# Continuous Competition on Recognition of Documents with Complex Layouts - RDCL

Christian Clausner and Apostolos Antonacopoulos
Pattern Recognition and Image Analysis Research Lab
University of Salford
United Kingdom
www.primaresearch.org

***Abstract*— This paper introduces a continuous competition and the underlying system that enables it based on the ICDAR Competition on Recognition of Documents with Complex Layouts – the most recent being RDCL2017. It is shown how researchers can perform the evaluation of their results using new functionality of the Aletheia system and how the outcome can be published on the competition website for comparison with other evaluated approaches.**

***Keywords- performance evaluation; page segmentation; region classification; layout analysis; OCR; 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).

The aim of the ICDAR Page Segmentation competitions (running 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 [1]. The used datasets have been selected from curated repositories [2][3] containing realistic and representative documents. The last edition (RDCL2017 [4]) is based on the same principles established and refined by previous competitions with its focus being on documents with complex layouts.

In addition to having snapshots of evaluation of methods at regular intervals (e.g. at ICDAR) it is important to enable and provide a continuous evaluation facility to track progress in the field and maintain a record of the performance of different approaches over a longer time period. In the rest of this paper, the continuous evaluation system and its use is presented, after an overview of the competition itself and its modus operandi.

## II. THE COMPETITION

RDCL has three objectives: 1) comparative evaluation of participating methods on a representative dataset; 2) detailed analysis of the performance in different scenarios; 3) placement of the methods into context by comparing them to commercial and open-source systems.

The initial competition (for ICDAR2017) proceeded as follows. The authors of candidate methods downloaded the *example* dataset (document images and ground truth). The *Aletheia* [5] ground-truthing system and code for outputting results in the required PAGE format [6] were also available. Three weeks before the deadline, participants downloaded the *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.


Figure 1. Three images from the example set.

The importance of realistic datasets for meaningful performance evaluation has been discussed and the authors have addressed the issue for contemporary documents by creating the PRImA Layout Analysis dataset with ground truth [2]. For this competition, the evaluation set consists of 75 images selected as a representative sample ensuring a balanced presence of different issues affecting layout analysis and OCR. In addition to the evaluation set, six images were selected as the example set that is provided to the authors with ground truth (Fig. 1).

The ground truth is stored in the PAGE XML format [6]. 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.

## III. THE EVALUATION SYSTEM

The performance analysis method [7] consists of two main parts. First, correspondences between ground truth and segmentation result regions are determined. Then, errors are identified, quantified and qualified in the context of 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*, *split*, *miss* / partial miss, and *false detection*. In terms of Region Classification, considering also the *type* of a region, *misclassification* can be determined as additional situation. Based on the above, the segmentation and classification errors are *quantified*. The amount (based on overlap area) of each single error is recorded

(raw evaluation data). The raw data (errors) are then *qualified* by their significance using two levels of error significance, expressed by a set of weights, referred to as an *evaluation profile* [7]. Each evaluation scenario has a corresponding profile.

For comparative evaluation, the weighted errors are combined to calculate overall error and success rates.

The complete evaluation procedure has been integrated into the Aletheia Document Analysis System [5]. A dedicated competition dialog (see Fig. 2) guides the user through the required steps, including:

- Downloading the evaluation set images
- Producing segmentation results in PAGE format
- Selecting the image and result folders
- Auto-validating the results (for completeness and XML correctness)
- Selecting one of the predefined evaluation scenarios
- Running the evaluation (takes a few minutes)
- Viewing / exporting results
- Submitting via email (optional)

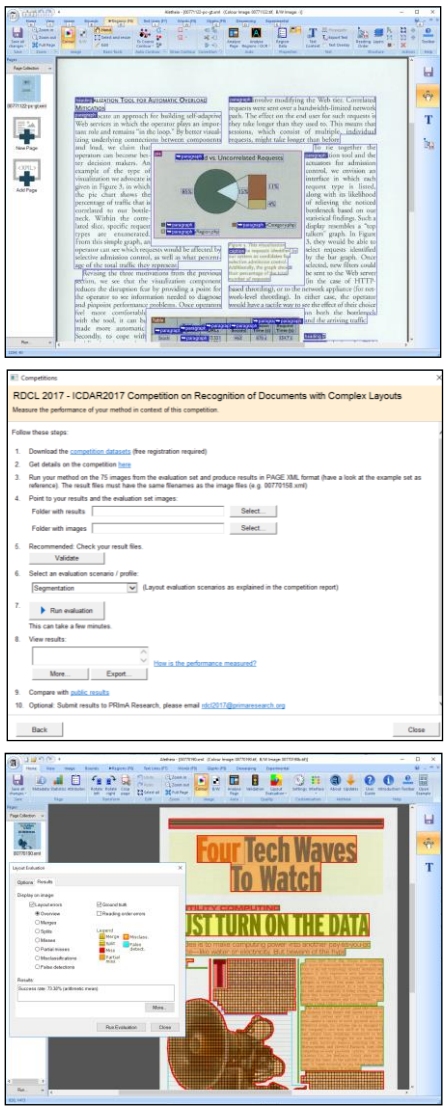


Figure 2. Aletheia with ground truth open (top), competition dialog (middle), and visual evaluation results for one page (bottom)

Detailed information and published results can be found on the competition website [8].

The user's evaluation results are presented in textual form in Aletheia and can be exported as comma-separated values (with per-page figures). The processing is performed locally on the user's system. An evaluation can therefore be repeated as often as required.

Aletheia also allows to evaluate segmentation results for individual pages, giving in-depth visual and textual feedback on different types of errors.

## IV. Discussion and Conclusion

The ICDAR competitions provide biennial snapshots of page recognition methods. The continuous RDCL competition builds upon that and adds the possibility for researchers to evaluate their systems at any time. For results to be published on the competition website (primaresearch.org/RDCL2017) the same rigor as in the ICDAR competition is used (validation by organisers). The ground truth of the evaluation dataset and the exact evaluation profile are kept secret for a fairer process. New results will be displayed alongside ICDAR2017 results but labelled clearly as 'new' since the original participants had limited time to finetune their methods.

A limit for how often results can be submitted has not been set but there will be a fair-use policy in place. A short method description and/or reference will be requested for each submission.

Aletheia and the competition dataset are publicly available at primaresearch.org.

## References

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

[2] 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.

[3] 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.

[4] C. Clausner, A. Antonacopoulos, S. Pletschacher, "ICDAR2017 Competition on Recognition of Documents with Complex Layouts – RDCL2017", *Proc. ICDAR2017*, Kyoto, Japan, 2017, pp. 1411-1416.

[5] 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.

[6] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, 2010, pp. 257-260.

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

[8] RDCL competition website: www.primaresearch.org/RDCL2017, January 2018