ICDAR2015 Competition on Recognition of Documents with Complex Layouts – RDCL2015[†]

A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher
Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford
Greater Manchester, M5 4WT, United Kingdom
www.primaresearch.org

Abstract—**This paper presents an objective comparative evaluation of page segmentation and region classification methods for documents with complex layouts. It describes the competition (modus operandi, dataset and evaluation methodology) held in the context of ICDAR2015, presenting the results of the evaluation of eight methods – four submitted, two state-of-the-art systems (one commercial and one open-source) and their two immediately previous versions. Three scenarios are reported in this paper, one evaluating the ability of methods to accurately segment regions and two evaluating both segmentation and region classification (one with emphasis on text and the other focusing only on text). The results indicate that an innovative approach has a clear advantage but there is still a considerable need to develop robust methods that deal with layout challenges, especially with the non-text content.**

Keywords - performance evaluation; page segmentation; region classification; layout analysis; recognition; datasets;

I. INTRODUCTION

Layout Analysis (Page Segmentation and Region Classification) is a critical step in the recognition workflow. Its performance significantly influences the overall success of a digitisation system, not only in terms of OCR accuracy but also in terms of the usefulness of the extracted information (in different use scenarios). It is also one of the most well-researched and active fields, indicating an appreciation that the problem is far from being solved. Frequently, methods are devised with a specific application in mind and are fine-tuned to the image dataset used by their authors. However, the variety of documents encountered in real-life situations (and the issues they raise) is far wider than the target document types of most methods.

In addition, OCR, largely abandoned by academic researchers, faces challenges in large-scale digitisation and is still not performing well enough to not require costly manual post-correction. Systematic evaluation is crucial to analyse the remaining obstacles and attempt to make progress.

The aim of the ICDAR Page Segmentation competitions (the oldest running ICDAR competition since 2001) has been to provide an objective evaluation of methods, on a realistic contemporary dataset, enabling the creation of a baseline for understanding the behaviour of different approaches in different circumstances. This is the only international layout analysis competition series that the authors are aware of. Other evaluations of page segmentation methods have been constrained by their use of indirect evaluation (e.g. the OCR-based approach of UNLV [1]) and/or the limited scope of the dataset (e.g. the structured documents used in [2]. In addition, a characteristic of previous reports has been the use of rather basic evaluation metrics. While the latter point is also true to some extent of early editions of this competition series, which used precision/recall type of metrics, the 5th edition of the ICDAR Page Segmentation competition (ICDAR2009) [3] made significant additions and enhancements. First, that competition marked a radical departure from the previous evaluation methodology. A new evaluation scheme was introduced, allowing for higher level goal-oriented evaluation and much more detailed region comparison. In addition, the datasets used since then have been selected from new repositories [4][5] that contain different instances of realistic documents.

This edition (RDCL2015) is based on the same principles established and refined by the 2011 and 2013 competitions on historical document layout analysis [6] but its focus is on documents with complex layouts. The evaluation scenarios selected for this competition reflect the significant need to identify robust and accurate methods for large-scale digitisation initiatives.

An overview of the competition and its modus operandi is given next. In Section 3, the evaluation dataset used and its general context are described. The performance evaluation methodology is described in Section 4, while each participating method is summarised in Section 5. Finally, different comparative views of the results of the competition are presented and the paper is concluded in Sections 6 and 7, respectively.

II. THE COMPETITION

RDCL2015 had the following three objectives. The first was a comparative evaluation of the participating methods on a representative dataset (i.e. one that reflects the issues and their distribution across library collections that are likely to be scanned). Delving deeper, the second objective was

[†] This work has been part-funded through the EU 7th Framework Programme grant SUCCEED (Ref. 600555) and the EU Competitiveness and Innovation Framework Programme grant Europeana Newspapers (Ref. 297380).

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

a detailed analysis of the performance of each method in different scenarios from the simple ability to correctly identify and label regions to a text recognition scenario where the reading order needs to be preserved. This analysis facilitates a better understanding of the behaviour of methods in different digitisation scenarios across the variety of documents in the dataset. Finally, the third objective was a placement of the participating methods into context by comparing them to leading commercial and open-source systems currently used in industry and academia.

The competition proceeded as follows. The authors of candidate methods registered their interest in the competition and downloaded the *example* dataset (document images and associated ground truth). The *Aletheia* [7] ground-truthing system (which can also be used as a viewer for results) and code for outputting results in the required PAGE format [8] (see below) were also available for download. Three weeks before the competition closing date, registered authors of candidate methods were able to download the document *images* of the *evaluation* dataset. At the closing date, the organisers received both the executables and the results of the candidate methods on the evaluation dataset, submitted by their authors in the PAGE format. The organisers then verified the submitted results and evaluated them.

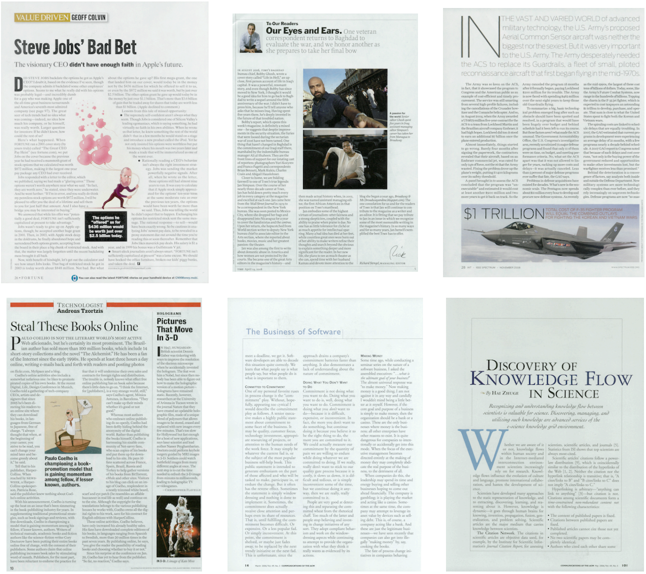

Figure 1. Page images in the example set.

III. The Dataset

The importance of the availability of realistic datasets for meaningful performance evaluation has been repeatedly discussed and the authors have addressed the issue for contemporary documents by creating the PRImA Layout Analysis dataset with ground truth [4] and making it available to all researchers. The overall dataset contains a wide selection of contemporary documents (with complex as well as simple layouts) together with comprehensive ground truth and extensive metadata. Particular emphasis is placed on magazines (mostly) and technical articles, which are likely to be the focus of digitisation efforts.

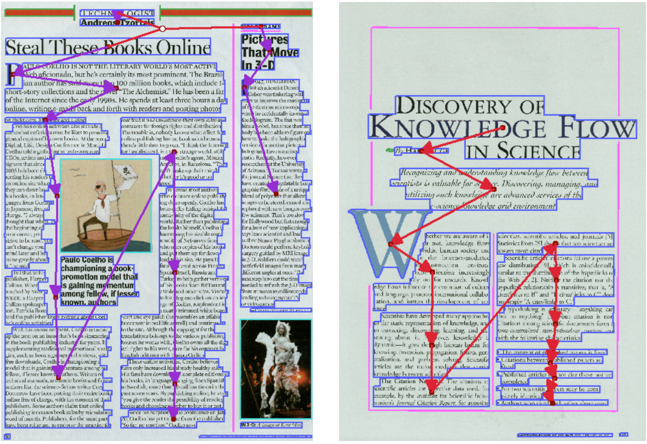
Steal These Books Online
Pictures That Move In 3-D
DISCOVERY OF KNOWLEDGE FLOW IN SCIENCE

Figure 2. Sample images showing the region outlines (blue: text, magenta: separator, green: graphic, cyan: image) and reding order.

For the purpose of this competition, the evaluation set consisted of 70 page images selected from the PRImA Layout Analysis dataset as a representative sample ensuring a balanced presence of different issues affecting layout analysis and OCR. Such issues include the presence non-rectangular shaped regions, varying text column widths, varying font sizes, presence of separators and regions of "reverse video" text (light-coloured text on a dark background). The presence of running headers and captions of illustrations/photos in addition to the main body of text, pose difficulties in the identification of the correct reading order of the page.

In addition to the evaluation set, six representative images were selected as the example set that was provided to the authors with ground truth. The pages from the latter can be seen in Fig. 1.

The ground truth is stored in the XML format which is part of the PAGE (Page Analysis and Ground truth Elements) representation framework [8]. For each region on the page there is a description of its outline in the form of a closely fitting polygon. A range of metadata is recorded for each different type of region. For example, text regions hold information about *language*, *font*, *reading direction*, *text colour*, *background colour*, *logical label* (e.g. heading, paragraph, caption, footer, etc.) among others. Moreover, the format offers sophisticated means for expressing reading order and more complex relations between regions. Sample images with ground truth description can be seen in Fig. 2.

IV. Performance Evaluation

A. Layout Analysis

The performance analysis method used for this competition [9] can be divided into three parts. First, all regions (polygonal representations of ground truth and method results for a given image) are transformed into an interval representation, which allows efficient comparison and calculation of overlapping/missed parts. Second, correspondences between ground truth and segmentation result regions are determined. Finally, errors are identified, quantified and qualified in the context of one or more use scenarios.

The region correspondence determination step identifies geometric overlaps between ground truth and segmentation result regions. In terms of Page Segmentation, the following situations can be determined:

- *Merger*: A segmentation result region overlaps more than one ground truth region.
- *Split*: A ground truth region is overlapped by more than one segmentation result region.
- *Miss (or partial miss)*: A ground truth region is not (or not completely) overlapped by a segmentation result region.
- *False detection*: A segmentation result region does not overlap any ground truth region.

In terms of Region Classification, considering also the *type* of a region, an additional situation can be determined:

- *Misclassification*: A ground truth region is overlapped by a result region of another type.

Based on the above, the segmentation and classification errors are *quantified*. This step can also be described as the collection of raw evaluation data. The amount (based on overlap area) of each single error is recorded.

This raw data (errors) are then *qualified* by their significance. There are two levels of error significance. The first is the implicit *context-dependent* significance. It represents the logical and geometric relation between regions. Examples are *allowable* and *non-allowable* mergers. A merger of two vertically adjacent paragraphs in a given column of text can be regarded as allowable, as the result will not violate the reading order. On the contrary, a merger between two paragraphs across two different columns of text is regarded as non-allowable, because the reading order will be violated. To determine the allowable/non-allowable situations accurately, the reading order, the relative position of regions, and the reading direction and orientation are taken into account.

The second level of error significance reflects the additional importance of particular errors according to the use scenario for which the evaluation is intended. For instance, to build the table of contents for a print-on demand facsimile edition of a book, the correct segmentation and classification of page numbers and headings is very important (e.g. a merger between those regions and other text should be penalised more heavily).

Both levels of error significance are expressed by a set of weights, referred to as an *evaluation profile* [9]. Each evaluation scenario has a corresponding evaluation profile.

Appropriately, the errors are also weighted by the size of the area affected (excluding background pixels). In this way, a missed region corresponding to a few characters will have less influence on the overall result than a miss of a whole paragraph, for instance.

For comparative evaluation, the weighted errors are combined to calculate overall error and success rates. A non-linear function is used in this calculation in order to better highlight contrast between methods and to allow an open scale (due to the nature of the errors and weighting).

B. *Text Recognition*

For the evaluation of OCR results a word-based method has been implemented. As in [10], the order of the words is not considered (Bag of Words) since the reading order of the submitted results is not known.

Since no participant submitted OCR results, only the Layout Analysis results have been evaluated.

V. Participating Methods

Brief descriptions of the methods submitted to the competition are given next. Each account has been provided by the method's authors and summarised by the organisers.

A. *The Fraunhofer Segmenter*

This method, submitted by the NetMedia Group at Fraunhofer IAIS (based at Sankt Augustin, Germany), is essentially the same as the Historical Newspaper Edition of the Fraunhofer Segmenter submitted to the HNLA2013 competition [6] (predecessor of this competition), where a detailed description of its processes can be found.

In summary, it is comprehensive approach where, after a page de-shadowing operation and a global or local binarisation (selection applied based on the computation of several features), black and white (logical) separators are identified. A hybrid page segmentation approach combines bottom-up component aggregation with top-down constraints in the form of *logical column layout* (determined from the lists of black and white separators identified earlier). Regions of text are separated from non-text based on a number of (text-like) characteristics of components within regions. Considering the textual regions only, text lines are computed and, using font information, paragraphs/columns are built containing text of similar font.

B. *The ISPL method*

This layout analysis system was submitted by Hyung Il Koo from Ajou University, Suwon, Korea and Dong Ju Jeong and Nam Ik Cho from Seoul National University, Seoul, Korea. It is a bottom-up approach that first extracts text-lines in images, and estimates paragraph structures using detected text-lines. Then, other regions (e.g., separators, pictures) are extracted from the non-text regions.

For the text-line extraction, connected components are extracted [11] and they are grouped into text-lines [12]. For the reliable extraction, the text/non-text classifier presented in [11] was adopted. From the detected text-lines, the method estimates the regions of drop caps as well as the paragraph structures. Finally, by applying the LSD algorithm [13] to non-text regions, vertical/horizontal separators are detected.

To detect the image regions, an inpainting method is applied to the detected text regions [14] and an intermediate result is generated where text-lines are removed (i.e., the image consists of non-text regions and background). To this intermediate image, a salient object detection algorithm is applied and the extracted objects are the image regions.

C. The MHS method

The *Page Segmentation Using Multilevel Homogeneous Structure (MHS)* method was submitted by Tuan Anh Tran, Hoai Nam Vu, In Seop Na and Soo Hyung Kim from Chonnam National University, Republic of Korea. It is a hybrid method involving both connected component analysis and white space (background) analysis. The method works on bitonal (black and white) images, obtained from the given grayscale ones by applying Sauvola binarisation with integral images. The threshold is fixed for window size ½ of the minimum of the width and the height of input image.

The method starts by identifying connected components and heuristically filters out all those that can be reliably deemed to be noise or non-text regions. On the remaining regions a multilevel classification is performed based on multilevel homogeneous regions and white space analysis to identify all text and non-text components. This is an iterative process that contains three main steps: Segmentation (subdividing regions), recursive filtering, and convergence. At the end of the process, all identified elements (text and non-text) are considered again to remove noise and merge the discrete components into regions.

The output of the above process consists of two images, one containing the text components and the other the non-text ones. On the image containing text, adaptive mathematic morphology is applied to obtain the text regions. The kernel is based on the size of white space and white lines in each homogeneous text region. On the non-text image, based on the properties of non-text elements, components are classified in the following order: line, table (use reverse image), separator, and image.

In a final refinement step, region boundaries are corrected, large text regions are segmented into paragraphs, and the functional labels of text regions are indentified (e.g. heading, page number etc.) based on the size and location of the regions. Further noise is removed by examining all small remaining regions and those on the birder of the image.

D. The PAL method

This bottom-up approach was submitted by Kai Chen, Fei Yin and Cheng-Lin Liu of the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation of the Chinese Academy of Sciences. First, using an idea similar to the one proposed in [15] to get the edge boxes of text proposals, each edge box is binarised using Otsu's algorithm, with high-value and low-value pixels placing in Image H and L, respectively. After extracting connected components (CCs) in both Image H and L, a first filtering is performed to identify those that are apparently not text (based on their geometric characteristics) and those that are text (based on similarity characteristics). The remaining CCs are classified into text and non-text using an SVM with features extracted from skeleton, stroke width colour etc.

Text lines are obtained in Image H and L independently, based on an analysis of the alignment of neighbouring CCs – grouping adjacent CCs into text lines if their aligning orientations are consistent. Subsequently, the method in [16] is used to extract background as whitespace rectangles in Image H and L separately. After appropriate filtering, the remaining (foreground) rectangles are grouped to form the final text lines.

Text lines are grouped into text blocks in both Image H and L, using the method in [16]. Text blocks and non-text CCs are then classified into different types heuristically in both Image H and L. Finally, the results from both Image H and L are combined.

VI. Results

Evaluation results for the above methods are presented in this section in the form of graphs with corresponding tables. For comparison purposes, the layout analysis and recognition components of a leading product, ABBYY FineReader® Engine 11 (FRE11), and that of the popular open-source system, Tesseract 3.03 are also included. For a comparison between versions, previous ones of FineReader (FRE10) and Tesseract (3.02) have also been evaluated. It must be noted that FineReader and Tesseract have been evaluated with no prior training or knowledge of the dataset.

Three scenarios have been defined for the competition, each with a corresponding evaluation profile. The first profile is used to measure the pure segmentation performance. Therefore, misclassification errors are ignored completely. Miss and partial miss errors are considered worst and have the highest weights. The weights for merge and split errors are set to 50%, whereas false detection, as the least important error type, has a weight of only 10%. Results for this profile are shown in Fig. 3.

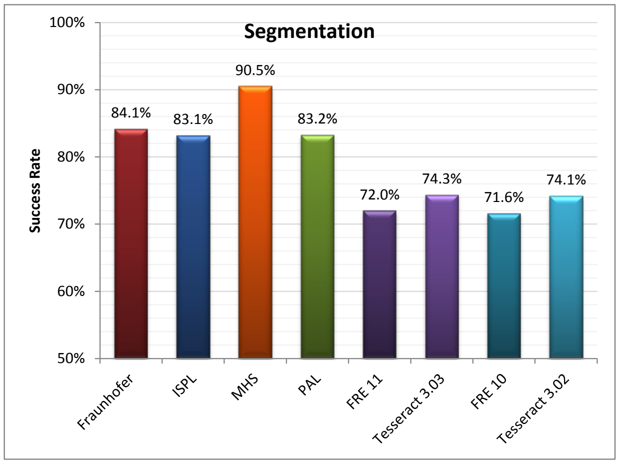

Figure 3. Results using the segmentation evaluation profile.

The second profile ("OCR") also evaluates region classification, in the context of a typical OCR system, focusing primarily on text but not ignoring the non-text regions. Accordingly, this profile is similar the first but misclassification of text is weighted highest and all other misclassification weights are set to 10%. Results for this profile are shown in Fig. 4. The third profile ("Text Only") is based on the OCR profile but focuses solely on text, ignoring non-text regions. Results for this profile are shown in Figure 5. A breakdown of the layout analysis errors made by each method (OCR scenario) is given in Fig. 6.

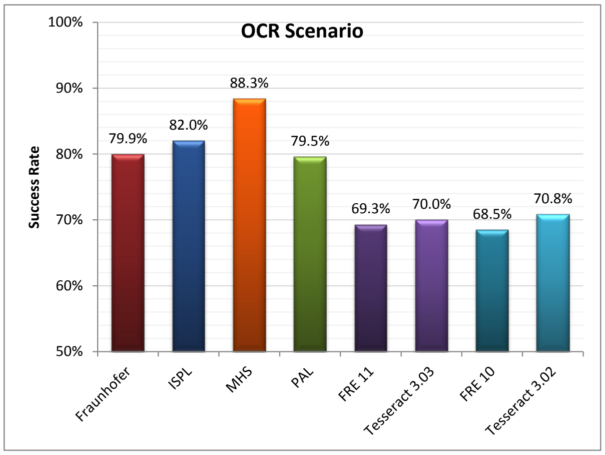

Figure 4. Results using the OCR-scenario evaluation profile.

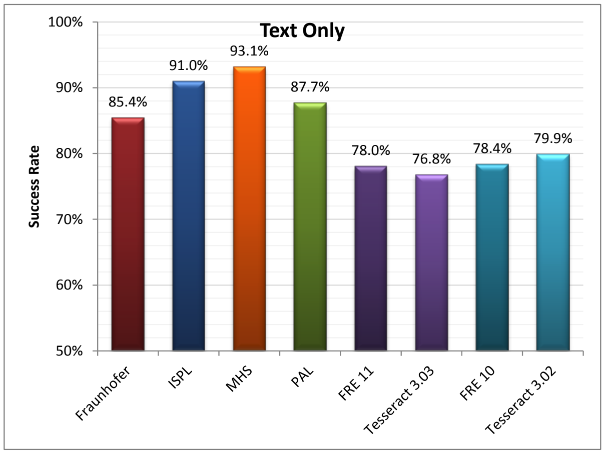

Figure 5. Results using the Text Only evaluation profile.

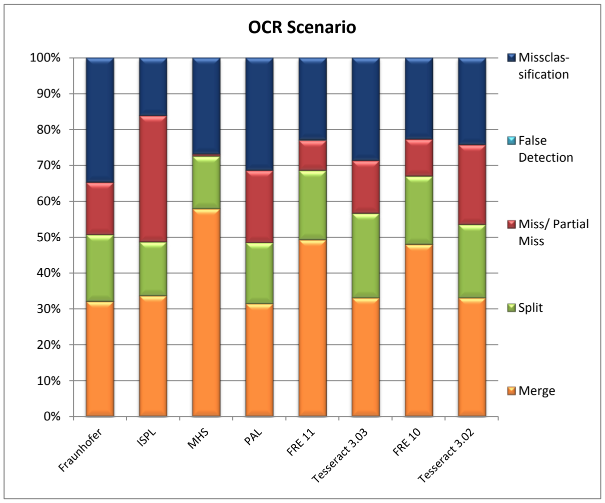

Figure 6. Breakdown of errors made by each method.

VII. Concluding Remarks

The Fraunhofer, ISPL and PAL submissions follow a similar bottom-up approach (especially Fraunhofer and PAL) and their performance is rather similar in the segmentation scenario. The hybrid approach of MHS, however, seems to have a clear advantage, especially in not missing regions. The latest versions of FineReader and Tesseract also present a marginal improvement over their previous versions. In terms of focusing primarily on text, or even ignoring non-text regions altogether, the results are more even with MHS still outperforming the others but ISPL close in second position. Both the latest Tesseract and FineReader seem to have slightly worse performance than their predecessors. It is also worth reporting that MHS and ISPL have the least standard deviation in all evaluation scenarios. Finally, from a closer analysis of the errors common to all methods, it is clear that there is still a considerable need to develop robust methods that deal with the issues such as accurate segmentation of non-text regions and significantly varying font sizes.

References

[1] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, “Automated Evaluation of OCR Zoning”, *IEEE PAMI,* 17(1), 1995, pp. 86-90.

[2] F. Shafait, D. Keysers and T.M. Breuel, “Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms” *IEEE PAMI,* 30(6), 2008, pp. 941–954.

[3] A. Antonacopoulos, S. Pletschacher, D. Bridson, C. Papadopoulos, “ICDAR2009 Page Segmentation Competition”, *Proc. ICDAR2009*, Barcelona, Spain, July 2009, pp. 1370-1374.

[4] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher, “A Realistic Dataset for Performance Evaluation of Document Layout Analysis”, *Proc. ICDAR2009*, Barcelona, Spain, July 2009, pp. 296-300.

[5] C. Papadopoulos, S. Pletschacher, C. Clausner, A. Antonacopoulos, “The IMPACT dataset of Historical Document Images”, *Proc. HIP2013*, Washington DC, USA, August 2013, pp. 123-130.

[6] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, “ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013”, *Proc. ICDAR2013*, Washington DC, USA, Aug 2013.

[7] C. Clausner, S. Pletschacher and A. Antonacopoulos, “Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments”, *Proc. ICDAR2011*, Beijing, China, 2011.

[8] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.

[9] C. Clausner, S. Pletschacher and A. Antonacopoulos, “Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods”, *Proc. ICDAR2011*, Beijing, China, Sept 2011.

[10] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, “ICDAR2013 Competition on Historical Book Recognition – HBR2013”, *Proc. ICDAR2013*, Washington DC, USA, Aug 2013.

[11] H. I. Koo and D. H. Kim, “Scene text detection via connected component clustering and nontext filtering”, *IEEE Transactions on Image Processing,* 22(6):2296–2305, June 2013.

[12] H.I.Koo and N.I.Cho “State estimation in a document image and its application in text block identification and text line extraction” *in ECCV*, volume 6312, pages 421–434, 2010.

[13] R. von Gioi, J. Jakubowicz, J.-M. Morel and G. Randall “LSD: A fast line segment detector with a false detection control”, *IEEE Trans PAMI*, 32(4):722–732, April 2010.

[14] H.I. Koo, Y.K. Baik, and B.S. Kim, *Efficient blending methods for ar applications*, Jan. 3 2013. US Patent App. 13/283,462.

[15] C.L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges”, *Computer Vision–ECCV 2014*. Springer International Publishing, 2014: 391-405.

[16] Kai Chen, Fei Yin and Cheng-Lin Liu, “Hybrid Page Segmentation with Efficient Whitespace Rectangles Extraction and Grouping”, Proc. ICDAR2013, Washington DC, USA, Aug 2013.