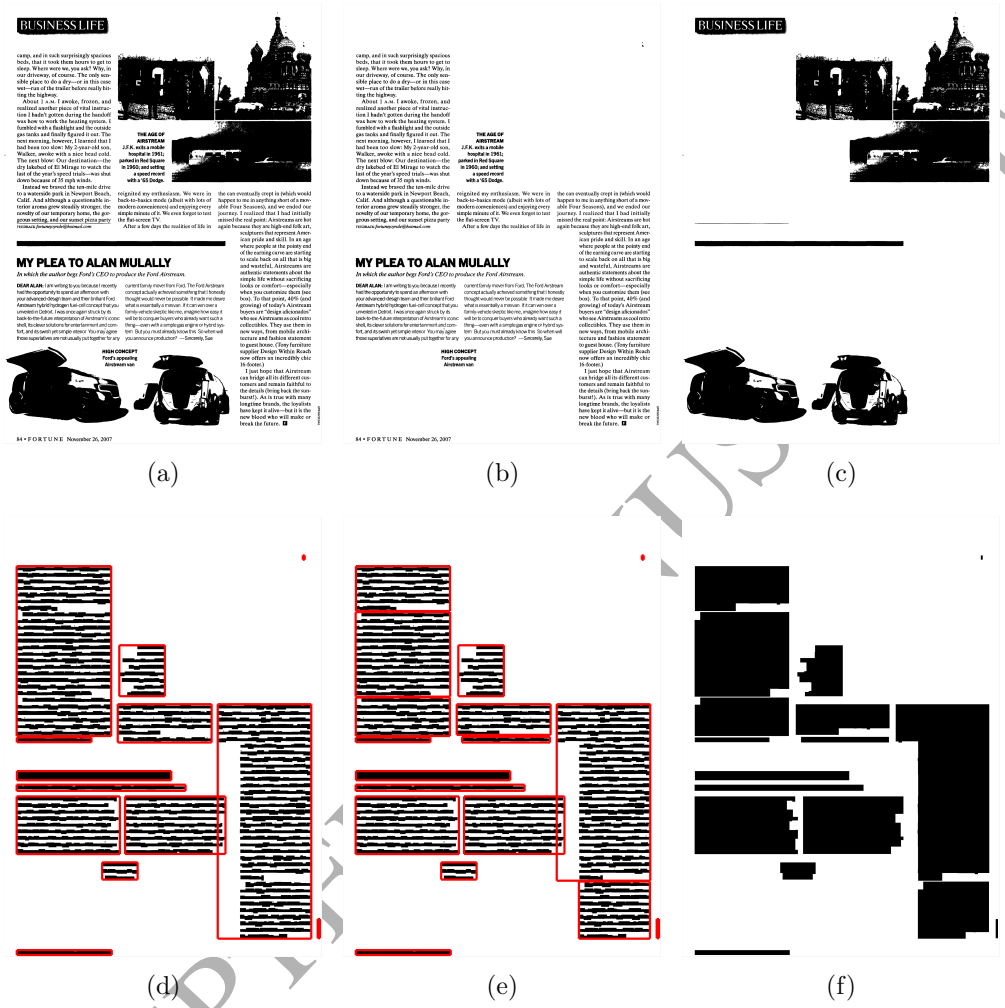
In addition to minimizing the computation time, the use of mathematical morphology in the text-line document (instead of the text document) helps us to reduce the errors cause by noise, non-allowable merge errors, and split errors (Clausner et al., 2011).

3.2.2. Non-text identification

In this section, a method for the identification of the components in the non-text document is presented. The five types of non-text regions are detected in the order of negative-text region, line, table, separator, and image.

a) Negative-text region detection

A negative-text region is an area that contains text in the background color surrounded by a rectangular block in the foreground color, as can be seen in Figure 12 (header/heading region), and Figure 9(a). The detection of this region proceeds as follows:



(a) (b) (c)

(d) (e) (f)

Figure 8: Example of (a) binary image, (b) text document, (c) non-text document (d) text line extraction and homogeneous regions HR_k (in red outlines), (e) the output of paragraph segmentation (HR_k^* in red outlines), and (f) the output of text region segmentation.

(a) Negative-text region candidate (b) Negatived image of (a)

Figure 9: Example of negative-text region and its negative-image.

A $CC_i \in CCs^N$ is suspected as a negative-text region, if $\lambda_i^F \approx 1$ (λ_i^F is the ratio of the CC_i that is filled with $\mathcal{A}_i^B$). Let $\mathcal{R}_i$ be the binary image region of the suspected CC_i, as in Figure 9(a). $\mathcal{R}_i$ is then negated (or inverted), so that pixel 0 becomes pixel 1, and vice-versa, as shown in Figure 9(b). The multi-layer classification is applied to the negative-image to detect the text and non-text elements. Let CCu^T, CCu^N are the sets of text and non-text elements respectively. If

$$\sum_{j=1,2,\ldots}^{CCu^T} \mathcal{A}_j > \sum_{t=1,2,\ldots}^{CCu^N} \mathcal{A}_t \tag{9}$$

$\mathcal{R}_i$ is identified as a negative-text region; otherwise, $\mathcal{R}_i$ is an ordinary (non-negative) image region.

It should be mentioned that, if $\mathcal{R}_i$ has more than one text line, the similarity between the white line (w_i) and the text line (b_i) is also guaranteed.

Lastly, as shown in Figure 12(a), all of the negative-text regions are added to the text document, whereas they are eliminated from the non-text document, as follows: $CCs^N = CCs^N \setminus CC_i, CCs^T = CCs^T \cup CC_i$.

b) Line detection

In the proposed system, the horizontal and vertical lines are identified separately because they will be used for detecting table regions. It is already known that the disparity between the height and the width of a line is very high; therefore, if $\gamma_i \leq 0.1$ and $W_i > H_i$ ($H_i > W_i$), the $CC_i \in CCs^N$ is classified as a horizontal (vertical) line.

c) Table detection

Detecting table regions in a document images is an essential problem because the tables are also important components of the document. Even though the use of table is common, most of the page segmentation methods disregard it; this leads to a low performance of segmentation accuracy for those documents that contain the table region (e.g., science journals, business magazines).

The tables in document image are generally separated into two main categories: ruling-line table (RL-T) and non-ruling line table (NRL-T), see Figure 10. The previous table detection methods usually use the one-side feature (text or non-text feature) to detect the table regions. Therefore,

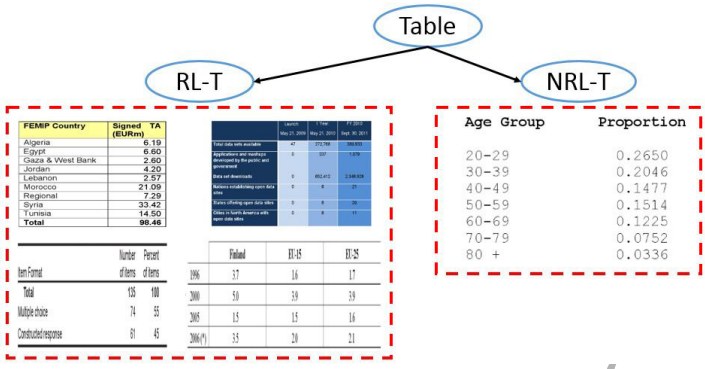


Figure 10: Table categories (in red bounding boxes).

they are ineffective in detecting the correct table region, especially with the document of complex layout. Besides, all of these methods use rectangular bounding box to describe the table region, so they cannot detect the tables from skewed documents.

To overcome this problem, an efficient method for the table detection is embedded in the proposed system. This method was also proposed by Tran et al. (2016a), and the effectiveness was proved on the UNLV and ICDAR 2013 table competition datasets. By using a new bounding box called "Random Rotation Bounding Box" (abbreviated to RB) to describe the table region, this method can detect any kind of tables, and even the skewed table in a non-skew document, as shown in Figure 13(a), Appendix A-Figure 18. As can be seen in Figure 11, our system considers both text and non-text features (based on the MHA algorithm's result). It consists of the following two main phases: RL-T detection and NRL-T detection. Moreover, a post-processing is added to the table detection algorithm to improve the performance as well as to reduce the errors caused by the noise in a document image. Here, the following two conditions are considered:

- First, for the case that the document image contains only one NRL-T region, the NRL-T detection process is applied again without the use of vertical homogeneous region. In other words, only the horizontal homogeneous regions of the document image are considered in this process of detecting the NRL-T region.

- Second, the detected NRL-T regions are eliminated if it contains images since NRL-T does not contain images.

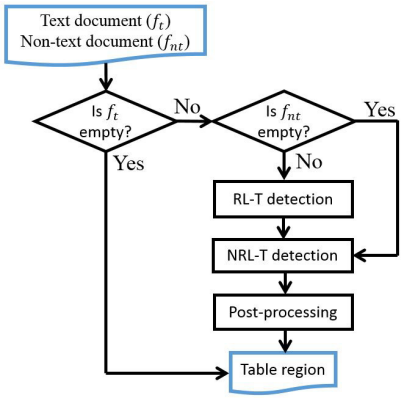


Figure 11: Flowchart for the table detection process.

This table detection method is designed to fit with the proposed document layout analysis system, as shown in Figure 1. Our algorithm provides an efficient and consistent method for table detection, while the increase of computation time is trivial.

d) Separator detection

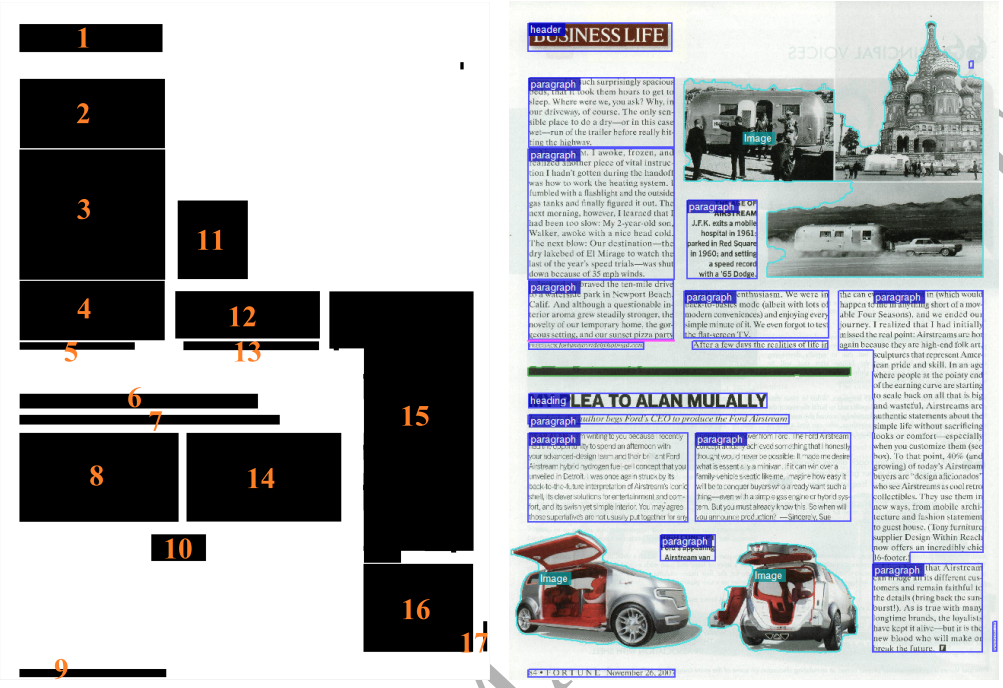
A separator is a combination of horizontal and vertical lines; these lines are often used to separate the text regions, so their bounding box area is usually gigantic, whereas its size is tiny; therefore, if λ_i is minuscule and its bounding box contains text elements, CC_i is classified as a separator.

e) Image (graphic) detection

After the classifications of negative-text regions, line regions, table regions, and separator regions, the rest of the non-text document comprises the image regions. A morphological dilation with a small disk kernel is applied to each image region to obtain their boundaries; here, the structuring element of the dilation is a small disk.

3.3. Region refinement and labeling

This section could be considered as a description of the post-processing of the proposed system. Firstly, all the regions (including text and non-text regions) are refined to eliminate the unexpected components and, if possible, to extract the rectangular shape of each region (see Figure 12(a)). Besides,



(a) Text refinement regions and its label (orange number)

(b) The final result of our system

Figure 12: An example of region refinement and region labeling.

the remaining regions are considered carefully to detect the noise. The details of this refinement process are described in Tran et al. (2016b).

After the region refinement, the final step for the labeling of each region is applied. Based on the position and text characteristics of each region, a simple approach with heuristic rules is used for the labeling of the text regions. For instance (see Figure 12(a)), if the considered region R_i is written vertically, it is labeled as credit (R_{17}); if most of the text elements in R_i are of the same height (capital letters), its label could be heading (R_6) or page-number based on the position, the number of text elements in the region, etc.

Due to the differences in the text structures of various languages, the labeling process of each language is slightly different. The Korean language, for example, does not comprise the capital letters, so the labeling of the heading region is based on its font size and position. Due to the complexity

of the page layout and the randomness of the text elements, the labeling of regions without OCR support is not a simple task. Instead of using heuristic rules, the application of training can also be considered in this step.

4. Experimental Results

4.1. System environment and database

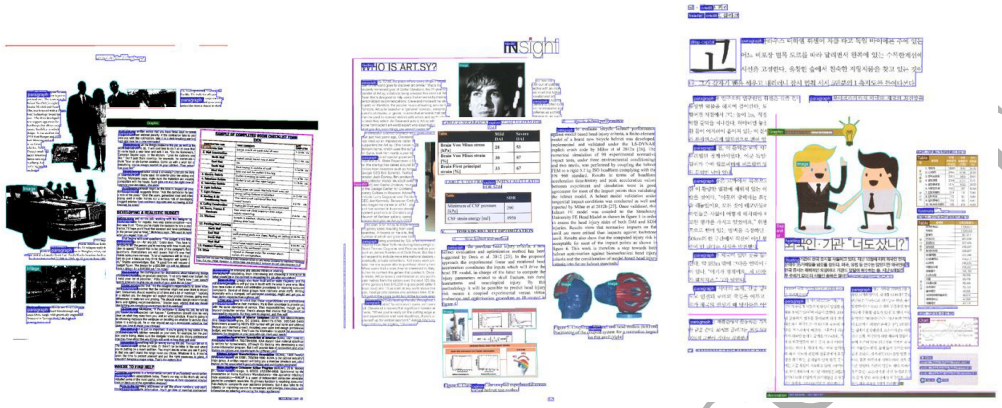
The proposed system was implemented using Matlab on a Core i5 3470 PC with 4 GB memory and running Windows 7. The first release version of the proposed system was converted to Visual C++ 2010 with OpenCV 2.4.3. To test the efficiency of the proposed method, experiments were conducted on four published databases.

The first dataset is the Recognition of Documents with Complex Layouts (RDCL-2015) competition dataset from ICDAR2015, and it belongs to PRImA Research Lab, University of Salford, UK (Antonacopoulos et al., 2015). This dataset contains 70 English-language color documents in a variety of layouts, and its focus is the diversity of the text regions and the complexity of the document layout.

The second dataset is the UNLV database (University of Nevada, Las Vegas, U.S.A.). The UNLV dataset contains a large variety of binary document images (more than 2000 scanned images), such as those from magazines, newspapers, and reports (Shafait & Smith, 2010) (ISRI-OCR, 2016). The authors' dataset was created from a selection of 50 images (named UNLV-A1) with complex layouts. The created dataset also contains images with a number of tables that are of a variety of structures. Different from the RDCL2015 dataset, the UNLV-A1 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.

To test the performance of the proposed system in a real application, 95 real-document images in different languages were scanned and collected. These images are categorized into the following two datasets: the CNU-Korean dataset contains 50 Korean-language color documents, and the CNU-English dataset contains 45 English-language color documents. These datasets include some skewed and noisy document images caused by the scanning process or low quality of the document, and some of the images include the local skew or distortion paragraphs.

With the use of the Athelia tool from the PRImA lab, the ground truths of the UNLV-A1, CNU-Korean, and CNU-English datasets were labeled manu-



(a) UNLV-A1 (b) CNU-English (c) CNU-Korean

Figure 13: Example of UNLV-A1, CNU-English, CNU-Korean datasets with its ground-truths.

ally in the same way that the RDCL-2015 dataset was made (Clausner et al., 2014). The details of these datasets can be found and downloaded in Tran (2016). An example of these datasets is shown in Figure 13.

4.2. Metric

The evaluation of the document layout analysis is always complicated because it depends on the database and its ground truth. As of this paper, numerous metrics have been proposed (Shafait et al., 2008) (Mao & Kanungo, 2002) (Clausner et al., 2011). The Scenario Driven In-Depth method (Clausner et al., 2011), for example, was verified as an efficient method, and it has been widely used in the performance evaluation regarding the document layout analysis. For this metric, five types of error (*merge, split, miss/partial-miss, miss-classification,* and *false positive*) are combined to extract the weights for the individual scenarios in the systemic performance. This metric was used in the performance evaluation for the HDLA2011, HDLA2013, and RDCL2015 competitions. In this paper, following six scenarios were obtained from the ICDAR competitions and used across all of the experiments: General Recognition, Segmentation, OCR, OCR*, Text, and Text* (Antonacopoulos et al., 2015).

The General Recognition (abbreviated to G-Recognition) scenario is the general recognition evaluation, in which all the errors and region types are

balanced weights (Clausner et al., 2011). The Segmentation scenario, as presented in Antonacopoulos et al. (2015), focuses on the miss and partial-miss errors while the miss-classification errors are ignored. Different from Segmentation scenario, the OCR scenario includes the evaluation of the miss-classification errors. For the Text scenario ("Text only"), only the text regions are considered. In the OCR and Text scenarios, however, the table region and the math expression, which is miss-classified as the text region, are not penalized. Also, any error in the miss-classification of the text-to-text region is ignored (e.g., a heading is miss-classified as a paragraph). The two other evaluation scenarios, OCR* and Text*, are therefore proposed to test the performance of the page segmentation systems.

Based on the OCR and Text evaluation scenarios, OCR* and Text* have been created so that all of the region miss-classifications are considered even though these regions are either a table, a math expression, or noise. These two scenarios are crucial because the use of the table and math expression are quite common in magazines and science journals.

4.3. Evaluation profiles

The proposed method has been tested on different databases and produced positive results, as can be seen in Appendix A.

First, to evaluate the effectiveness of the proposed system, the authors participated in the Recognition of Document with Complex Layouts (RDCL-2015) section of the ICDAR2015 competition. In this competition, the participants are required to submit an executable program, and the organizer tests it independently. The following eight systems were evaluated in this competition: Tesseract 3.02 (TO 3.02), Tesseract 3.03 (TO 3.03), ABBYY FineReader Engine 10 and 11 (FRE 10, FRE 11), PAL (winner HDLA2013 competition), ISPL, Fraunhofer (winner of the ICDAR2009 page segmentation competition), and the proposed system, MHS. Because the errors in the detection of table and math expressions are not considered in this competition, the following three scenarios are used: Segmentation, OCR, and Text Only. The final evaluation profiles of this contest are presented in Figure 14. According to the announcement from the organizer, the proposed system (named "MHS - Multilevel Homogeneity Structure") achieved the highest result - winner (Antonacopoulos et al., 2015). Figure 14 also includes the results of the two systems, the AOSM system (Ha et al., 2016) and the updated MHS version MHS-2016. Even though these methods did not participate in

the RDCL-2015 competition, their performances were also evaluated by the competition organizer, the PRImA lab.

Second, in terms of the UNLV-A1 dataset, Figure 15 shows the evaluation of the MHS-2016 system alongside those of the commercial system ABBYY FineReader 12 Professional (FRP-12), TO-3.03, PAL, ISPL, and AOSM. All the six scenarios were used to test the performance of each method for this dataset.

Lastly, in terms of the two real datasets (CNU-English and CNU-Korean), the segmentation results of the proposed system and its comparison with AOSM, TO-3.03, PAL, and FRP-12 are presented in Figure 16 and Figure 17.

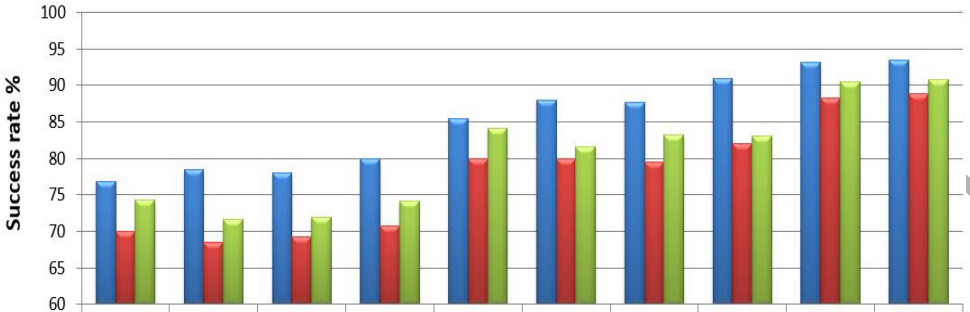
For each dataset, the layout analysis errors that are calculated in the OCR scenario (Antonacopoulos et al., 2015) are also presented for an analysis of the errors (*merge, split, miss/partial miss, misclassification, false detection*).

One of the essential assessments is the comparison of processing times, although that evaluation is difficult because most of the algorithms do not provide their computing time. Also, the algorithm timings cannot be directly compared because of the differences between the computing environments. In this study, the computing time is also considered as a critical metric. Because any image resolution can be used for the running of the proposed system, an image with a high resolution will be re-sized to reduce the time consumption. The experimental results show that the success rate of the proposed algorithm is still highly precise for those images with a resolution that is greater than or equal to 3.0 MP; here, the success rate is similar to that of the original image. The average processing time of the proposed system on Visual C++ is 3.46 sec, which includes 2.49 sec for the page segmentation and 0.97 sec for the table detection.

4.4. Discussion and analysis

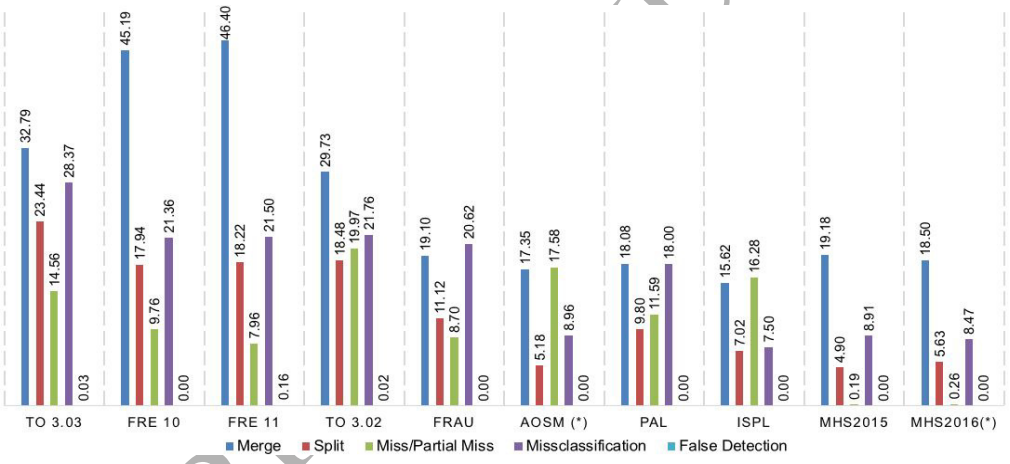
4.4.1. Performance analysis

All the evaluations in this paper, conducted on a variety of datasets, show that the proposed method yields a superior performance. The proposed system can work with many different document languages, and its robustness has been proved on various datasets. According to the analysis of the RDCL-2015 competition, the proposed system still outperforms the other methods even when the primary focus is the text, or when the non-text regions are ignored (Antonacopoulos et al., 2015).



	TO 3.03	FRE 10	FRE 11	TO 3.02	Fraunhofer	AOSM (*)	PAL	ISPL	MHS - RDCL2015	MHS - 2016 (*)
Text	76.8	78.4	78	79.9	85.4	88	87.7	91	93.1	93.4
OCR	70	68.5	69.3	70.8	79.9	79.9	79.5	82	88.3	88.8
Segmentation	74.3	71.6	72	74.1	84.1	81.6	83.2	83.1	90.5	90.8

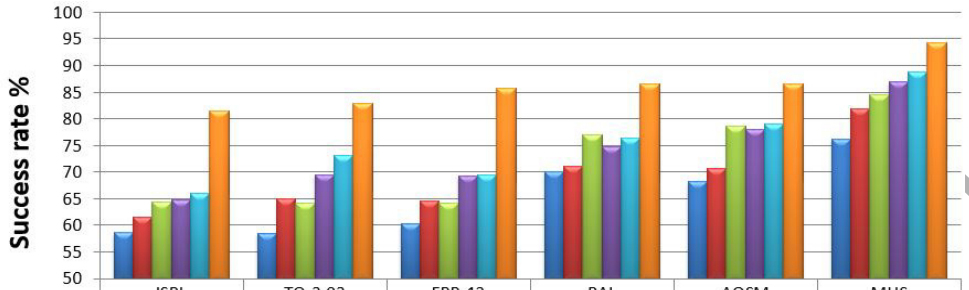
(a) Success rates of the ten methods with the Segmentation, OCR, Text scenarios.



(b) Breakdown errors of the layout analysis based on the OCR scenario.

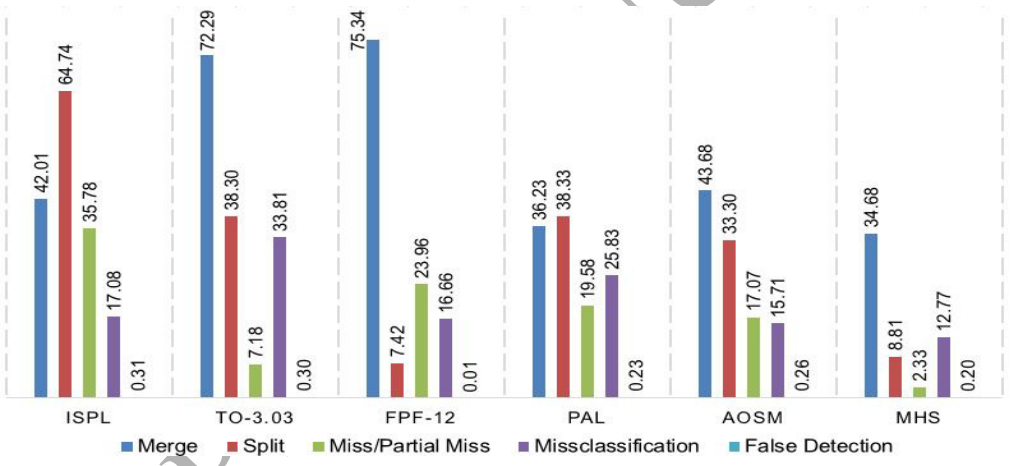
Figure 14: Performance evaluation of the eight participants in the RDCL-2015 competition, and another 2 systems(*) proposed in 2016.

The ISPL, PAL, AOSM and Fraunhofer methods exhibited positive performances for the RDCL-2015, and this is especially the case for the two runner-up systems, ISPL and PAL. The difference between their performances and that of the proposed method for this dataset is small (approximately 2% to 6%). However, the gap is increased significantly for the real-scan images or the binary images, such as UNLV-A1, CNU-English, and CNU-Korean datasets (approximate 6% to 20%). The errors in the



	ISPL	TO-3.03	FRP-12	PAL	AOSM	MHS
Text*	58.65	58.47	60.25	70.13	68.31	76.17
OCR*	61.47	64.93	64.51	71.15	70.76	81.84
Text	64.38	64.11	64.11	76.94	78.68	84.57
OCR	64.68	69.54	69.33	74.73	77.91	86.86
Segmentation	65.93	73.06	69.43	76.43	78.94	88.75
G-Recognition	81.52	82.98	85.73	86.5	86.58	94.3

(a) Success rate for the 6 scenarios.

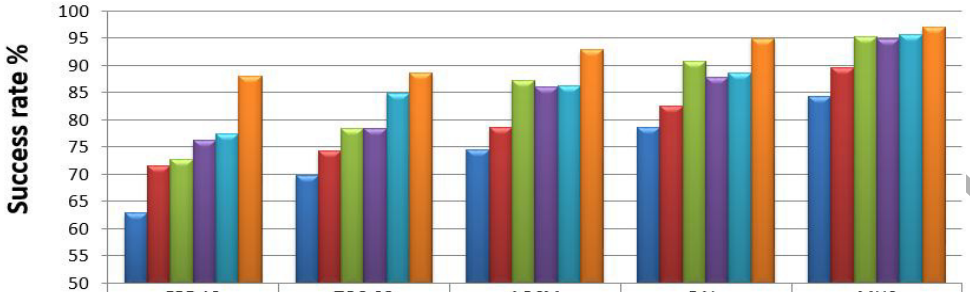


(b) Breakdown errors of the layout analysis based on the OCR scenario.

Figure 15: Evaluation profile for the six systems on the UNLV-A1 dataset.

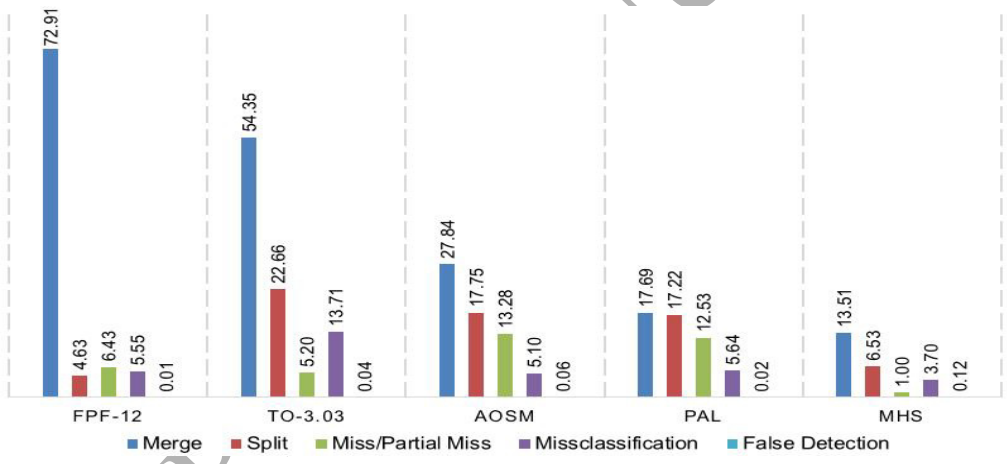
miss/partial-miss of these methods are increased with the images (binary image) that contain a lot of noise or complex text regions, see Figure 15. The efficiency in the classification of the text and non-text elements of the ISPL, PAL, and AOSM methods is significantly compromised as can be seen in Appendix A-Figure 18.

Regarding the commercial systems of FRE 10 and 11 and FRP-12, the partial-miss and merge errors happen frequently in the text regions. In FRP-



	FPF-12	TO3.03	AOSM	PAL	MHS
Text*	62.83	69.7	74.46	78.59	84.3
OCR*	71.52	74.32	78.54	82.46	89.54
Text	72.65	78.48	87.17	90.75	95.23
OCR	76.25	78.32	86.03	87.93	94.84
Segmentation	77.42	84.97	86.17	88.69	95.63
G-Recognition	88.03	88.53	92.89	94.95	97.06

(a) Success rate of each method for the 6 scenarios.

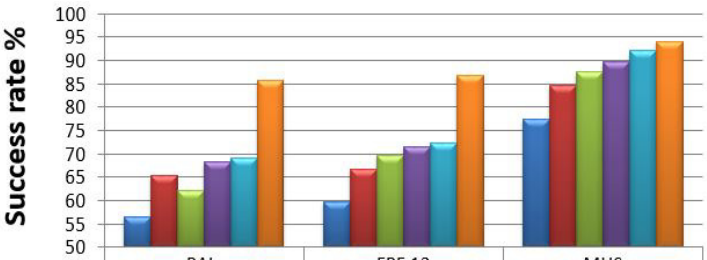


(b) Breakdown errors of the layout analysis based on the OCR scenario of each method.

Figure 16: Evaluation profile for the five systems on CNU-English dataset.

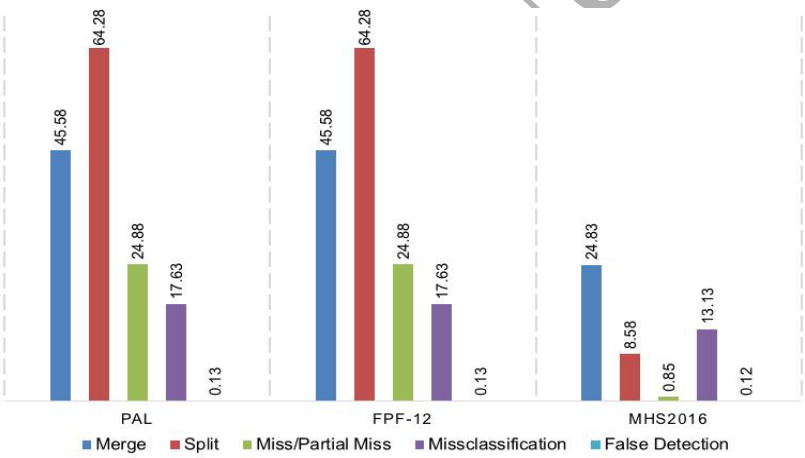
12 system, the separator identification and text region labeling (all of the text regions share the same label) are not considered, while the image detection causes miss-classification errors in the text region. With Tesseract OCR systems (3.02 and 3.03), however, a lot of noise occur frequently during the detection and classification of the zones, even though the regions are labeled and segmented.

Although the proposed system achieves a high success rate for the seg-



	PAL	FPF 12	MHS
Text*	56.57	59.79	77.55
OCR*	65.35	66.8	84.7
Text	62.1	69.75	87.6
OCR	68.29	71.4	89.75
Segmentation	69.25	72.46	92.09
G-Recognition	85.8	86.95	94.11

(a) Success rate for the 6 scenarios.



(b) Breakdown errors of the layout analysis based on the OCR scenario.

Figure 17: Evaluation profile for the three systems on CNU-Korean dataset. With this dataset, the language mode of FRP-12 was changed to Hangul (Korean language).

mentation and the recognition, some limitations remain. As is presented in Figure 14(b), Figure 15(b), Figure 16(b), and Figure 17(b), the proposed system is clearly advantageous in terms of the regions that are not missing in the document image, while the merge is the primary source of errors. Most of these errors due to the mistakes in the merged image components that are of a large size, as can be seen in Appendix A-Figure 20(b) (two bicy-

cles) or Figure 12(b) (three top images). This mistake, however, does not cause a major problem in the reading and understanding of the document. The results also show that the merge error of the proposed method in the RDCL-2015 dataset is slightly higher than those of ISPL (3%), PAL (0.5%), and AOSM (1.15%) methods. However, the mistakes of the proposed system are significantly less than those of other methods, in particular with the miss/partial miss (8% to 17%). It should be noted that, with UNLV-A1, CNU-English, and CNU-Korean datasets, the number of merge errors in the proposed system is significantly less than those of the other methods.

The binarization quality can be a cause of error in the proposed system; furthermore, the negative-text regions are not controlled in some difficult cases (e.g., in complex image regions), and the non-text element of a small size is not identified accurately. The combination of the text line and homogeneous region extraction can reduce the merging errors between two adjacent text zones, but the errors can still happen between two 'narrow' text lines. The labeling method based on the size of the text element and its arrangement of the text region is still heuristic. As already known, the text regions in the document are not only segmented and labeled according to their position, but also according to their meaning; but without the OCR process, this is a tough task. In the future work, the learning approach will be considered to overcome this problem.

The accuracy for UNLV-A1 dataset, for all the methods, is unsatisfactory (e.g., the vertical-heading regions are often labeled wrong). Although the least standard deviation in the RDCL-2015 competition was attributed to the proposed method, the standard deviation in all of the evaluation scenarios regarding the proposed method for the UNLV-A1 and CNU-Korean datasets did not match the expectations. This fact shows that the proposed system still needs to be improved to reduce the errors; for example, the negative-text regions in a complex image region and the labeling of the Korean language should be reconsidered.

In fact, we covered a broad range of page segmentation systems. In other words, we compared the performance of our system with many other methods. For instance, Ha et al. (2016) compared their AOSM method with many older methods such as Voronoi (Kise et al., 1998), Docstrum (O'Gorman, 1993), Tab-stop (Smith, 2009) on the ICDAR2009 dataset. Although the performance of AOSM is better than the others, there is a big gap between our system and AOSM. As we can see in our evaluation profiles, the accuracy of AOSM always less than our method in all datasets.

Table 1: Area-ratio-based evaluation profiles (Shafait & Smith, 2010) for table detection of the MHS-2016, FRP-12, and TO-3.03.

Dataset	Method	No. T	Correct	Partial	Over	Under	False	Miss	Precision	Recall
RDCL2015	MHS	3	3	0	0	0	0	0	99.9	99.9
UNLV-A1	MHS	48	45	0	1	2	0	0	98.7	98.6
	FRP		30	4	2	8	0	4	93.7	85.2
	TO		23	11	4	8	1	2	88.6	85.6
CNU-E	MHS	13	12	0	0	0	0	1	97.5	91
	FRP		10	0	0	2	1	1	97.7	89.3
	TO		5	4	0	0	4	4	66.5	74
CNU-K	MHS	31	31	0	0	0	0	1	97.9	99.9
	FRP		29	1	0	0	1	7	88	96.5

Table 2: Performance comparison of the MHS-2016 system and the version that was proposed by Tran et al. (2016b).

		Success rate				
Dataset	Method	Segmentation	OCR	Text	OCR*	Text*
ICDAR2009	MHS-2016	92.01	89.93	94.66	82.19	84.83
	Tran et al. (2016b)	83.26	79.76	81.32	72.4	73.56
UNLV-A1	MHS-2016	88.75	86.86	84.57	81.84	76.17
	Tran et al. (2016b)	78.53	75.16	70.92	69.79	62.12
CNU-Korean	MHS-2016	92.09	89.75	87.6	84.7	77.55
	Tran et al. (2016b)	74.15	72.36	67.48	65.54	57.32

It should be noted that, except for ABBYY FindReader and Tesseract OCR, all the existing algorithms do not provide methods for the detection of table regions or mathematics expression. For this work, the proposed system also yields a superior performance. This is proved by the evaluation profile presented in Table 1.

4.4.2. System analysis

The primary goal of this article is to present a complete system for the whole document layout analysis process. The first version of our system is presented in Tran et al. (2016b). That paper mostly focuses on the MHA algorithm regarding the text and non-text classification in English and Korean document. Therein, the MHA algorithm is one of the most important processes of our system. Even though the initial classification result is positive, the MHA algorithm at that time is not stable as we expected. The heuristic filter still causes some errors for noisy documents or diacritic language docu-

ment, whereas the multilevel classification is not strong enough. Other parts of the system such as layout text segmentation step often cause the merge errors (including non-allowable merge errors) and split errors. The success rate of the OCR evaluation is less than 80%, while the segmentation success rate is not more than 84% for the three datasets in Table 2. Therefore, an overall upgraded version, the MHS-2016, is presented in this paper. It is the latest and complete version in which all of the strengths have been included, and the weaknesses of the previous versions have been eliminated. The experimental results showed this improvement in our MHS system. As we can see in Table 2, the success rate of every scenario always higher than the old version from 9% to 20%.

As mentioned previously, two notable updates are in the classification and in the segmentation processes. In the classification step, the heuristic filter is refined to work well with English/Vietnamese and Korean/Chinese languages. Through the completion of many experiments, the performance of the new heuristic filter in this paper is stable with both Korean and English documents. It is also previously mentioned that, due to the role of this filter, it can be removed without sacrificing the accuracy in the classification of text and non-text components. The time consumption, however, is increased because the MLL classification based on the recursive filter requires more processing time, especially for large connected components. Besides, the efficiency of the heuristic filter relating to the removal of tiny components (noise) was also proved.

As mentioned in Section 3.2.1, an overall improvement of the text region clustering in the segmentation process has been achieved. Instead of the use of horizontal projection, a white-space analysis is utilized for the extraction of homogeneous regions, thereby reducing the errors that can be easily caused by the projection (see Figure 6). This improvement is critical because, in real applications, the input document is often skewed (small angle) or noise, or it contains distorted paragraphs. The combination of text line extraction and mathematical morphology in text region segmentation also facilitates the deduction of more effective results. The use of paragraph segmentation in text segmentation helps us to reduce the merge errors, whereas the use of text line extraction reduces the split errors. This effectiveness is achieved not only for English language documents but also for Korean language documents.

An essential upgrade of the MHS-2016 system is the table detection step. It should be noted that the system used for the RDCL-2015 competition is essentially an MHS-2016 system for which the table detection step had not

been embedded; furthermore, some minor programming errors have been fixed in the current version.

Like ABBYY FineReader system, the proposed method also allows the user to select the language of the input document (default is English). Due to the differences in the character structures, this selection can improve the performance of the system. In future work, the heuristic filter of the classification process will be considered for the design process for the two main language types, Latin (English, Vietnamese, etc.) and non-Latin (Korean, Chinese, Arabic, etc.). Also, it can be designed for each type of document language. The dependency of the proposed system on the document language is slight, most of the system processes (92%) are fixed for all languages. Due to the diversity in language structure, the difference mainly located in the segmentation process (e.g., paragraph segmentation).

Different from the two runner-up participants in the RDCL-2015, the training process is not required for the proposed system. In the future, however, machine learning can be considered to overcome the problems with the labeling and the classification of the language in the input document.

5. Conclusions

This paper addressed a whole system for the document layout analysis. The proposed system is designed for consistency with the following three main stages: a classification of text and non-text elements via multilevel/multi-layer homogeneity structure; a text segmentation via a text line extraction and mathematical morphology, and a non-text segmentation with the RB for table detection; a region refinement via a rectangular extraction and noise removal, and the implementation of a labeling method via the heuristic rules based on the region's position and size. As can be seen in the experiment results, the MHS-2016 system is stable and more efficient than not only the other page segmentation systems but also the older versions of the proposed system.

Compared with the other page segmentation methods, the proposed method provides four major contributions:

First, one of the major differences between MHS and existing systems is the ability to work with different document languages; not only English documents but also non-Latin documents (Korean). The experiment shows that our system always achieves a higher success rate than other systems with any document language. The proposed system is designed to work with

document images in various languages whereas the other systems almost cannot (or tough) to do that. One of the reasons is the strict dependency on the document language in the classification process.

Second, using a new technique, called MLL classification, MHS shows a higher performance in classifying text and non-text elements with various document languages. By combining Multilevel and Multi-layer classification as well as by reducing the influence of the heuristic filter, the new version of MHA algorithm in our system proved the robustness in identifying text and non-text element in many different document languages. As mentioned above, this is the first and most important process of any page segmentation method, but all the other page segmentation algorithms do not focus enough on this problem. They often fail in those documents having a complex layout (non-Manhattan).

Third, a new technique using a combination of text line extraction and mathematical morphology is proposed for text segmentation. Based on this method, we can overcome the weak points of traditional methods in text clustering, such as the errors cause by projection in the skew document or distortion region, errors in merging text elements, etc.

Fourth, the proposed system contains a robust method for table detection. Except for our system, there is no page segmentation method which comprises a robust table detection process. The technical limitations in their approaches may be a reason.

The focus of future improvement of MHS-2016 is on the errors that were observed in the experiment. First, the expansion of the system so that it can work with Arabic, Vietnamese, and other particular document language is an important consideration. The page segmentation on documents with multi-language and random text directions (vertical, diagonal, etc.) is also an important task, while the text extraction from complex layouts should be considered to improve the performance regarding those document images with complex backgrounds.

The author's goal is to complete a robust system that can be used for the extraction of the information from a document image. Even though the proposed system can be applied to the OCR process on a scanning device or a computer system, the mobile-device application still requires some improvement, e.g., estimation and correction of perspective and non-linear document distortions.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A3A01018993).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version.

References

Agrawal, M., & Doermann, D. S. (2009). Voronoi++ A dynamic page segmentation approach based on Voronoi and Docstrum features. *Proceedings of the* 10^{th} *ICDAR*, 1011-1015.

Antonacopoulos, A., Pletschacher, S., Bridson, D. & Papadopoulos, C. (2009). ICDAR2009 page segmentation competition. *Proceedings of the* 10^{th} *ICDAR*, 1370-1374.

Antonacopoulos, A., Bridson, D., Papadopoulos, C. & Pletschacher, A., (2009). A Realistic Dataset for Performance Evaluation of Document Layout Analysis. *Proceedings of the* 10^{th} *ICDAR*, 296-300.

Antonacopoulos, A., Clausner, C., Papadopoulos, C. & Pletschacher, S., (2015). ICDAR2015 Competition on Recognition of Documents with Complex Layouts - RDCL2015, *Proc. of the* 13^{th} *ICDAR*, 1151-1155.

Baird, H., Jones, S. & Fortune, S. (1990). Image segmentation by shape-directed covers. *Proceedings of the ICPR*, 820-825.

Bukhari, S. S., Al Azawi, A., Shafait, F., & Breuel, T. M. (2010). Document image segmentation using discriminative learning over connected components. *Proceeding of 8th IAPR International Workshop on Document Analysis Systems*, 183-190.

Bukhari, S. S., (2011). Improved document image segmentation algorithm using multiresolution morphology. *Document Recognition and Retrieval XVIII*, SPIE, 7874, 1-10.

Caponetti L., Castiello, C., & Gorecki P. (2008). Document page segmentation using neuro-fuzzy approach. *Applied Soft Computing*, 8, 118-126.

Chen, K., Yin, F. & Liu, C.-L. (2013). Hybrid Page Segmentation with Efficient Whitespace Rectangles Extraction and Grouping. *Proceedings of the* 12^{th} *ICDAR*. 958-962.

Cheng, H. & Bouman, C.A. (2001). Multi-scale Bayesian Segmentation Using a Trainable Context Model. *IEEE Transactions on Image Processing*, 10(4), 511-525.

Cinque, L., Lombardi, L., & Manzini G. (1998). A multiresolution approach for page segmentation. *Pattern Recognition Letters*, 19, 217-225.

Clausner, C., Pletschacher, S., & Antonacopoulos, A. (2011). Scenario driven in-depth performance evaluation of document layout analysis methods. *Proceedings of the* 11^{th} *ICDAR*, 1404-1408.

Clausner, C., Pletschacher, S., & Antonacopoulos, A., (2014). Efficient ocr training data generation with Aletheia. *Proceedings of the 11th Workshop on Document Analysis System*, ACM 178.

DIOTEK Mobile Software Development Company, http://www.diotek.com/eng/

Ferilli, S., Basile, T.M.A, & Esposito, F. (2010). A histogram based technique for automatic threshold assessment in a run length smoothing-based algorithm. *Proceedings of the DAS*, ACM 349-356.

Gioi, R.-G. V., Jakubowicz, J., Morel, J.-M., & Randall, G., (2010). LSD: A fast line segment detector with a false detection control. *IEEE PAMI*, 32(4), 722-732.

Ha, J., Haralick, R.M. & Phillips, I.T. (1995). Recursive X-Y Cut Using Bounding Boxes of Connected Components. *Proceedings of the* 3^{rd} *ICDAR*, 952-955.

Ha, D-T., Nguyen, D-D., & Le, D-H. (2016). An adaptive over-split and merge algorithm for page segmentation. *Pattern Recognition Letters*, 80, 137-143.

Isri ocr evaluation tools, https://code.google.com/p/isri-ocr-evaluation-tools/downloads/list

Jain, A.K., & Yu, B. (1998). Document Representation and Its Application to Page Decomposition. *IEEE PAMI*, 20(3), 294-308.

Kise, K., Sato, A., & Iwata, M. (1998). Segmentation of page images using the area Voronoi diagram. *Computer Vision Image Understanding*, 70(3), 370-382.

Koo, H. I. & Kim, D.H., (2013). Scene text detection via connected component clustering and non-text filtering. *IEEE Transactions on Image Processing*, 22, 2296-2305.

Lee, S.-W. & Ryu, D.-S. (2001). Parameter - Free Geometric Document Layout Analysis. *IEEE PAMI*, 23(11), 1240-1256.

Mao, S., & Kanungo, T., (2001). Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms, *IEEE PAMI*, 23(3), 242-256.

Mao, S., & Kanungo, T., (2002). Software architecture of pset: a page segmentation evaluation toolkit. *Int. J. Doc. Anal. Recogn.*, 4, 205217.

Nagy, G., Seth, S. & Viswanathan, M. (1992). A prototype document image analysis system for technical journals. *Computer*, 25(7), 1022.

O'Gorman, L. (1993). The document spectrum for page layout analysis. *IEEE PAMI*, 15(11), 1162-1173.

Okamoto, M., & Takahashi, M. (1993). A hybrid page segmentation method. *Proceedings of the* 2^{nd} *ICDAR*, 743-746.

Pan, Y., Zhao, Q. & Kamata, S. (2010). Document Layout Analysis & Reading Order Determination for a Reading Robot. *Tencon2010-2010 IEEE Region* 10^{th} *Conference*, 1607-1612.

Papamdreou, A., & Gatos, B.,(2011). A Novel Skew Detection Technique Based on Vertical Projections. *Proceedings of the* 11^{th} *ICDAR*, 384-388.

Sebastien, E., Petra, G.-K., & Ogier, J.-M.,(2017). A comprehensive survey of mostly textual document segmentation algorithms since 2008. *Pattern Recognition*, 64, 1-14.

Shafait, F., Keysers, D. & Breuel, T.M. (2008). Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithm. *IEEE PAMI*, 30(6), 941-954.

Shafait, F., & Smith, R., (2010). Table detection in Heterogeneous Documents. *Proc. of the 9th IAPR DAS*, 65-72.

Simon, A., Pret, J.C., & Peter Johnson, A. (1997). A fast algorithm for bottom-up document layout analysis. *IEEE PAMI*, 19(3), 273-277.

Smith, R. (2009). Hybrid Page Layout Analysis via Tab-Stop Detection. *Proceedings of the* 10^{th} *ICDAR*, 241-245.

Sun, H.M. (2005). Page segmentation for Manhattan and non-Manhattan layout documents via selective CRLA. *Proceedings of the* 8^{th} *ICDAR*, 116-120.

Tran, T. A., Na, I. S., & Kim, S. H. (2015). Separation of Text and Non-text in Document Layout Analysis using a Recursive Filter. *KSII Transaction on Internet and Information Systems*, 9, 4072-4091.

Tran, T.A., Tran, H.T., Na, I.S., Lee, G. S., Yang, H. J., & Kim, S. H, (2016). A mixture model using Random Rotation Bounding Box to detect table region in document image. *International Journal of Visual Communication and Image Representation*, 39, 196-208.

Tran, T. A., Na, I. S., & Kim, S. H. (2016). Page Segmentation using Minimum Homogeneity Algorithm and Adaptive Mathematical Morphology, *Int. Jour. Doc. Ana. Recog.*, 19(3), 191-209.

Tran, T. A., The multi-language datasets for document layout analysis. *CNU dataset*, (2016). Download here.

Wahl, F.M., & Wong, K.Y. & Casey, R.G. (1982). Block segmentation and text extraction in mixed text/image documents. *Graphical Models and Image Processing*, 20(4), 375-390.

Zitnick, C., & Dollr, P., (2014). Edge boxes: Locating object proposals from edges, *Computer Vision ECCV 2014*, 391-405.