ICDAR2019 Competition on Recognition of Documents with Complex Layouts – RDCL2019

C. Clausner, A. Antonacopoulos, and S. Pletschacher
Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Science, Engineering and Environment, 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 ICDAR2019, presenting the results of the evaluation of twelve methods – nine submitted, three state-of-the-art systems (commercial and open-source). 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. Text recognition was a bonus challenge and was not taken up by all participants. 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-textual content.**

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). 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. Systematic evaluation is crucial to study the issues involved in order to make progress.

The aim of the ICDAR Page Segmentation competitions (the longest 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. 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 other competition reports has been the use of rather basic evaluation metrics. Since the 2009 edition of the ICDAR Page Segmentation competition a more extensive evaluation scheme has been used [3], allowing for higher-level goal-oriented evaluation and much more detailed region comparison, going far beyond simple precision/recall metrics. In addition, the used datasets have been selected from curated repositories [4][5] containing realistic and representative documents. This edition (RDCL2019) is based on the same principles established and refined by the 2011, 2013, 2015, and 2017 competitions on historical and contemporary document layout analysis [6] but its focus is on documents with complex layouts. The evaluation scenarios selected for this competition reflect the need to identify robust and accurate methods for large-scale digitisation initiatives.

An overview of the competition is given next. In Section 3, the evaluation dataset 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.

II. THE COMPETITION

RDCL2019 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). The second objective was 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 to place 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 (images and 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.

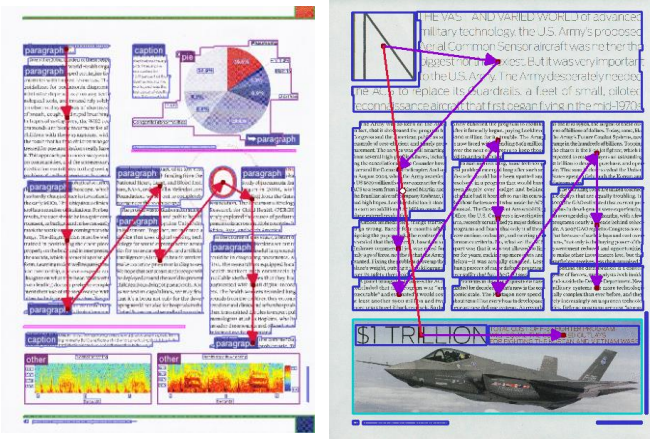

Figure 2. Sample images showing the region outlines (blue: text, purple: chart, green: graphic, cyan: image, magenta: separator) and reding order.

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. Emphasis is placed on magazines (mostly) and technical articles, which are likely to be the focus of digitisation efforts.

For this competition, the evaluation set consisted of 85 images. These included ten new scans taken from IEEE Spectrum magazines and 75 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 of 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/photographs in addition to the main body of text, pose difficulties in the identification of the correct reading order of the page.

Figure 3. New content / challenges in RDCL2019 (top: Map regions as introduced in PAGE XML version 2018, bottom: full-page reverse video).

In addition to the evaluation set, 15 representative images were selected as the example set that was provided to the authors with ground truth. 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. Structured content can be modelled with nested regions (regions within regions). For this competition only nested text was taken into account (table cells, text on images, chart labels etc.). Sample images with ground truth description can be seen in Fig. 2. New types of content and challenges are shown in Figure 3.

The dataset is available for download at www.primaresearch.org/datasets.

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 (not fully) overlapped by a 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 a different type.

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).

This raw data (errors) are then *qualified* by their significance using 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. In contrast, 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). 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. To this end, bitonal images are produced using the Sauvola method (window size 20, weight 0.4).

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

Nested regions (regions within regions) require special handling. Top-level regions (parent regions) and nested regions (child regions) are thereby treated as being in different layers. For each ground truth region, two error values are calculated: one in the same layer (top-level to top-level) and one across layers (top-level to nested or nested to top-level). The lower of the two is then used as final value for the region. More information on the handling of nested regions can be found in [6].

B. Text Recognition

For the evaluation of OCR results, character-based and word-based measures were used. The former gives a detailed insight into the recognition accuracy of a method while the word-based approach is more realistic in terms of use scenarios such as keyword-based search.

A major problem for the evaluation is the influence of the reading order of text regions. For simple page layouts, the order is obvious, but for more complex layouts, the reading order can be ambiguous. In such cases, measures that are affected by the reading order are less meaningful. An OCR method might recognise all characters perfectly, but if it does not return the regions in the same order as in the ground truth, it will get a very low performance score. Special care was therefore taken when selecting the evaluation measures.

The Character Accuracy [11] is based on the edit distance (insertions, deletions and substitutions) between ground truth and OCR result. The method was extended by the authors to reduce the influence of the reading order. The edit distance is thereby calculated for parts of the texts, starting with good matches and marking matched parts as “visited” until the

whole text is processed (unmatched parts count as deletion or insertion errors). The extended measure is called Flex Character Accuracy.

The word-based measure called *Bag of Words* (see [10]) disregards reading order entirely since it only looks at the occurrence of words and their counts, not at the context or location of a word.

All evaluation methods (and datasets) are available at the authors' website [12].

V. Participating Methods

An overview of the methods submitted to the competition is given in TABLE I. The individual descriptions were provided by the method's authors and summarised by the organisers. The full method descriptions are available on the competition website: www.primaresearch.org/RDCL2019.

TABLE I. Participating Methods and State-of-the-art off-the-shelf Systems

Method	Description
BINYAS	Showmik Bhowmik, Soumyadeep Kundu, Ram Sarkar - Department of Computer Science and Engineering, Jadavpur University, India. Mainly based on connected component analysis and morphology.
BKZA	Duc Nguyen, Cuong Ha - Ho Chi Minh City University of Technology. Using deep learning to segment page and heuristic algorithms for post-processing.
DSPH	Tan Lu, Ann Dooms - Vrije Universiteit Brussel. Document Segmentation with Probabilistic Homogeneity, using: binarization, text / non-text classification, text region extraction, non-text structure extraction.
JBM	Klára Janoušková, Michal Bušta and Jiří Matas - Czech Technical University, Prague. The method uses following steps: obtaining segmentation maps from a CNN, post-processing based on polygons, OCR (CNN-based).
LingDIAR	Dou Haobin - Lingban Tech Co., Ltd. Based on a multi-task deep network model trained with synthetic document images.
MHS	Tuan Anh Tran[a], Nam Quan Nguyen[a], Quoc Thang Nguyen[a], Hai Duong Nguyen[b], Soo Hyung Kim[c] - (a): HoChiMinh National University City - HCMUT, Viet Nam & Cinnamon AI, (b): Concordia University, Canada, (c): Chonnam National University, Gwangju, Republic of Korea. Based on following steps: negative/positive image detection, binarization, text / non-text classification, text segmentation, image classification, region refinement + labelling, OCR.
MICS	Yassine Ouali, Céline Hudelot - MICS, CentraleSupélec, France. A method similar to Pyramid Scene Parsing Network in combination with training on augmented data and final inference / post-processing steps.
TAQ	Nam Quan Nguyen, Tran Hai Anh Vo, Quoc Thang Nguyen - Cinnamon AI Lab Inc. Based on following steps: text / non-text classification, text region improvements / smoothing, non-text classification.
ZLCW	Chendi Zang, Hui Li, Xinfeng Chang, Yaqiang Wu - Lenovo Research. Using FCN (fully convolutional networks) trained with 15 original training images, augmented to 3960 images (using cropping, gaussian blur, adding random noise as well as colour jittering).
FRE11	ABBYY FineReader Engine 11 with PRImA FineReader-to-PAGE wrapper.
FRE12	ABBYY FineReader Engine 12 with PRImA FineReader-to-PAGE wrapper.
Tess.4	Tesseract 4 with PRImA Tesseract-to-PAGE wrapper.

VI. Results

Evaluation results for the above methods are presented in this section in the form of graphs and tables. For comparison purposes, the layout analysis and recognition components of a leading product, ABBYY FineReader® Engine (versions 11 and 12), and that of the popular open-source system, Tesseract 4 are also included. It must be noted that FineReader and Tesseract have been evaluated with no prior training or knowledge of the dataset.

All layout evaluation results are aggregated in TABLE II. arranged by scenario and test set. The larger set (named "all") contains all 85 pages of the evaluation set. The smaller set ("new") contains only the ten pages that were added for this year's run of the competition.

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 Figure 4.

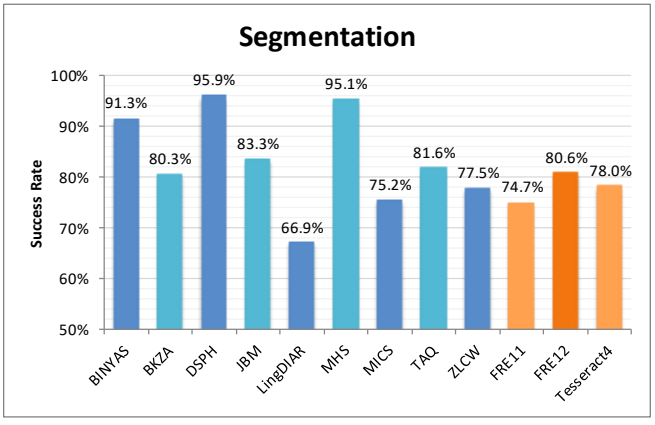

Figure 4. Results for "Segmentation" evaluation profile.

The second profile ("Segmentation + Classification") also evaluates region classification, in the context of a typical OCR system, focusing on text but not ignoring the non-text regions. Accordingly, this profile is similar to 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 Figure 5.

The third profile ("Text regions only") is based on the previous profile but focuses solely on text, ignoring non-text regions. Results for this profile are shown in Figure 6.

A breakdown of the layout analysis errors made by each method (Segmentation + Classification) is given in Figure 7.

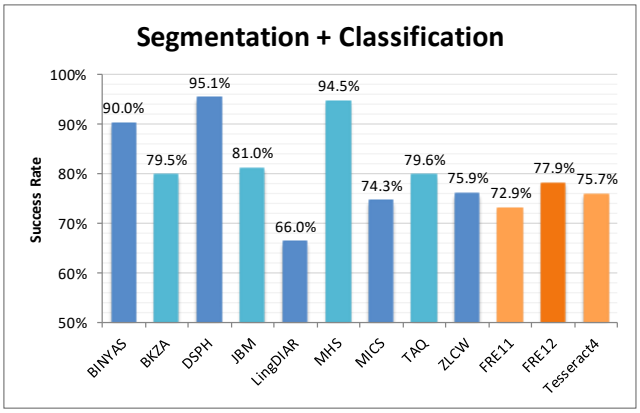

Figure 5. Results for "Segmentation + Classification" evaluation profile.

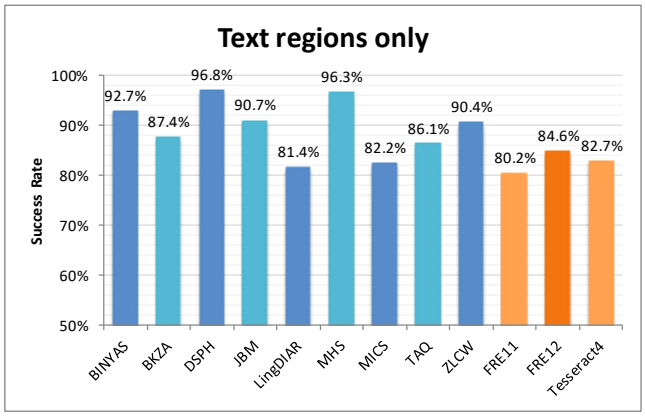

Figure 6. Results for "Text regions only" evaluation profile.

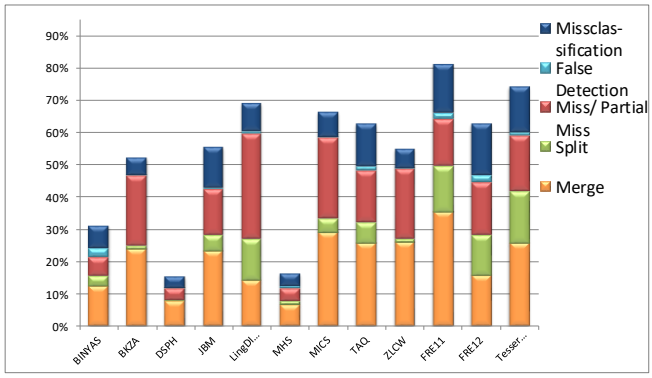

Figure 7. Breakdown of errors made by each method.

Fig. 8 shows the images that turned out the most and the least challenging across all participating methods. The image on the left received an average success rate of 60% (segmentation + classification scenario). The main reason for problems seems to be the heading that is made up of small components. The image on the right represents the most basic form of content – a two-column page with text only. The methods scored 99.2% on average.

Looking again at TABLE II. , it can be seen that the DSPH method performed the most consistent when comparing the results for "All" and "New". This points to good generalisation capabilities of the method (the ten pages in "New" were added this year and completely unseen before the competition). The runner-up method (MHS) performs very well on the 85 pages, on average, but scores 15% worse on the 10 new pages.

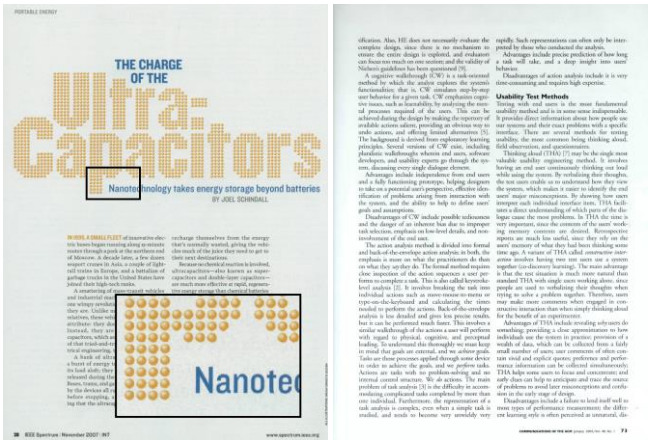

Figure 8. Most challenging (left) and least challenging (right) image of the evaluation set.

TABLE II. LAYOUT EVALUATION RESULTS PER SCENARIO ("ALL"=85 PAGES, "NEW"=10 PAGES ADDED IN 2019; VALUES ARE PERCENT SUCCESS; BEST PER COLUMN IN BOLD)

METHOD	SCENARIO					
	Segmentation		Segm. + Classification		Text regions only	
	All	*New*	*All*	*New*	*All*	*New*
BINYAS	91.26	76.56	89.96	72.24	92.69	85.46
BKZA	80.32	68.87	79.51	62.93	87.40	78.41
DSPH	**95.86**	**92.96**	**95.08**	**91.91**	**96.83**	**95.27**
JBM	83.28	73.36	80.97	70.55	90.74	88.61
LINGDIAR	66.86	60.26	65.99	59.74	81.41	84.40
MHS	95.10	79.87	94.50	78.52	96.30	87.77
MICS	75.20	68.66	74.30	64.24	82.20	74.69
TAQ	81.60	73.85	79.60	70.35	86.10	85.27
ZLCW	77.52	66.08	75.87	63.66	90.35	89.80
FRE11	74.73	72.37	72.90	68.56	80.19	86.95
FRE12	80.62	71.83	77.88	66.16	84.58	79.29
TESSERACT4	78.00	67.55	75.70	58.80	82.70	71.19

TABLE III. TEXT EVALUATION RESULTS (BEST PER COLUMN IN BOLD)

METHOD	BAG OF WORDS	CHAR ACC.	FLEX CHAR ACC.
BKZA	94.89%	47.98%	94.77%
LINGDIAR	91.32%	73.02%	89.81%
MHS	92.43%	61.33%	94.62%
FRE11	**97.49%**	48.24%	95.90%
FRE12	97.07%	**80.39%**	**96.12%**
TESSERACT4	95.68%	47.28%	95.27%

Text recognition – a bonus challenge – was submitted by three participants (BKZA, LingDIAR, and MHS). An overview of all results is provided in TABLE III. Fig. 9 shows the results using the bag of words evaluation method and Fig. 10 shows the character accuracy.

All methods achieve scores of 90% or better for bag of words. FineReader performs best in this setup. One interesting observation can be made from the differences of traditional character accuracy and flex character accuracy. FineReader Engine 12 seems to have big improvements with regards to reading order, when compared to version 11. The closer the character accuracy is to the flex character accuracy, the closer the OCR reading order is to the ground truth reading order.

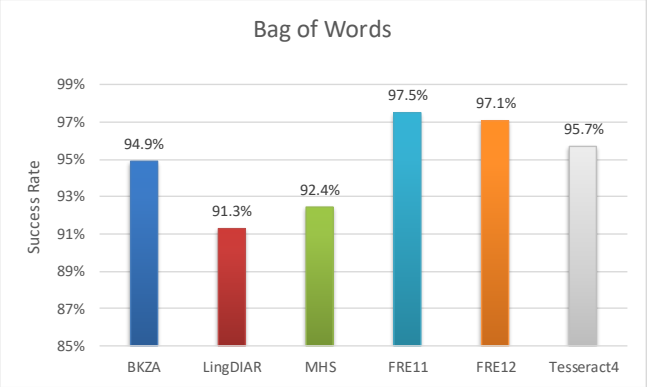

Figure 9. OCR evaluation result using bag of words method.

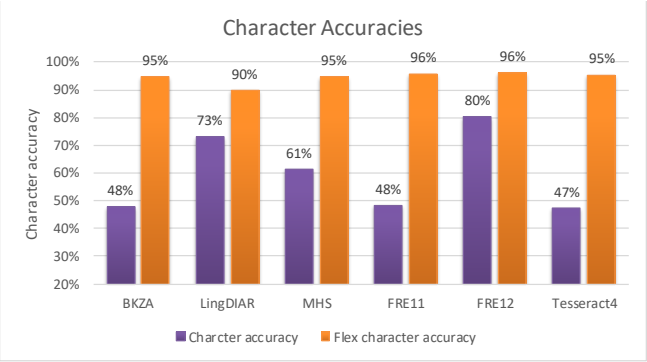

Figure 10. OCR evaluation result using character accuracy methods.

VII. Concluding Remarks

Despite the difficulty (e.g. table regions, inverse video etc.), good page segmentation results were achieved by the submitted methods. For general segmentation and region classification, the DSPH wins with the MHS method being on a close second place.

All methods perform better in the text-only scenario. Some however, have a clear focus on textual content. LingDIAR and ZLCW have almost 15% difference between the "Segmernation + Classification" and the "Text Regions Only" scenarios.

The off-the-shelf systems have made progress in page segmentation, but the dedicated, trained methods clearly outperform Tesseract and FineReader on the given dataset.

The breakdown by error type is particularly useful to identify areas of improvement for an algorithm. Mergers and misses represent the largest problem across all methods. False detection, on the other hand, is negligible.

While methods are maturing overall there is still room for improvements. Even the winning method suffers low scores for a few pages. More consistency would be desirable. The DSPH authors are certainly on the right track, they have – by some margin – the lowest standard deviation across the evaluation set (5.7%, the next best being MHS with 9.3%).

Methods based on neural networks are particularly impacted by the relatively small training set (example set). Data augmentation seems a popular choice to increase the number of training samples. Augmentation approaches include mirroring, cropping, and various image operations. Another option is to use third-party datasets. Following these strategies, both the JBM and the ZLCW methods performed well (e.g. in the Text Regions Only scenario).

More information, including longer method descriptions, can be found on the competition website: primaresearch.org / RDCL2019.

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, "ICDAR2017 Competition on Recognition of Documents with Complex Layouts – RDCL2017", *Proc. ICDAR2017*, Kyoto, Japan, November 2017, pp. 1404-1410.

[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] S.V. Rice, "Measuring the Accuracy of Page-Reading Systems", PhD thesis, University of Nevada, Las Vegas December 1996.

[12] PRImA Performance Evaluation Tools http://www.primaresearch.org/tools/PerformanceEvaluation