2011 International Conference on Document Analysis and Recognition

Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods*

C. Clausner, S. Pletschacher and A. Antonacopoulos

PRImA Lab, School of Computing, Science and Engineering, University of Salford,
Greater Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

Abstract - This paper presents an advanced framework for evaluating the performance of layout analysis methods. It combines efficiency and accuracy by using a special interval based geometric representation of regions. A wide range of sophisticated evaluation measures provides the means for a deep insight into the analysed systems, which goes far beyond simple benchmarking. The support of user-defined profiles allows the tuning for practically any kind of evaluation scenario related to real world applications. The framework has been successfully delivered as part of a major EU-funded project (IMPACT) to evaluate large-scale digitisation projects and has been validated using the dataset from the ICDAR2009 Page Segmentation Competition.

I. INTRODUCTION

Layout Analysis is a fundamental step in the process of Document Image Analysis. Its aim is to extract the underlying geometric and logical structure of a document from its low-level image data. The first step is to segment the document page into regions of interest (Page Segmentation or Zoning). The identified regions then need to be classified according to their type of content (Region Classification or Labelling). Furthermore, information of higher levels may be investigated, such as relations between regions (e.g. nested regions or reading order).

As all subsequent tasks in a document analysis system are based on layout analysis, the correctness of its output is crucial for the whole process. Numerous layout segmentation and classification methods have been reported over previous decades and new approaches are still emerging. Although many layout analysis systems have proven their success for rather constrained types of documents, there is still a high demand for robust methods, capable of dealing with a broad spectrum of layouts found in historic and contemporary documents.

The need for objective, comparative and detailed evaluation on realistic and large datasets is more pressing than ever. Past approaches mostly focus solely on benchmarking based on simple measures such as precision and recall. However, in order to find specific strengths and weaknesses of methods and give leads for developers to improve their algorithms, more sophisticated metrics are required. The ICDAR Page Segmentation Competition series confirms this view [1][2].

Depending on the application or context that layout analysis systems are intended for, different aspects of segmentation and classification may be of interest. Therefore, an evaluation system has to be customizable to be capable of meeting different, sometimes contradictory, expectations. Moreover, evaluation systems are expected to be as accurate as possible and efficient at the same time; especially in the case of large scale evaluation scenarios where the number of documents being processed may be considerably high.

Early approaches [3] followed the idea of indirect evaluation based on OCR results to measure the performance of the layout analysis stage of a system. Although this is useful for comparative means of whole workflows, it lacks giving deeper insight into the segmentation performance (further discussed in [4][5]).

Most later approaches concentrate on comparing region characteristics of ground truth and segmentation results. These approaches can be divided into geometry and pixel comparison based methods. The former [5][6][7] try to find geometric correspondences between page elements (regions, text lines, characters). To represent a region, most of these methods use bounding boxes. Ground truth is easy to produce and can be stored very efficiently. Nevertheless, there is a considerable disadvantage regarding complex layouts (e.g. text flowing around an image), where the use of bounding boxes can lead to overlapping regions. This problem has been addressed in the literature [4][8].

Pixel comparison approaches [1][9][10] use labelling on pixel level and are therefore predestined to be able to accurately handle complex layouts. The disadvantages are that ground truth production is more laborious [11] and storage demanding. Furthermore, processing the layout data is less efficient as it is for geometric representations.

II. THE FRAMEWORK

The presented framework for layout analysis performance evaluation is the natural successor of the method that has already proven its potential in the ICDAR2009 Page Segmentation Competition [2][12]. Advances have been made in almost all aspects of the system (efficiency, weighting, performance measures, XML format and user interface). It is component of a whole performance analysis infrastructure comprising XML formats, ground-truthing tools, evaluation modules, validators, converters, viewers, on-line datasets, etc., developed in context of the IMPACT project (IMProving ACcess to Text, EU-funded project

* This work has been supported in part through the EU 7th Framework Programme grant IMPACT (Ref: 215064).

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.282

aiming at improving technologies for mass digitisation of historical documents).

The framework avoids a compromise between accuracy and efficiency by using a geometric approach based on polygonal region outlines. By decomposing the polygons into an interval based representation, regions can still be compared very efficiently. Moreover, the framework encompasses an evaluation metric beyond simple scores based on cumulative errors. Together with an interactive result viewer it can support researchers and developers in improving their segmentation methods. Furthermore, the system achieves a high customizability by allowing the definition of tailor-made profiles for a wide range of evaluation scenarios.

Ground truth and segmentation results are stored according to an XML schema which is part of the PAGE (Page Analysis and Ground truth Elements) format framework [13]. Each region of a page is described by its outline in form of a polygon. In addition, a range of metadata can be recorded (e.g. for text elements there is among others: language, font, reading direction, text colour and sub-type such as paragraph or heading). For text regions, further structure levels are available in form of text lines, words and glyphs. The format also supports text content for ground truth and OCR results, a flexible reading order definition, region layers and region hierarchies. The PAGE XML format can be considered mature. It is already being used for large datasets of modern and historical documents in the scope of the IMPACT project as well as past ICDAR Page Segmentation Competitions and the contemporary dataset presented in [14].

III. PERFORMANCE ANALYSIS SYSTEM

The performance analysis method can be divided into three parts. Firstly, all regions have to be transformed into a special interval representation, which allows efficient further processing. Secondly, correspondences between ground truth and segmentation result regions are determined. Finally, errors are identified, quantified and qualified in the context of an application scenario.

A. Region Representation

The region representation is the key for an efficient and accurate evaluation. Polygons provide sufficient accuracy. However, in order to achieve efficiency as well, the polygons are internally converted into an interval representation (fig. 1). Therefore, the polygons are initially transformed into isothetic format (only horizontal and vertical edges). Due to the raster-based nature of digital document images, this transformation does not lead to a loss of information regarding the shape. Then, an interval representation is calculated. An interval is defined as a maximal rectangle that can be fitted horizontally inside a region [15]. An interval representation is the decomposition of a shape into a set of vertically adjacent horizontally orientated rectangles. Simple regions will be decomposed into only a few intervals, whereas complex polygons result in more intervals (the simplest shape, a box, is decomposed into one interval).

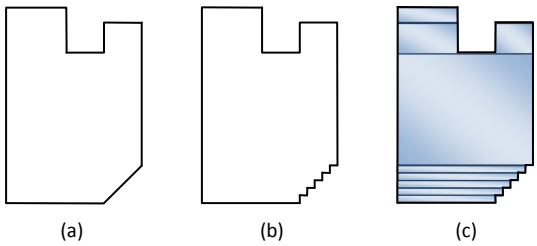

Figure 1. Process of interval decomposition: (a) Original polygon; (b) isothetic polygon; (c) Interval Representation

B. Region Correspondence Determination

The region correspondence determination is the step to find geometric overlaps between ground truth and segmentation result regions. For efficiency, all overlap candidates are determined by using bounding boxes. To decide if two (or more) regions overlap, a combined interval representation of the regarded regions is calculated. This way, it can be determined if and by exactly how much two regions overlap. Due to the nature of the representation only a small amount of operations is needed to compute the combined structure (less than pixel based approaches) [16]. The overlap information is then assembled in two look-up tables, containing a list of overlaps for each region (one-to-many relations). The first table has an entry for each ground truth region stating which segmentation result regions it overlaps with. The second table denotes for each segmentation result region which ground truth regions it overlaps with. Together they are used to identify the following conditions (PRImA measure, see [2]):

- Merge: 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 / partial miss: A ground truth region is not or not completely overlapped by a segmentation result region.
- False detection: A segmentation result region overlaps no ground truth region.

Considering also the type of a region, an additional measure for classification can be formulated:

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

In addition to these region based measures further logical aspects can be evaluated. Examples are absolute and relative region count deviation (as indicator for over- or under-segmentation) and reading order. The latter is novel in its complexity. The PAGE XML format allows a hierarchical definition of reading order relations between text regions, based on ordered and unordered groups. The evaluation therefore resembles a graph theoretical problem. For completeness and to allow the comparison to more simplistic evaluation approaches, the widely used measures precision, recall and F-measure are evaluated as well.

Evaluation can be carried out on all levels of text elements. However, for text lines, words and glyphs a reduced metric is used, since a misclassification on levels below regions is not possible (as there is only one element type per level).

C. Error Quantification and Qualification

Based on the aforementioned measures, the segmentation and classification errors are being 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. Using this raw data, the errors are then qualified by their significance. There are two types 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 merges. A merge of two vertically adjacent paragraphs of one column can be regarded as allowable, as a possible OCR result will barely be affected. A merge between two (not subsequent) paragraphs of two different columns is regarded as non-allowable, because an OCR engine is likely to produce results with a confused text flow. To determine the allowable/non-allowable data accurately, reading order, relative region position, reading direction and orientation are taken into account.

The second significance is user-defined and reflects the scenario the evaluation is intended for. For instance, in a table-of-contents recovery application the page number and heading regions are most important whereas graphic regions can be ignored completely. The significances are expressed by a set of weights, referred to as evaluation profile (dealt with in the next section).

For more realistic results, the errors are also quantified by the involved area. This way, a small missed region has less influence on the overall result than a miss of a whole paragraph for instance. The area based errors can be calculated in two ways: either by using the whole involved area or only the foreground area (combined area of foreground pixels). The latter has the advantage that the result is independent of the shape of the regions. Different segmentation methods may deliver regions with different sizes and margins (more loosely or closely wrapped around the objects of interest), still marking the region correctly. By using the foreground area only, these differences are ignored.

To take into account applications in which all regions have the same importance (regardless of their size), count based errors are calculated in addition to the described area based errors. The count based metric handles regions involved in errors as units (one missed region counts as one, a region split into three regions counts as three etc.).

For comparative evaluation, the weighted errors are combined to overall error and success rates by using a non-linear function in order to maximise contrast and allow an open scale (due to the nature of the errors and weighting).

D. Evaluation Scenarios and Profiles

An evaluation profile is a set of weights and settings representing a specific evaluation scenario in the context of a document analysis application. There are two different types of weights:

- Region type weights reflect the overall influence of specific types of regions, such as image, table and text (further divided into sub-types, e.g. paragraph, page number, etc.)
- Error type weights reflect the importance of the evaluated segmentation and classification errors (merge, split, miss, false detection, misclassification and reading order). The weights are further divided into region types (equal to the aforementioned weights). That way, scenarios can be described in great detail. It is for instance possible to individually set the weight for the merge of a text region representing the page number with a text region representing a heading.

Where applicable, weights are sub-divided into allowable and non-allowable weights. The evaluation raw data only contains an allowable flag, indicating if a merge or split was allowable according to the reading order and geometric context. The actual significance of allowable and non-allowable is defined by corresponding values.

In collaboration with major European libraries (in the scope of IMPACT) several profile presets have been developed for frequently used evaluation scenarios, including the following:

- General recognition: A scenario for common recognition tasks, including all error and region types with finely balanced weights.
- Full text recognition: A scenario specialized on text recognition tasks. Only text regions are evaluated.
- Text indexing: A scenario for indexing or keyword extraction applications. The focus lies on text regions only and errors like merge and split are considered less significant as the text flow is of less importance.
- Images, Graphics and Charts: A scenario to evaluate the performance of image and graphic detection methods. All other region types are ignored.

E. Result Format

A new XML schema has been developed for the evaluation framework. It is used to save evaluation profiles, complete records of evaluation results and additional metadata. It has also been included in the PAGE Format Framework. Evaluation results are always stored together with the profile in whose context they were produced. This is a benefit for applications in research as well as in production environments, since results are reproducible and extendable. A new evaluation task can be carried out by simply selecting an existing evaluation result as profile source.

IV. TOOLS

The framework is equipped with a command line tool and a graphical interface. The command line tool is intended for batch processing and integration into other systems (e.g. via web services, as already done in the context of IMPACT). The results are stored in the aforementioned

XML format and can additionally be output as comma separated values (CSV) for automated processing of larger datasets. The graphical tool aims at evaluating selected documents and inspecting the results. It can also be used to create or modify profiles for evaluation scenarios. Due to the very high level of detail the profiles offer (over 800 weights), they are presented hierarchically in form of a tree widget. This allows the modification of weights at different depths. If for instance merge errors as a whole need to be adjusted, this can be achieved by simply using a slider control on the highest level, changing all weights in lower levels at once. If however the single weight for a merge between a paragraph and a heading is to be changed, this can be done by expanding the tree accordingly and adapting the slider at the appropriate level. Fig. 2 illustrates the creation of an evaluation profile.

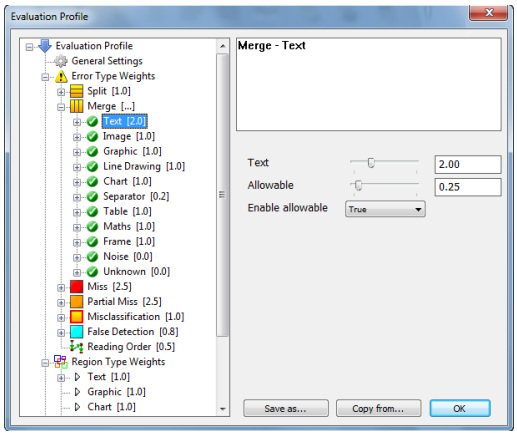

Figure 2. Creation of evaluation profiles.

Evaluation results are presented interactively in several ways (see fig. 3):

- Graphical: Transparent, colour and pattern coded overlay on the document image.
- Tree-like: Expandable tree, structured hierarchically by page element level (region, text line, word and glyph), error type, statistics and performance.
- Detailed report: Error and performance information.

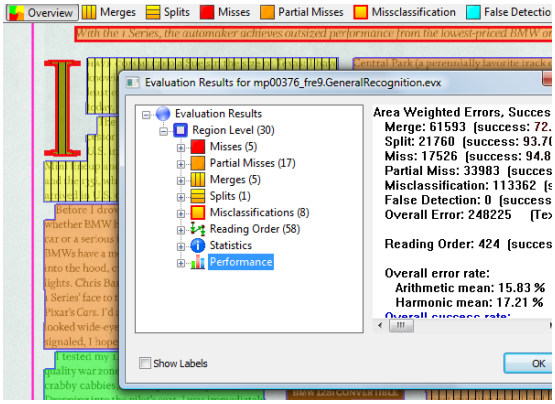

Figure 3. Inspecting evaluation results with the graphical user interface.

While the graphical presentation provides an overview of the general performance, the tree-like and textual report views offer easy-to-navigate in-depth information. Errors can be traced back to their source (regions) by simply selecting them (the involved page elements are then highlighted within the document image). The user interface also includes features such as dialogues and overlays for regions and reading order. Furthermore, the evaluation results can be fully saved and loaded using the previously mentioned XML format.

V. Experiments and Results

The new framework has been used to re-evaluate the results of the ICDAR2009 Page Segmentation Competition, confirming the evaluation results of the old system. Although the absolute performance values were slightly different (due to different weighting), the tendencies stayed the same.

The average processing time per document was 2.1 seconds (on a system with Intel CPU at 2.4 GHz, 3 GB RAM). The average number of regions per document is 25 for the ground truth and 19 for the segmentation results. The document images have a size of about 7 mega pixels (e.g. 2400*3000px). The previous version of the tool, in comparison, has a runtime of 6.1 seconds per document.

To investigate the influence of evaluation scenarios, segmentation results of the state-of-the-art system ABBYY FineReader Engine 9.0 (as implemented in the IMPACT tools framework) have been evaluated using four different evaluation profiles. The tested dataset comprises 23 contemporary documents also used in the ICDAR2009 Page Segmentation Competition. It consists of scanned magazine and journal pages with complex as well as simple layouts. Fig. 4 shows the average performance of the layout analysis system for each applied scenario. It can be observed that the total success rate varies considerably depending on the employed scenario (and thus on the intended use of the tool). FineReader performs best in text oriented scenarios. This is understandable, as text recognition is the main focus of the system.

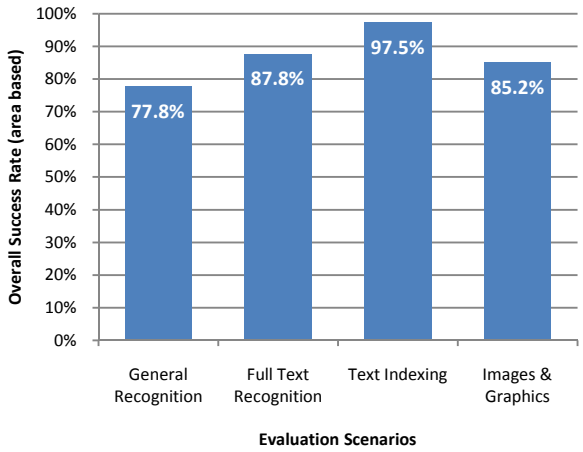

Figure 4. Average success rates for four different evaluation scenarios.

Fig. 5 shows the error rates for each error type of the PRImA measure, exemplarily for the evaluation results using the general recognition scenario. The numbers suggest that misclassification errors are the main problem of the method, after merge and miss errors. False detection on the other hand is practically non-existent.

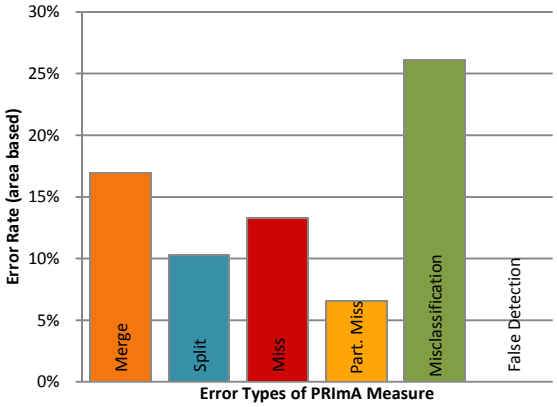

Figure 5. PRImA measure error rates for the general recognition scenario.

VI. Conclusion and Future Work

It has been shown that the presented framework is a powerful instrument for performance evaluation, meeting demands for efficiency, accuracy, profundity and flexibility. The key for the efficiency lies in the usage of the geometry based interval representations of polygonal region outlines. Accuracy is achieved by using arbitrary polygons. The use of an elaborate set of measures and the level of detail of the data being recorded gives a very profound insight into the layout analysis methods being examined. By enabling users to define evaluation profiles, a highly customizable environment is provided, fit for coping with a wide range of evaluation scenarios. Comparative large scale evaluation can be carried out by using the command line tool. A graphical user interface allows an efficient in-depth analysis of small datasets. Results are stored in XML format, which allows a straightforward integration into most infrastructures. Always storing the applied evaluation profile together with the results helps making experiments reproducible and extendable.

The framework is part of a complete performance analysis infrastructure and is being employed in major projects as IMPACT and the ICDAR Page Segmentation Competition. The potential of the system has been shown by evaluating the results of a state-of-the-art layout analysis system for a realistic dataset.

Plans for future development comprise the support of layers and nested regions (already part of the ground truth and segmentation result XML schema) as well as the integration of OCR text evaluation modules. The latter has already been done on glyph level in an experimental manner. Furthermore, in collaboration with major European libraries, more ready-made and finely balanced evaluation profiles are going to be developed for the most common scenarios.

References

[1] A. Antonacopoulos, B. Gatos and D. Bridson, "ICDAR2005 Page Segmentation Competition", Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005), Seoul, South Korea, August 29–September 1, 2005, pp. 75–79.

[2] A. Antonacopoulos, S. Pletschacher, D. Bridson and C. Papadopoulos, "ICDAR2009 Page Segmentation Competition", Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), Barcelona, Spain, July 26-29, 2009, pp.1370-1374.

[3] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated evaluation of OCR zoning" IEEE Transactions on Pattern Analysis and Machine Intelligences, Vol. 17 (1995), pp. 86–90.

[4] A. Antonacopoulos and A. Brough, "Methodology for Flexible and Efficient Analysis of the Performance of Page Segmentation Algorithms", Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR1999), Bangalore, India, September 20–22, 1999, pp. 451–454.

[5] M. Thulke, V. Märgner and A. Dengel, "A General Approach to Quality Evaluation of Document Segmentation Results", Proc. of the 3rd IAPR Workshop on Document Analysis Systems (DAS98), Nagano, Japan, Nov. 4–6, 1998, Springer LNCS (1655), pp 43–57.

[6] S. Mao and T. Kanungo, "Software Architecture of PSET: A Page Segmentation Evaluation Toolkit" International Journal of Document Analysis and Recognition, Vol. 4 (2002), pp. 205–217.

[7] A.K. Das, S.K. Saha and B. Chanda, "An empirical measure of the performance of a document image segmentation algorithm" International Journal of Document Analysis and Recognition, Vol. 4 (2002), pp. 183–190.

[8] A. Antonacopoulos, F. Coenen, "Region Description and Comparative Analysis using a Tesseral Representation", Proc. of the 5th International Conference on Document Analysis and Recognition (ICDAR1999), Bangalore, India, Sep. 20–22, 1999, pp. 193–196.

[9] B. Yanikoglu and L. Vincent, "Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation" Pattern Recognition, Vol. 31 (1998), pp. 1191–1204.

[10] F. Shafait, D. Keysers and T.M. Breuel, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images", Proc. of the 18th International Conference on Pattern Recognition (ICPR2006), Hong Kong, China, Aug. 20–24, 2006, pp. 872–875.

[11] J. Kanai, "Automated Performance Evaluation of Document Image-Analysis Systems: Issues and Practice" International Journal of Imaging Systems and Technology, Vol. 7 (1996), pp. 363–369.

[12] A. Antonacopoulos and D. Bridson, "Performance Analysis Framework for Layout Analysis Methods" Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR2007), Curitiba, Brazil, September 2007, IEEE Computer Society Press, pp. 1258-1262.

[13] S. Pletschacher, A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.

[14] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), IEEE Computer Society, Los Alamitos, USA, pp.296-300.

[15] A. Antonacopoulos and R.T. Ritchings, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles", Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR95), Montreal, Canada, August 14–15, 1995, pp. 1132–1135.

[16] D. Bridson and A. Antonacopoulos, "A Geometric Approach for Accurate and Efficient Performance Evaluation of Layout Analysis Methods", Proceedings of the 19th International Conference on Pattern Recognition (ICPR2008), Tampa, Florida, USA, December 7-11, 2008, IEEE-CS Press.