Goal-Oriented Performance Evaluation Methodology for Page Segmentation Techniques

Nikolaos Stamatopoulos, Georgios Louloudis and Basilis Gatos
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-153 10 Agia Paraskevi, Athens, Greece
{nstam, louloud, bgat}@iit.demokritos.gr

***Abstract*—Document image segmentation is a fundamental step in the document image analysis pipeline as it affects the accuracy of subsequent processing steps. An objective and realistic evaluation of page segmentation techniques is crucial for a quantitative comparison among them. In this paper, a goal-oriented performance evaluation methodology that calculates a comprehensive evaluation measure *SR* (Success Rate) is presented. *SR* measure reflects the entire performance of a page segmentation technique in a concise quantitative manner. It is a pixel-based approach which avoids the dependence on a strictly defined ground-truth. The proposed evaluation measure *SR* deals only with text regions and is correlated with the percentage of the text information in which the subsequent processing (e.g. text line segmentation and recognition) can be applied successfully.**

Keywords—page segmentation; performance evaluation; performance metric; document image analysis

I. INTRODUCTION

Page segmentation is a crucial processing step in a document image analysis system. It is the process of identifying the areas of interest in a document page image [1 – 3]. The performance of subsequent processing such as text line segmentation and optical character recognition (OCR) heavily depends on the accuracy of page segmentation techniques.

The automatic evaluation of page segmentation algorithms is an important issue both for quantitative comparisons among different techniques as well as for qualitative analysis of segmentation results. In this paper, a goal-oriented performance evaluation methodology is proposed that reflects the percentage of the text information in which the subsequent processing, such as text line segmentation and recognition, can be applied successfully. It is a pixel-based approach which deals only with text regions. Moreover, the proposed evaluation technique avoids the dependence on a strictly defined ground-truth since the ground-truth for page segmentation is quite ambiguous and may differ between users.

The remainder of the paper is organized as follows. In Section II the related work is discussed. Section III focuses on the proposed performance evaluation methodology. The advantages of the proposed method are discussed in Section IV while conclusions are drawn in Section V.

II. RELATED WORK

Several page segmentation competitions [4, 5] have been organized in order to address the need of comparative performance evaluation under realistic circumstances. The performance analysis method used for these competitions is based on a geometric approach using polygon region outlines [6]. The ground-truth creation for such approaches is quite ambiguous. Kanai et al. [7] use an indirect evaluation based on OCR results. The advantage of this method is that it requires only transcription ground-truth and, hence, does not require defining ground-truth regions. However, it cannot give an accurate indication of page segmentation performance as it is dependent on the OCR engine. In [8], Mao and Kanungo propose a textline based performance metric that examines geometric correspondences of text lines. The main drawback of this method is that it requires ground-truth at text line level and it deals only with deskewed document images. Liang et al. [9] describe a region area based metric in which different weights are assigned to each type of matching (one-to-one, many-to-one, etc.). In a similar way, Shafait et al. [10] use a weight bipartite graph called pixel-correspondence graph [11] in order to calculate the total number of over-segmented and under-segmented regions as well as the missed regions and false alarms. In [12], the evaluation method is based on a set of simple rules concerning the main body text regions, the auxiliary text regions and the non-text regions. Finally, Agrawal et al. [13] consider a result region as correctly detected if its foreground pixels overlap with those of ground-truth above a user specified threshold. All the above mentioned performance evaluation methods are highly dependent on the ground-truth which should be strictly defined. The proposed evaluation framework avoids the dependence on a strictly defined ground-truth and it is based on simple and clear guidelines given to the users.

III. PERFORMANCE EVALUATION METHODOLOGY

A detailed description of the distinct stages of the proposed evaluation methodology is presented in this section. First, an overview of ground-truth requirements and related issues is given and then, the proposed performance metric is presented. The proposed evaluation methodology deals only with text regions and it requires the binary version of the document image since it is a pixel-based approach.

978-1-4799-1805-8/15/$31.00 ©2015 IEEE

A. Ground-thruth creation

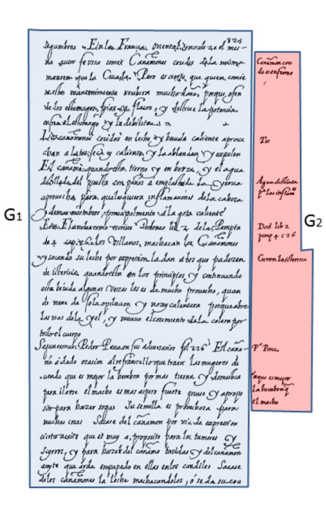
The first step for the performance evaluation of a page segmentation algorithm is the ground-truth creation. However, ground-truth is quite ambiguous and may differ between users. At the proposed evaluation framework, the ground-truth creation is based on two very simple and clear guidelines for the users. Our goal is to create ground-truth regions in which the subsequent text line segmentation stage can be applied successfully. Different ways of ground-truthing, for example a text column marked as one region or as separate paragraphs, do not affect the proposed evaluation metric.

Ground-truth text regions are represented by polygons. Let B be a binary document image and $P(G) = \{G_i, i = 1,2,\cdots,\#P(G)\}$ be a set of ground-truth polygons, where # denotes the cardinality of a set. Each ground-truth text region G_i should be consistent with the following two guidelines:

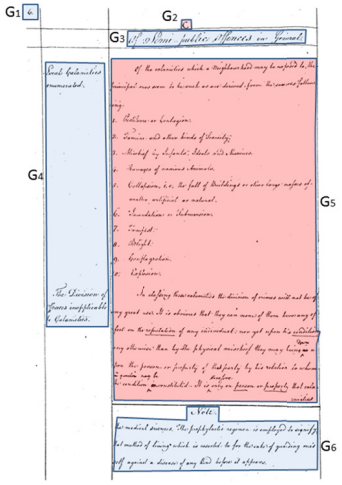
1. It should not contain text lines with horizontal overlap (e.g. text lines of different columns or marginal notes).

2. It should not contain non-text elements (separator lines, drawings, images etc.).

If a text region follows the above mentioned guidelines, then the subsequent processing such as text line segmentation, can be applied successfully. Figure 1 depicts document images with the corresponding acceptable ground-truth regions while Figure 2 presents examples of ground-truth regions that are not consistent with the abovementioned guidelines.

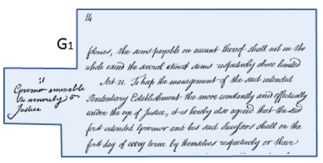


(a)

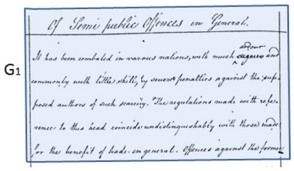


(b)

Fig. 1. Document images with the corresponding acceptable ground-truth regions.



(a)



(b)

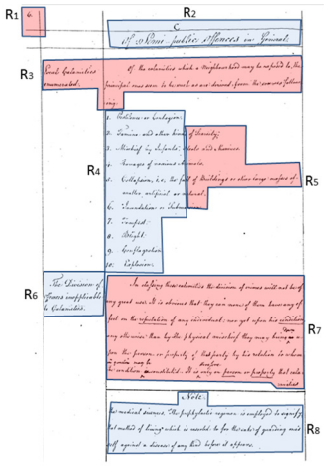
Fig. 2. Examples of ground-truth regions which are not consistent with the guidelines. The ground-truth region contains: (a) text lines of different columns with horizontal overlap, (b) separator lines.

B. Performance Metric

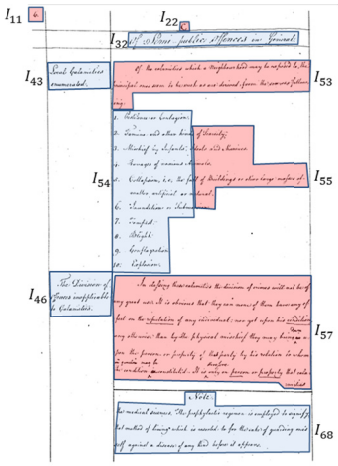
Let $P(R) = \{R_j, j = 1,2,\cdots,\#P(R)\}$ be a set of polygons produced by an automatic page segmentation algorithm. We define the set of intersection regions $P(I)$ of the ground-truth and the segmentation result as follows:

$$P(I) = \left\{ I_{ij} \middle| \begin{array}{l} I_{ij} = G_i \cap R_j, if\ G_i \cap R_j \neq \emptyset\ and\ \dfrac{F(I_{ij})}{F(R_j)} > Th \\ I_{ij} = \emptyset, \qquad otherwise \end{array} \right\} \quad (1)$$

where $F(\cdot)$ a function which counts the foreground pixels of a region. The condition of Eq. (1) assures that the overlap between a ground-truth and a result region is significant. In our experiments, we set the threshold Th equal to 0.01. A page segmentation result of the document image shown in Fig. 1(b) as well as the corresponding intersection regions are presented in Fig. 3.



(a)



(b)

Fig. 3. (a) A page segmentation result of the document image shown in Fig. 1(b), (b) the corresponding intersection regions.

We define the overall quantitative evaluation measure SR (Success Rate) as follows:

$$SR = \frac{\sum_{i=1}^{\#P(G)} \sum_{j=1}^{\#P(R)} w_{ij} \times F(I_{ij})}{\sum_{i=1}^{\#P(G)} F(G_i)} \times 100 \quad (2)$$

where w_{ij} corresponds to a weight for each intersection region I_{ij} ranging in the interval [0,…,1]. As it can be observed, the maximum value of the numerator is the sum of the foreground pixels of all intersection regions (in the case that all weights are equal to one) and the denominator represents all the foreground pixels of the ground-truth. The proposed evaluation measure SR ranges in the interval [0,…,100] and the higher the value of the SR, the better is the performance of the page segmentation algorithm.

In the sequel, we define the corresponding weight w_{ij} for each of the following conditions: (i) the ground-truth region G_i has been detected correctly, (ii) the ground-truth region G_i has been split, (iii) the result region R_j has been overlapped by two or more ground-truth regions (merge) and finally (iv) non-text elements have been included in the result region R_j. If more than one condition is satisfied, the weight with the smaller value is selected.

(i) Correct Detection:

When the ground-truth region G_i is overlapped completely by the result region R_j and vice versa ($G_i \cap B = R_j \cap B = I_{ij} \cap B$, where B is the binary image) this means that the given region is correctly detected in the segmentation result. In this case the corresponding weight w_{ij} is equal to one, so all the foreground pixels of ground-truth region G_i are considered as correctly detected. An example of a correctly detected region is presented in Fig. 1(b) and Fig. 3 for correlating G_6 and R_8 ($w_{68} = 1$).

(ii) Split ground-truth region:

In the case that the ground-truth region G_i is overlapped by two or more result regions, it is consider as split. We check if the corresponding intersection regions have horizontal overlap and treat each case accordingly.

- Splitting without having horizontal overlap

If the intersection region I_{ij} does not overlap horizontally with any other intersection region $I_{ij'}$, produced by the same ground-truth region G_i, we set the corresponding weight w_{ij} equal to one. The text lines of this region have not been split; as a result, they can be detected correctly in the subsequent processing steps. An example of this case is presented in Fig. 4.

- Splitting having horizontal overlap

In the case that the intersection region I_{ij} overlaps horizontally with one or more regions $I_{ij'}$, some text lines of this region may have been split. As a result, the subsequent text line segmentation stage will not be able to detect correctly these text lines. Our goal is not to reject all the foreground pixels but to detect the subregions of I_{ij} that do not overlap horizontally with other regions (the text lines which have not been split) in order to consider the foreground pixels of them as correctly detected (see Fig. 5).

First, we define the set of subregions of the region I_{ij} without horizontal overlap as follows:

$$P^{HO}(I_{ij}) = \{H_q^{ij}, q = 1,2,\dots,\#P^{HO}(I_{ij}) \mid H_q^{ij} \subset I_{ij} \; H_q^{ij} \neq H_{q'}^{ij}, \forall q' \neq q \; H_q^{ij} \; doesn't \; overlap \; horizontally \; \forall I_{ij'} \; j \neq j' \} \quad (3)$$

Figure 5 depicts the subregions of the region I_{54} of the example presented in Fig. 3(b). As it can be observed, these subregions include five text lines, which are considered as correctly detected since they can be detected in the subsequent text line segmentation stage.

The corresponding weight w_{ij} of the region I_{ij} with horizontal overlap can be defined as the ratio of the foreground pixels of all subregions without horizontal overlap over the total foreground pixels of the region as follows:

$$w_{ij} = \frac{\sum_{q=1}^{\#P^{HO}(I_{ij})} F(H_q^{ij})}{F(I_{ij})} \quad (4)$$

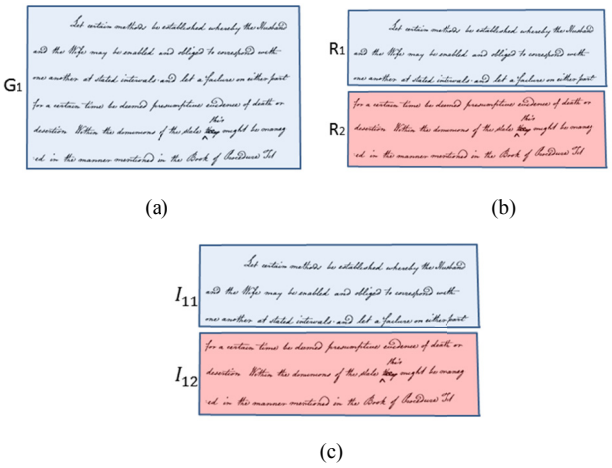


Fig. 4. An example of a ground-truth region that has been split into regions without horizontal overlap. (a) Ground-truth region (b) result regions (c) intersection regions without horizontal overlap ($w_{11} = 1$, $w_{12} = 1$). All the foreground pixels of ground-truth region G_1 are considered as correctly detected.

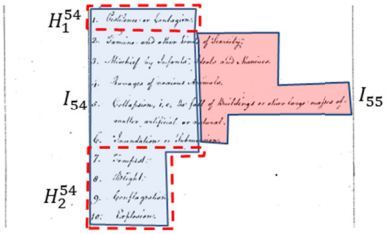


Fig. 5. Subregions H_1^{54}, H_2^{54} (dashed line) of the region I_{54} without horizontal overlap of the example presented in Fig. 3(b). The first text line as well as the four last text lines are considered as correctly detected since they can be detected in the subsequent text line segmentation stage.

(iii) Merged ground-truth regions:

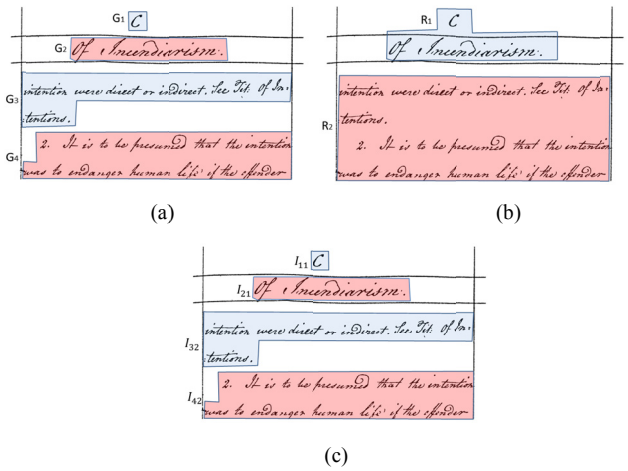
When the result region R_j is overlapped by two or more ground-truth regions it means that these ground-truth regions have been merged in the page segmentation result. We check if the corresponding intersection regions have horizontal overlap and treat each case accordingly.

- Merging without having horizontal overlap

In the case that the intersection region I_{ij} does not overlap horizontally with any other region $I_{i'j}$, produced by the same result region R_j, non-text elements may have been included at the result region (see Fig. 6). We set a penalty for the corresponding region according to the percentage of non-text foreground pixels. The corresponding weight w_{ij} is defined as follows:

$$w_{ij} = \frac{F(I_{ij})}{F(R_j) - \sum_{\substack{i'=1 \\ i' \neq i}}^{\#P(G)} F(I_{i'j})} \quad (5)$$

As it can be observed, the weight is equal to one only if non-text elements are not included, so all the foreground pixels of the region I_{ij} are considered as correctly detected. This is the case where, for example, two paragraphs have been marked as one or two different ground-truth regions. For both cases, the proposed evaluation metric does not set a penalty.



(a) (b)

(c)

Fig. 6. An example of a set of ground-truth regions that have been merged into regions without horizontal overlap. (a) Ground-truth regions (b) result regions (c) intersection regions without horizontal overlap. w_{11}<1 and w_{21}<1 since non-text elements are included. On the other hand w_{32} =1 and w_{42} =1, so all the foreground pixels of ground-truth regions G_3 and G_4 are considered as correctly detected.

- Merging having horizontal overlap

If the intersection region I_{ij} overlaps horizontally with one or more other regions $I_{i'j}$, the text lines of ground-truth region G_i may have been merged with text lines of different ground-truth regions (e.g. text lines of a multi column document), so they cannot be detected correctly in the subsequent processing steps. For the calculation of the weight w_{ij} we follow the same procedure as in the case of splitting with horizontal overlap described previously. Our goal is to detect the subregions of I_{ij} without horizontal overlap (see Eq. 3) in order to consider the foreground pixels of them as correctly detected (see Eq. 4).

(iv) Noise has been included:

The final possible condition refers to the case that the result region R_j overlaps only with one ground-truth region but non-text elements (noise, image, borders, separator lines etc.) have been also included. In this case, we set a penalty according to the percentage of non-text foreground pixels. The corresponding weight w_{ij} is defined as follows:

$$w_{ij} = \frac{F(I_{ij})}{F(R_j)} \quad (6)$$

IV. Discussion

In this section we discuss the main advantages of the proposed performance evaluation framework: (i) independence from a strictly defined ground-truth and (ii) tolerance to insignificant errors. Representative examples are also given.

A first advantage of the proposed evaluation methodology is that it avoids the dependence on a strictly defined ground-truth. Ground-truth creation for page segmentation is quite ambiguous and may differ among users. For example, if there is a large blank gap between two paragraphs of the same column, these paragraphs may be marked as one or two different ground-truth regions. Figure 7 presents an example of a document image with two different page segmentation ground-truth approaches, which both follow the aforementioned guidelines (see Sect. III-A), as well as a corresponding page segmentation result. The proposed evaluation metric *SR* will be the same regardless the ground-truth approach used.

The second advantage is that the proposed evaluation measure is not very strict concerning possible page segmentation errors which do not adversely affect the subsequent processing. For example, if the main text zone of a document image is merged with a small marginal note, this will be considered by our approach as a partial error. Following the proposed approach, subregions which can be processed successfully by the subsequent text line segmentation stage are detected (see Fig. 8 - regions with dashed lines) letting the SR measure reflect the percentage of text lines which are not affected by the page segmentation error.

V. Conclusions

A goal-oriented performance evaluation methodology which defines a comprehensive evaluation measure is presented. It is a pixel-based approach which deals only with text regions and the evaluation measure reflects the percentage of the text information in which the subsequent processing (e.g. text line segmentation) can be applied successfully. The main advantages of the proposed performance evaluation framework concern its independence from a strictly defined ground-truth and its tolerance to insignificant page segmentation errors.

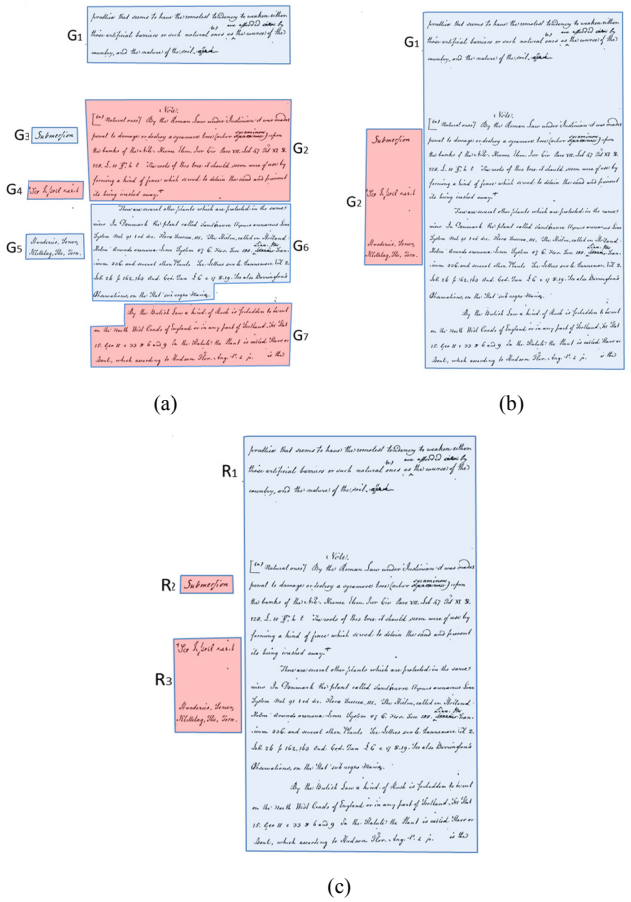


Fig. 7. An example of a document image with two different page segmentation ground-truth approaches which are in consistent with our guidelines (see Sect. III.A). (a) First ground-truth approach (7 regions) (b) second ground-truth approach (2 regions) (c) a page segmentation result for which the proposed evaluation metic *SR* is equal to 100 regardless the ground-truth approach used.

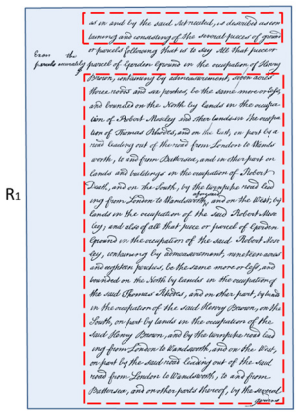


Fig. 8. Possible insignificant segmentation error produced by a page segmentation tecnique. The proposed evaluation measure *SR* is equal to 92.46 since the majority of the text lines (27 out of 31) can been detected correctly (dashed line).

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 600707 (project tranScriptorium). This work has been also supported by the OldDocPro project (ID 4717) funded by the GSRT.

REFERENCES

[1] L. O'Gorman, "The document spectrum for page layout analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 11 pp. 1162-1173, 1993.

[2] S.W. Lee and D.S. Ryu, "Parameter-Free Geometric Document Layout Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 1240-1256, 2001.

[3] M. Benjelil, S. Kanoun, R. Mullot and A.M. Alimi, "Complex documents images segmentation based on steerable pyramid features," International Journal on Document Analysis and Recognition (IJDAR), vol. 13, no. 3, pp. 209-228, 2010.

[4] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013," 12th International Conference on Document Analysis and Recognition (ICDAR'13), pp. 1454-1458, Washington DC, USA, August 2013.

[5] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "Historical Document Layout Analysis Competition," 11th International Conference on Document Analysis and Recognition (ICDAR'11), pp. 1516-1520, Beijing, China, September 2011.

[6] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods," 11th International Conference on Document Analysis and Recognition (ICDAR'11), pp. 1404-1408, Beijing, China, September 2011.

[7] J. Kanai, S.V. Rice, T.A. Nartker, and G. Nagy, "Automated Evaluation of OCR Zoning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 86-90, 1995.

[8] S. Mao and T. Kanungo, "Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms," IEEE Transaction on Analysis and Machine Intelligence, vol. 23, no. 3, pp. 242-256, 2001.

[9] J. Liang, I.T Phillips and R.M Haralick, "Performance Evaluation of Document Structure Extraction Algorithms," Computer Vision and Image Understanding, vol. 84, no 1, pp. 144-159, 2001.

[10] F. Shafait, D. Keysers, and T.M. Breuel, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images," 18th International Conference on Pattern Recognition (ICPR'06), pp. 872-875, 2006.

[11] T.M. Breuel, "Representations and metrics for off-line handwriting segmentation," 8th International Workshop on Frontiers in Handwriting Recognition (ICFHR'02), pp. 428–433, Ontario, Canada, 2002.

[12] K. Kise, A. Sato and M. Iwata, "Segmentation of Page Images Using the Area Voronoi Diagram," Computer Vision and Image Understanding, vol. 70, no 3, pp. 370-383, 1998.

[13] M. Agrawal and D. Doermann, "Voronoi++: A Dynamic Page Segmentation Approach Based on Voronoi and Docstrum Features," 10th International Conference on Document Analysis and Recognition, (ICDAR'09), pp.1011-1015, July 2009.