2009 10th International Conference on Document Analysis and Recognition

ICDAR2009 Page Segmentation Competition[†]

A. Antonacopoulos, S. Pletschacher, D. Bridson and C. Papadopoulos

Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford, Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

Abstract

This paper presents an objective comparative evaluation of layout analysis methods in realistic circumstances. It describes the Page Segmentation competition (modus operandi, dataset and evaluation methodology) held in the context of ICDAR2009 and presents the results of the evaluation of four submitted methods. Two state-of-the art methods are also compared as well as the three methods from the ICDA2007 Page Segmentation competition. The results indicate that although methods continue to mature, there is still a considerable need to develop robust methods that deal with everyday documents.

1 Introduction

Layout Analysis is the first major step in a Document Analysis workflow where, after Image Enhancement, a higher (than pixel-level) representation of the page structure is obtained. Homogeneous printed regions are identified (Page Segmentation) and labelled according to the type of their content (Region Classification). The correctness of the output of Page Segmentation is crucial as it forms the basis for all subsequent analysis and recognition processes.

Page Segmentation is one of the most well-researched fields in Document Image Analysis, yet new methods continue to be reported in the literature, hinting that the problem is far from being solved. Successful methods have certainly been reported but, frequently, those 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.

The aim of the ICDAR Page Segmentation competitions (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 page segmentation competition series that the authors are aware of. While other evaluations of page segmentation methods have been presented in the literature, they have been rather 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. This latter point is also true of the previous editions of this competition series which used a variant of the established precision/recall type of metrics. These provided a useful but rather limited insight to the performance of page segmentation methods.

This 5th edition of the ICDAR Page Segmentation competition series incorporates significant additions and improvements. First, this competition marks a radical departure from the previous evaluation methodology [3]. The new evaluation scheme allows for higher level goal-oriented (e.g. violation of reading order) evaluation and much more detailed region comparison. Second, the dataset used has been selected from the new recently expanded PRImA dataset [4] that contains even more and different instances of realistic documents. Third, to place the competition results in context, selected state-of-the-art methods have also been evaluated. Finally, to track progress between this and the previous competition, the participating methods of both competitions have also been evaluated on the ICDAR2007 competition dataset.

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 method and metrics are described in Section 4, while each of the participating methods 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.

2 The competition

The page segmentation competition had the following three objectives. The first was a comparative evaluation of the participating methods on a realistic dataset (i.e. one that reflects commonly occurring everyday documents that are likely to be scanned). Delving deeper, the second

[†] *The work presented in this paper has been partly supported through the EU 7th Framework Programme grant IMPACT (Ref: 215064).*

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.275

objective was a detailed analysis of the performance of each method in different scenarios from the simple ability to correctly identify regions to a text recognition scenario where the reading order needs to be preserved. This analysis facilitates a better understanding of the behaviour of different methods in the variety of situations occurring in the dataset. Finally, the third objective was a placement of the participating methods into context by comparing them to state-of-the-art methods and to the participating methods in the ICDAR2007 competition.

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). One week 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 (new for this competition) and the results of the candidate methods on the evaluation dataset, submitted by their authors in a pre-defined format. The organisers then evaluated the submitted results.

3 The dataset

Over the years there has been scarce availability of ground truth for the evaluation of methods analysing complex layouts (e.g., having non-rectangular regions). It was not possible, therefore, to evaluate methods under realistic circumstances (as opposed to specific structured document types e.g. technical articles). A realistic dataset was first created for the ICDAR2003 Page Segmentation competition and progressively evolved into the PRImA contemporary dataset [4]. The overall dataset contains a wide selection of contemporary documents (with complex as well as simple layouts) together with accurate ground truth and extensive metadata. Particular emphasis is placed on magazines and technical journals which are likely to be the focus of digitisation efforts.

The ground truth is stored in a new XML Schema which is part of the PAGE (Page Analysis and Ground truth Elements) image representation framework [5]. For each region on the page there is a description of its outline in the form of an isothetic polygon (i.e. a polygon having only horizontal and vertical edges). Such a representation enables a very accurate and efficient geometric description, especially for complex-shaped regions. 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.

Figure 1. Sample evaluation set images.

A viewer was developed for examining the images and the corresponding ground-truth, and was distributed to the competition participants. A sample image with the ground truth description of regions can be seen in Fig. 2.

The types of regions defined for the competition (simplified from the total number of different types in the general dataset) are: (i) *text*, (ii) *graphics*, (iii) *line art*, (iv) *separator*—graphical line segments between regions, and (v) *noise*.

Figure 2. Image showing the region outlines (magenta: text, blue: image, green: separator).

4 Performance evaluation

A new performance evaluation methodology is used in this competition, as opposed to the simpler pixel-based precision/recall-type method used in previous ICDAR Page Segmentation competitions [3].

The new evaluation system [6] comprises three stages:

1. *Region representation:* Ground truth and segmentation regions are transformed into an interval-based representation.
2. *Region correspondence determination:* Using the interval-based representation, correspondence between parts of ground truth, segmentation and background regions is established.
3. *Error qualification and quantification:* Errors in correspondence between ground truth and segmentation regions are examined in the context of application scenarios and their significance is established.

The following conditions are identified:

1. A segmentation region that has no overlap with any ground truth region (wrongly detected region)
2. A ground truth region that has been completely overlapped by a segmentation region (correctly detected region)
3. A ground truth region that has been overlapped – completely or partially – by more than one segmentation region (split region)
4. More than one ground truth region has been overlapped – completely or partially – by a single segmentation region (merged regions)
5. A ground truth region that has not been completely (or not at all) overlapped by any number of segmentation regions (partially or wholly missed region)

The degree of success of a layout analysis method directly depends on the *type* as well as on the *quantity* of errors it makes. In terms of page segmentation, the five types of error (as listed above) can have different significance depending on *context* (within the document) and/or the *application scenario* (user defined).

According to context, errors (in particular mergers and splits) can be classified as *allowable* or *non-allowable*. Typical examples are:

- A merger of text regions within the reading order (e.g. two paragraphs within a single column of text) is allowable.
- A merger that violates the reading order (e.g. a paragraph of body text and a figure caption, or two paragraphs across different columns) is a non-allowable.
- A merger between regions of different type (e.g. a text paragraph and an image) is non-allowable.
- A split that violates the reading order (e.g. a paragraph split creating two columns) is non-allowable.
- A split that does not violate the reading order (a text line split off a paragraph) is allowable.

Error significance according to application scenario supplements the above. Typical situations include:

- A split graphic or a merger between two image regions may not be significant in a text extraction / recognition scenario.
- A missed heading or page number region can be very significant in an indexing scenario.
- A missed separator may not be as significant an error in a recognition (any region type) scenario.
- All errors in regions of a particular type can be far more significant in general than errors in regions of other types.

The significance of context as well as application scenarios is expressed by corresponding adjustable weights.

5 Participating methods

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

5.1 The DICE system

Chang An, Sui-yu Wang and Henry Baird, of Lehigh University, in Pennsylvania, USA submitted a method that performs pixel classification rather than region segmentation "in order to avoid the arbitrariness and restrictiveness of limited families of region shapes", as the authors state. The Document Image Content Extraction (DICE) system comprises two main steps. First individual pixels are classified primarily into *machine-print* text, *handwriting* text and *photograph* [7]. Next, a post-classification methodology [8] is used which enforces local uniformity without imposing a restricted class of region shapes.

In order to produce the polygonal region description required for the competition the following sequence of mathematical morphology operations was applied to the results of DICE using MATLAB®. First, masks are extracted for each content type. Second, isolated pixels are cleaned. Third, iterated open and close operations are used to remove small regions. Finally, interior pixels are removed and contours of polygons are extracted.

It should be noted that the DICE system is designed as a first step, executed earlier than "classical" layout-analysis methods. Once all machine print has been extracted, say by masking with content-specific regions, then other specialized methods would be used to decompose text into blocks, columns etc. Therefore, the precision/recall type of evaluation metrics are more applicable than the higher-level metrics that consider region structure and penalise the violation of reading-order.

5.2 The Fraunhofer Newspaper Segmenter

This system was submitted by Iuliu Konya, Stefan Eickeler and Christoph Seibert of the Fraunhofer Institute

for Intelligent Analysis and Information Systems at Sankt Augustin, Germany.

The method applies the following modules to the image, in sequence:

1. *Pre-processing*. Global optimal binarisation is applied to the input greyscale image.
2. *Black separator detection.* First, the quality of the horizontal and vertical separators is improved [9] before being extracted [10]. A subsequent triage of the separators is performed by using information about the dominant character size on the page.
3. *White separator detection.* Maximally empty rectangles are detected [11], but they must also satisfy certain conditions, e.g. their height must be large enough in relation to the dominant character size.
4. *Page segmentation.* A hybrid approach is applied comprising a bottom-up process [12] guided by top-down information given in the form of *logical column layout* of the page (determined by means of dynamic programming using the lists of separators). Text regions are separated from non-text ones using statistical properties of text (e.g. characters aligned on baselines).
5. *Text line and region extraction.* Exact textlines are detected again in the raw text regions detected in the previous step using a method similar to [12]. Font characteristics (e.g. stroke width, x-height, italics) are computed for each textline and used to derive the text regions with similar properties.

5.3 The REGIM-ENIS method

This page segmentation system was submitted by Mohamed Benjelil of REGIM at ENIS in Sfax, Tunisia. It is designed primarily for degraded multi-script multi-lingual complex official documents, containing also tabular structures, logos, stamps, handwritten text and images. It works in two stages.

First, the page image is segmented based on a steerable pyramid transform. The features extracted from pyramid sub-bands serve to locate and classify regions into text and non-text regions. The second stage performs script identification to the printed and handwritten regions.

5.4 The Tesseract method

Ray Smith of Google Inc. submitted this method which is a new addition to the Tesseract OCR system he has been developing over the years. The page layout analysis method uses bottom-up methods, including binary morphology and connected component analysis, to estimate the type (text, image, separator, or unknown) of connected components. Two of the key methods employed include neighbourhood stroke-width measurement, and appropriateness of overlap between adjacent connected components. (Measuring how well they fit on a text line.)

The method uses these preliminary type labels to detect the tab-stops that were used to mark out column boundaries, indents, table columns etc. when the document was formatted.

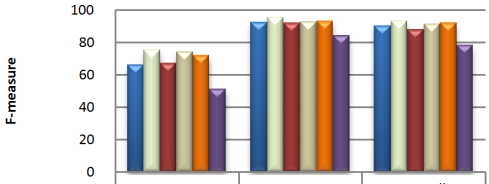
By taking various exceptions into account, the column layout of the page is determined from the detected tab-stops. The column layout constrains the construction of partitions of the page – approximating text lines – that are then gathered together into flows that make text regions.

An analogous process is applied to strips of image regions, so that they may also be wrapped into a poly-rectangle shape where text is flowed around non-rectangular images. The column layout also defines the physical reading order for the detected regions.

The method is described more fully in [13].

6 Results

Evaluation results are presented in this section, on the whole evaluation dataset (unless otherwise mentioned), in the form of graphs with corresponding tables. Two well-known state-of-the-art systems (ABBYY FineReader® Engine 8.1 and OCRopus 0.3.1) are also included for comparison. First, a simple F-measure report is given in Fig. 3 for the four submitted and the two state-of-the-art methods. Second, a detailed per-region-type (full-recognition scenario – all region types equally weighted) report expressed by the PRImA measure described in this paper is given for the four submitted methods in Fig. 4. In Fig. 5 the overall PRImA measure is reported (equally-weighted and text-weighted) for the four submitted and the two state-of-the-art methods. Finally, the PRImA measure (text, non-text and overall) of the four submitted methods is compared in Fig. 6, on the 32-image ICDAR2007 evaluation set, with the state-of-the-art methods and the ICDAR2007 competition methods [3].



	Non-text	Text	Overall
DICE	66.22	92.21	90.09
Fraunhofer	75.15	95.04	93.14
REGIM-ENIS	67.13	91.73	87.82
Tesseract	74.23	92.50	91.04
FineReader	71.75	93.09	91.90
OCRopus	51.08	84.18	78.35

Figure 3. F-measure of the four submitted and two state-of-the-art methods.

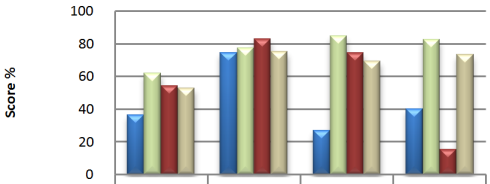
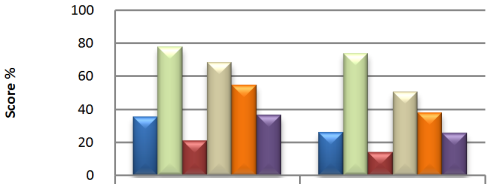


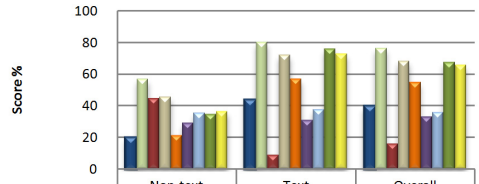
	Image	Noise	Separator	Text
DICE	36.46	74.71	27.08	40.02
Fraunhofer	61.83	77.32	84.51	82.37
REGIM-ENIS	54.42	82.77	74.53	15.44
Tesseract	52.95	75.03	69.42	73.24

Figure 4. PRImA measure for different region types for the four submitted methods.



	Overall	Overall text-weighted
DICE	35.31	25.97
Fraunhofer	77.55	73.74
REGIM-ENIS	20.81	13.94
Tesseract	68.15	50.30
FineReader	54.52	37.90
OCRopus	36.19	25.59

Figure 5. PRImA measure (standard and text-weighted) of the four submitted and two state-of-the-art methods.



	Non-text	Text	Overall
DICE	19.58	43.16	39.12
Fraunhofer	56.93	80.15	76.45
REGIM-ENIS	44.70	9.14	16.08
Tesseract	45.85	72.06	68.44
FineReader	21.44	57.00	54.90
OCRopus	29.26	31.03	32.98
BESUS	35.63	37.68	35.85
Tsinghua1	34.91	76.12	67.59
Tsinghua2	36.58	72.93	65.70

Figure 6. PRImA measure for the four submitted, two state-of-the-art and the three ICDAR2007 competition methods.

7 Concluding remarks

The aim of the ICDAR2009 Page Segmentation competition was to evaluate new and existing page segmentation methods using a realistic dataset and objective performance analysis. The dataset used comprised both technical articles (complex and not) and magazine pages. A new performance evaluation method was used with scenario and context-based measures in addition to simple precision / recall metrics (used in previous competitions). The competition ran in an off-line mode and evaluated the performance of four segmentation systems. The results show that the Fraunhofer Newspaper Segmenter method has an overall advantage, improving on both the state-of-the-art methods and the methods of the ICDAR2007 competition. It is interesting to note the discrimination ability of the new evaluation methodology (while the F-measure results are close, the PRImA measure is clearer and more informative). Hopefully this will enable more detailed examination of cases for improvement.

References

[1] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated Evaluation of OCR Zoning", *IEEE PAMI,* 17(1), January 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), June 2008, pp. 941–954.

[3] A. Antonacopoulos, B. Gatos and D. Bridson, "ICDAR2007 Page Segmentation Competition", *Proc. ICDAR2007*, Curitiba, Brazil, Sept. 2007, pp. 1279-1283.

[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.

[5] http://schema.primaresearch.org/PAGE/

[6] A. Antonacopoulos and D. Bridson, "Performance Analysis Framework for Layout Analysis Methods, *Proc. ICDAR2007*, Curitiba, Brazil, Sept. 2007, pp. 1258-1262

[7] H.S Baird, M.A. Moll and C. An, "Document Image Content Inventories", *Proc. IS&T/SPIE Document Recognition and Retrieval XIV,* San Jose, USA, 2007.

[8] C. An, H.S. Baird and P. Xiu, "Iterated Document Content Classification", *Proc. ICDAR2007*, Curitiba, Brazil, 2007.

[9] B. Gatos, D. Danatsas, I. Paratikakis and S.J. Perantonis, "Automatic Tbale Detection in Document Images", *Proc. ICAPR2005, Bath, UK, 2005, pp. 612-621.*

[10] Y. Zheng, C. Liu, X. Ding and S. Pan, "Form Frame Line Detection with Directional Single-Connected Chain", *Proc. ICDAR2001,* Seattle, USA, 2001.

[11] T.M. Breuel, "Two Algorithms for Geometric Layout Analysis", *Proc. DAS2002, Princeton, USA, 2002.*

[12] A.K. Jain and B. Yu, "Document Representation and Its Application to Page Decomposition", *IEEE PAMI, 20(3), 1998, pp. 294–308.*

[13] Ray Smith, "Hybrid Page Layout Analysis via Tab-Stop Detection", *Proc. ICDAR2009*, Barcelona, Spain, 2009.