2013 12th International Conference on Document Analysis and Recognition

# The Significance of Reading Order in Document Recognition and its Evaluation†

C. Clausner, S. Pletschacher and A. Antonacopoulos
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*—Reading order detection and representation is an important task in many digitisation scenarios involving the preservation of the logical structure of a document. The corresponding need for the evaluation of reading order results generated by layout analysis methods poses a particular challenge due to potential deviations between ground truth and actually detected segmentation of the page. To this end a novel evaluation approach that responds to this problem by incorporating region correspondence analysis is proposed. Furthermore, a sophisticated reading order representation scheme is presented and used by the system allowing the grouping of objects with ordered and/or unordered relations. This is a typical requirement for documents with complex layouts such as magazines and newspapers. The evaluation method has been validated using the results of two state-of-the-art OCR / layout analysis systems and a basic top-to-bottom reading order detection algorithm applied on representative samples from the PRImA contemporary and the IMPACT historical document datasets.**

***Keywords—document layout analysis; reading order detection; reading order evaluation; performance evaluation; document structure;***

## I. INTRODUCTION

Layout analysis is an active topic both in terms of research and in practice due to the significant number of ongoing digitisation efforts of libraries and other organisations aiming to make accessible the world heritage of printed documents. Besides segmentation and classification of layout elements, further logical information is required for many applications. The reading order describes the sequence in which to address textual elements and, therefore, is a key requirement with regard to a document's logical structure. This information is crucial, for instance, for conversion tasks needing to preserve the original text flow (e-books, PDF etc.)

Whereas several approaches to evaluate segmentation and classification methods have been reported in literature, evaluation systems that attempt to also evaluate the reading order (for example [1]) are rare and incomplete. In particular, the problem of region correspondence, caused by variances between ground truth and actual segmentation, is not being addressed. Malerba et al [2] describe a graph-based method for partially ordered elements, but do not address the region correspondence problem either.

† This work has been part funded through the EU Competitiveness and Innovation Framework Programme grant Europeana Newspapers (Ref. 297380).

Reading order, even though the term might suggest differently, is not restricted to a purely sequentially ordered list. While a sequence is usually sufficient for simple layouts such as book pages, more complex layouts (e.g. magazine and newspaper pages) require a more powerful description.

Most existing layout description formats and analysis systems ([1][3][4][5]) only support partial reading order (as multiple 'chains' of regions), but not the comprehensive concept of more complex grouping. Some layout analysis systems (e.g. [6]) use additional metadata description formats (e.g. METS [7]) to describe the reading order (still relatively simply but possibly spanning multiple pages).

The PAGE (Page Analysis and Ground Truth Elements) format [8] offers high flexibility by allowing for groups of ordered or unordered elements which can also be nested. The reading order can therefore be interpreted as tree structure with nodes representing groups and regions as leaf elements. Fig. 1 shows an example document with complex reading order. Fig. 2 shows the same document with a reading order result from a layout analysis system.

The field of document structure recognition (see for example [9]) has certain parallels to reading order detection. However, it focuses more on finding and recreating links between functional elements (such as the table of contents and chapter headings). The structure is commonly analysed for whole documents (for instance to retrieve the table of contents of a book) and not for single pages. Even though this may be the intent for reading order detection as well, the intra-page problems have to be solved first, which is the scope of this work.

The advantage of incorporating region correspondences between ground truth and segmentation result for reading order evaluation is that ground truth has to be created only once. Vice versa, approaches based solely on reading order have the disadvantage that the correct order (for comparison) has to be marked manually for each new segmentation result first. Alternatively, the quality of reading order can be assessed indirectly by evaluating results of optical character recognition (OCR), if available (see for instance [10]). Such an approach, however, lacks precision and in-depth information (results are correlated with the performance of the OCR method and errors cannot be located). More fundamentally, it requires recognised text of reasonable quality in the first place (this may not always be possible, especially for historical documents or currently unsupported languages). Yalnitz and Manmatha propose a method that

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.141

aligns ground truth text and OCR results using unique words [11]. That way, missing chunks of text do not cause the evaluation to break down and the use for indirect evaluation of reading order becomes therefore more viable. The inherent problems of not being able to pinpoint the source of errors and the need for recognised text remain, however.

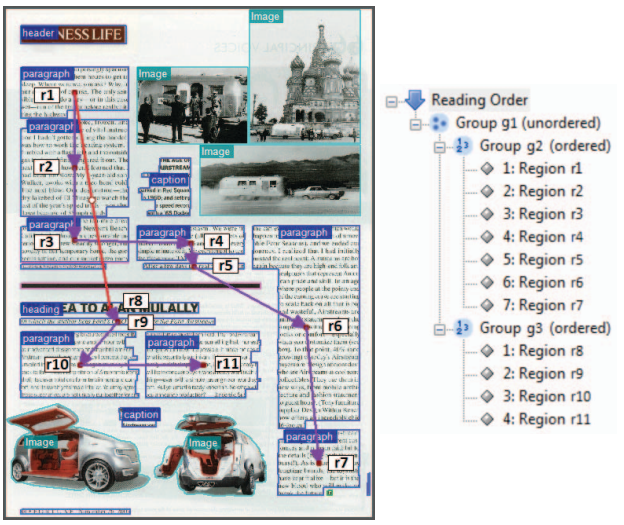



Figure 1. Reading order ground truth for a magazine page. There is no intrinsic information in which order to read the two articles, hence they are assigned to an unordered group. The paragraphs of each article are described by a sequence (ordered group). Headers, footers and captions are not part of the reading order (in accordance with the guidelines for this dataset).

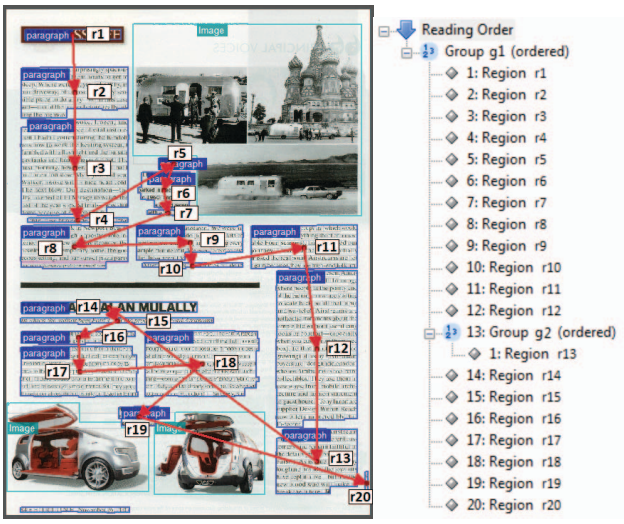



Figure 2. Automatically detected reading order. The segmented regions do not match the ground truth regions (see Fig 1), complicating a direct comparison of the corresponding reading order trees.

## II. Evaluation Method

Even though the description of the reading order has the form of a tree structure, a direct comparison of the ground truth reading order tree and the detection result tree is not possible because there is no one-to-one relation between the elements of both layouts. This is due to ambiguous interpretations or errors in segmentation. Fig. 3 illustrates this. To evaluate the relation between the segmentation result regions S1 and S2, two relations from the ground truth are involved because two ground truth regions (G1 and G2) overlap S1.

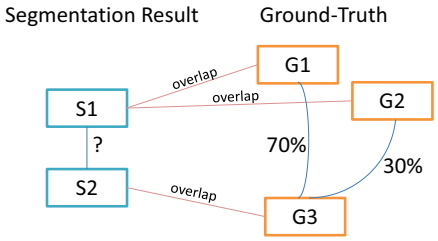
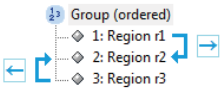
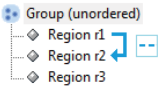
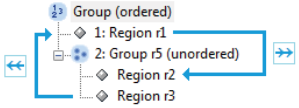
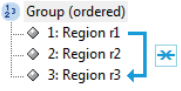
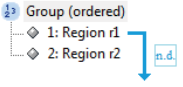



Figure 3. Example for composite reading order relation due to differences in the segmentation (percentages denoting penalty weights based on relative region overlap).

The proposed method compares the relation of each pair of regions of the layout analysis result against the set of relations extracted from the reading order ground truth. This set of relations is determined by the region correspondence between segmentation result and ground truth.

The possible relation types between two layout regions are defined in Table I.

TABLE I. Reading Order Relation Types

| | | |
|---|---|---|
| → | Direct successor |  |
| ← | Direct predecessor | |
| -- | Fully unordered relation (e.g. both in same unordered group) |  |
| →→ | Somewhere before (but unordered group involved) |  |
| ←← | Somewhere after (but unordered group involved) | |
| -x- | Neither direct nor unordered relation |  |
| n.d. | Relation not defined (one or both regions not in reading order tree) |  |

The relation between two regions is calculated by following the paths from the closest common ancestor group of both regions to their position in the reading order tree. Starting with the set of all relation types (direct, unordered, etc.), impossible types are sorted out progressively while traversing and comparing the two paths.

Based on the reading order from the layout analysis result this process delivers exactly one type of relationship as defined above. To find the ground truth relation between two segmentation result regions, the relations between all corresponding ground truth regions have to be determined and weighted with the relative overlap area. The result is a set of (possibly different) elementary relations forming a composite relation.

To calculate the penalty for a single relation from the layout analysis result it is compared to each elementary relation from the corresponding composite ground truth relation. Penalty values are retrieved from a matrix that has an entry for each possible combination of relations. Fig. 4 shows this matrix with proposed penalties which lead to reasonable contrast in the success rate. For tuning to specific use scenarios, however, deviating penalties can be defined notwithstanding the general evaluation approach.

| Result \ Ground Truth | → | ← | -- | -x- | n.d. | →→ | ←← |
|---|---|---|---|---|---|---|---|
| → | 0 | 40 | 10 | 20 | 0 | 0 | 10 |
| ← | 40 | 0 | 10 | 20 | 0 | 10 | 0 |
| -- | 20 | 20 | 0 | 10 | 0 | 10 | 10 |
| -x- | 20 | 20 | 10 | 0 | 0 | 10 | 10 |
| n.d. | 20 | 20 | 10 | 0 | 0 | 10 | 10 |
| →→ | 0 | 20 | 5 | 5 | 0 | 0 | 10 |
| ←← | 20 | 0 | 5 | 5 | 0 | 10 | 0 |

Figure 4. Penalty matrix. The columns represent the relations for the ground truth and the rows represent the relations for an analysis result. A single cell denotes the penalty for a misclassified relation between two layout regions.

Furthermore, each penalty is weighted according to the previously recorded relative overlap area of the involved regions.

The overall reading order error is the sum of all weighted penalties. Since the maximum of this value depends on the number of layout regions of a document, it is favourable to calculate a relative error or success measure in the form of a percentage. This can be achieved by relating the error value to the highest possible error value. Due to the unconstrained nature of layout analysis results a definitive maximum cannot be determined. There is for instance no limit to the number of overlapping/stacked regions. Instead, a non-linear success function (1) is used which has a parameter ($e_{50}$) representing an error value that corresponds to a success rate of 50%. Given an error value $e$ the success $s$ is defined as follows:

$$s = \frac{1}{e * \frac{1}{e_{50}} + 1} \tag{1}$$

The parameter $e_{50}$ is calculated using the maximum of a single penalty ($p_{max}=40$, according to the proposed penalty matrix) and the number of text regions in the ground truth document ($n_{GT}$) divided by two (since $e_{50}$ represents 50% success/error):

$$e_{50} = p_{max} * n_{GT}/2 \tag{2}$$

## III. Experiments and Results

Experiments have been carried out on a diverse set of 80 historical and contemporary documents, divided into four subsets with 20 documents each:

- single column book pages
- two column technical article pages
- magazine pages with complex layout (see Fig. 1)
- newspaper pages

The documents were chosen from the IMPACT Dataset [12] and the PRImA Contemporary Dataset [13]. Layout ground truth in PAGE format was readily available, only the reading order had to be annotated in some cases. Both the layout and the reading order ground truth have been produced using *Aletheia* [14], a ground-truthing and correction tool. Figure 5 shows examples for a book page, an article page and a newspaper page.

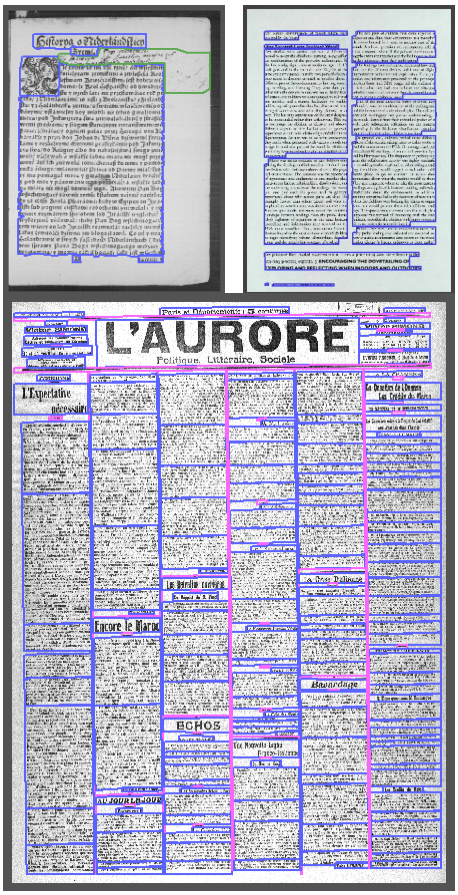



Figure 5. Example documents (layout regions highlighted) for single column books (top left), two column articles (top right) and newspapers (bottom).

For completeness it should be noted that, according to the ground-truthing guidelines which where specified in the IMPACT project, certain types of text were excluded from the reading order as they represent metadata or additional information that is usually not being read in any particular

order. These types are: page number, page header/footer, captions, footnotes, marginalia and signature marks. Alternatively, such elements could be placed inside unordered groups.

Results of the state-of-the-art layout analysis systems ABBYY FineReader Engine (version 10) and Tesseract (version 3.02) [15] were compared against a basic top-to-bottom approach (sequential reading order by sorting all text regions according to their vertical position). Both FineReader and Tesseract were interfaced directly through their API and the analysis results were exported to PAGE format. Text regions were excluded according to the same rules that apply to the ground truth reading order (where sufficient type information was available). In the case of the open source OCR engine Tesseract, the reading order detection has also been made easily accessible by integration into the aforementioned tool *Aletheia* which is publicly available.

Figure 6 shows the reading order success rate for FineReader (FRE 10), Tesseract, the top-to-bottom approach and for the trivial result of no detected reading order (see Table II for detailed values). Focusing on the overall result for all 80 documents, it can be observed that the ranking of the approaches is as it was to be expected according to the different levels of method sophistication (state-of-the-art systems best, no reading order worst).

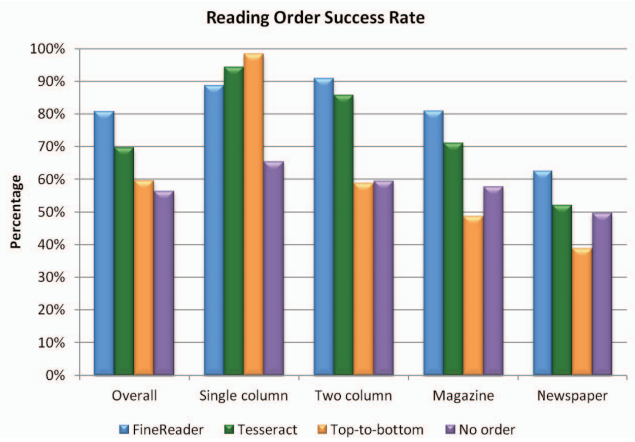



Figure 6. Reading order evaluation results grouped by subset.

The results per subset however, differ significantly. In general it can be stated that, in all cases, FineReader and Tesseract are superior to having no detected reading order (bearing in mind that an upside-down reading order result can be considered worse than no result). Possibly unexpected is that the basic top-to-bottom approach outperforms both state-of-the-art systems for the single column book pages. This can be explained by the fact that the quality of the segmentation result influences the evaluation of the reading order to some extent (see Fig. 7). The documents without reading order are based on the ideal segmentation (ground truth). FineReader and Tesseract on the other hand, perform their own layout analysis and segmentation errors are to be expected. For two-column layouts, however, it can be seen that the state-of-the-art systems clearly outperform the others.

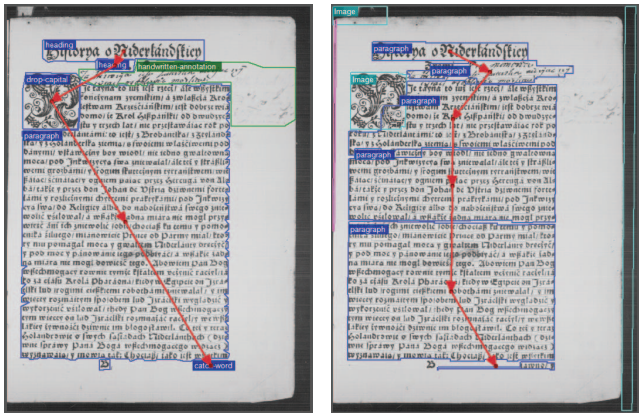


Figure 7. Example document with one-column layout. Left: Ground truth; Right: Tesseract result.

In general, the results confirm the natural assumption that the more complex the layout of a document is, the more difficult is the task of reading order detection. For very complex layouts, such as newspaper pages, the reading order success of Tesseract is almost as low as the result for no reading order. Moreover, the top-to-bottom approach, as expected, should only be applied to one-column layouts (where it actually achieved the best result among all methods).

On the whole it can be seen that in cases where the reading order detection approach is not suited for the layout complexity, it is preferable to deliver a blank result. That way, observers will not mistakenly rely on a wrong result and the task can be left for better detection methods in future.

TABLE II. READING ORDER EVALUATION RESULTS

| | Dataset | | | | |
|---|---|---|---|---|---|
| Method | Overall | Single column | Two column | Magazine | News-paper |
| FRE 10 | **80.8%** | 88.7% | **90.9%** | **81.0%** | **62.7%** |
| Tesseract | 69.8% | 94.4% | 85.8% | 71.2% | 52.2% |
| Top-to-bottom | 59.7% | **98.5%** | 58.9% | 48.7% | 38.9% |
| No order | 56.3% | 65.4% | 59.6% | 57.8% | 49.7% |

## IV. CONCLUDING REMARKS

In this work the significance of reading order detection has been highlighted and a novel evaluation approach has been presented for the evaluation of methods producing reading order results. The evaluation approach is capable of coping with complex relations between layout regions, including nested groups of ordered and/or unordered page elements. The problem of discrepancies in region segmentation between ground truth and analysis result is overcome by using composite reading order relations.

The evaluation approach has been validated by measuring the quality of the detected reading order produced by the state-of-the-art layout analysis systems ABBYY FineReader and Tesseract combined with a comparison against a

straightforward top-to-bottom approach and no reading order at all. Since the evaluation results are in line with the capability of each method for the given scenario, it can be concluded that the proposed metric is appropriate for the given task.

The presented method has been implemented as part of a comprehensive range of layout evaluation tools (see [16] for an overview) and is publicly available.

Future work will include the implementation of a configurable penalty matrix via evaluation profiles. Furthermore, the influence of the segmentation quality on the reading order evaluation could be decreased by incorporating the concept of allowable merge and split errors, which attract lower penalties than 'full' errors (as described in [16]).

## REFERENCES


[1] 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.

[2] D. Malerba, M. Ceci and M. Berardi, "Machine Learning for Reading Order Detection in Document Image Understanding", Machine Learning in Document Analysis and Recognition Studies in Computational Intelligence Volume 90, 2008, pp 45-69

[3] C. H. Lee and T. Kanungo, "The Architecture of TRUEVIZ: A GroundTRUth/Metadata Editing and VISualiZing Toolkit," Pattern Recognition, vol. 36, no. 3, pp. 811-825, 2003..

[4] ALTO (Analyzed Layout and Text Object) XML Schema, http://www.loc.gov/standards/alto/techcenter/structure.php

[5] Abbyy FineReader XML schema version 10.1, http://www.abbyy.com/FineReader_xml/FineReader10-schema-v1.xml

[6] CCS docWorks, http://www.ccs-digital.info/en/products/docworks

[7] METS (Metadata Encoding and Transmission Standard), http://www.loc.gov/standards/mets/METSOverview.v2.html

[8] 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.

[9] A. Doucet, G. Kazai, J.-L. Meunier, "ICDAR 2011 Book Structure Extraction Competition", Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, September 18-21, 2011, pp. 1501-1505

[10] S. V. Rice, "Measuring the Accuracy of Page-Reading Systems", PH.D. Dissertation, UNLV, Las Vegas

[11] I. Z. Yalniz, R. Manmatha, "A fast alignment scheme for automatic OCR evaluation of books", Proceedings of 2011 International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, September 2011, pp. 754–758

[12] IMPACT (IMProving ACcess to Text), http://www.impact-project.eu/, http://www.digitisation.eu/

[13] 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), Barcelona, Spain, July 2009, pp.296-300.

[14] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments", *Proc. ICDAR2011*, Beijing, China, September 2011, pp. 48-52

[15] Tesseract OCR Engine, http://code.google.com/p/tesseract-ocr/

[16] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods", *Proc. ICDAR2011*, Beijing, China, September 2011, pp. 1404-1408.