2011 International Conference on Document Analysis and Recognition

Historical Document Layout Analysis Competition†

A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher
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***Abstract*—This paper presents an objective comparative evaluation of layout analysis methods for scanned historical documents. It describes the competition (modus operandi, dataset and evaluation methodology) held in the context of ICDAR2011 and the International Workshop on Historical Document Imaging and Processing (HIP2011), presenting the results of the evaluation of four submitted methods. A commercial state-of-the-art system is also evaluated for comparison. Two scenarios are reported in this paper, one evaluating the ability of methods to accurately segment regions and the other evaluating the whole pipeline of segmentation and region classification (with a text extraction goal). The results indicate that there is a convergence to a certain methodology with some variations in the approach. However, there is still a considerable need to develop robust methods that deal with the idiosyncrasies of historical documents.**

Keywords-layout analysis; performance evaluation; page segmentation; region classification; datasets; historical documents

I. INTRODUCTION

Layout Analysis is the first major step in a Document Image Analysis workflow where, after Image Enhancement, a descriptive 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 and Region Classification is crucial as the resulting representation forms the basis for all subsequent analysis and recognition processes.

Layout Analysis is one of the most well-researched fields in Document Image Analysis, yet new methods continue to be reported in the literature, indicating 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 early 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. The 5th edition of the ICDAR Page Segmentation competition series (ICDAR2009) [3] incorporated significant additions and enhancements. First, that competition marked a radical departure from the previous evaluation methodology. A new evaluation scheme was introduced, allowing for higher level goal-oriented evaluation and much more detailed region comparison. In addition, the dataset used was selected from a new PRImA contemporary dataset [4] that contains different instances of realistic documents.

This 6th edition is based on the same established principles but with considerable changes and improvements. First, the complete Layout Analysis workflow is evaluated (both Page Segmentation and Region Classification). Second, focus has shifted on to historical documents to reflect the significant need to identify robust and accurate methods for the many current and future library digitisation initiatives. Appropriately, this edition of the competition is co-sponsored by both ICDAR2011 and HIP2011 (International Workshop on Historical Document Imaging and Processing). Finally, the evaluation system has been refined both in terms of increased options for detailed definition of penalties and weights and in reflecting the overall impact of methods in real world application scenarios [5].

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.

† This work has been funded through the EU 7th Framework Programme grant IMPACT (Ref. 215064)

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.301

II. THE COMPETITION

The Historical Document Layout Analysis competition 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). Delving deeper, 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 a placement of the participating methods into context by comparing them to a leading commercial method currently used by digitisation service providers.

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). The *Aletheia* [6] ground-truthing system (which can also be used as a viewer for results) and code for outputting results in the PAGE format [7] (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 a pre-defined format. The organisers then verified the submitted results and evaluated them.

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 a dataset with ground truth [4] and making it available to all researchers. In comparison, representative datasets of historical documents are even more difficult to collect (from different libraries) and to ground truth (due to the nature and variety of the texts).

A comprehensive dataset of historical document images is being created as part of the IMPACT project [8]. At the time of writing, the dataset contains approximately 700,000 images (with associated metadata) from 14 different content holders, including most national and major libraries in Europe. This dataset is being created having in mind not only the conditions and artefacts of historical documents that affect document analysis, but also the needs and priorities of the libraries, in terms of what types of documents (representative of their holdings) dominate their digitisation plans. The complete dataset consists of printed documents of various types, such as books, newspapers, journals and legal documents, in 17 different languages and 11 scripts. With regard to the age of the content available, the documents range from the 17th to the early 20th century.

The unique value of this dataset though is not only in how well it represents the major libraries' collections and the occurrence and distribution of the various issues found in historical documents, but also the availability of a considerable volume of detailed ground truth. At the time of writing, 15,000 images selected from different libraries have been ground truthed at the level of regions (equivalent to paragraphs, illustrations, separators etc.), with the aim being to have up to 25,000 in total. In addition to the description of region outlines, the text contained in each (textual) region has been re-keyed under strict rules, preserving typographic conventions, including, abbreviations, ligatures etc.

Figure 1. Sample evaluation set images (not shown to scale).

For the purpose of this competition, 100 images were selected as a representative sample ensuring the presence of different document types and ages and the issues affecting layout analysis are adequately covered. Such issues include dense printing (minimal spacing), irregular spacing, varying text column widths, marginal notes etc. as can be seen in the examples in Fig. 1. It is worth noting that the images for this competition were selected so as not to suffer from significant artefacts (e.g. severe page curl or arbitrary warping) that would require a separate image enhancement step before layout analysis (this competition relates to layout analysis and not to an end-to-end workflow).

The ground truth is stored in the XML format which is part of the PAGE (Page Analysis and Ground truth Elements) representation framework [7]. 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. Sample images with ground truth description can be seen in Fig. 2.

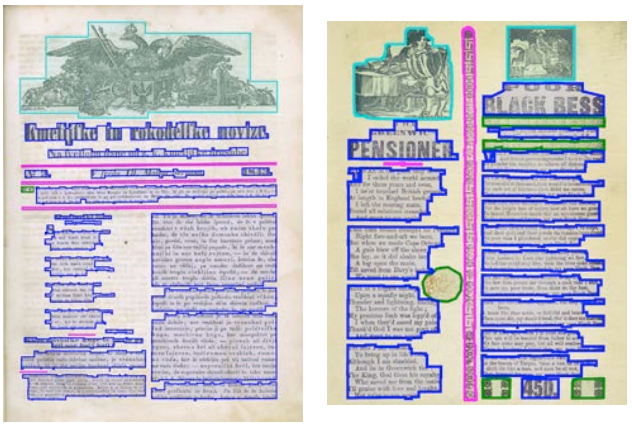

Figure 2. Images from the evaluation set showing the region outlines (blue: text, magenta: separator, green: graphic, cyan: image).

IV. Performance Evaluation

The performance analysis method used for this competition 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 [5], 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 application 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 (or not completely) overlapped by a segmentation 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 another type.

Based on the above, the segmentation and classification errors are *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.

Having this raw data, the errors are then *qualified* by their significance. There are 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 of applying OCR on the merged region will not violate the reading order. On the contrary, a merger between two paragraphs across two different columns of text is regarded as non-allowable, because the reading order will be violated in the OCR result. 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 application 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* [5]. For each application scenario to be evaluated there will be a corresponding evaluation profile.

Appropriately, the errors are also weighted by the size of the area affected (excluding background pixels). In this way, 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.

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

V. 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.

A. The MEPhI Method

This bottom-up component aggregation approach was submitted by Aleksey Vilkin of the Moscow Engineering Physics Institute (MEPhI), Russia. It starts by performing image pre-processing: (i) adaptive binarisation, combining local and global window (threshold calculated by Otsu method), and (ii) skew correction, by Hough transform.

Subsequently, the document is segmented into zones by first identifying text regions and filtering the rest (non-text). Connected components are identified and filtered according to size (e.g. text must be visually distinguishable, therefore very small components are filtered out) and complexity (number of extraneous components within the bounding box – characters should have very few). Words, text lines and text blocks are built by aggregating components based on their horizontal and vertical proximity.

To increase the quality of segmentation, a number of restrictions were placed on the properties of connected components and the resulting groups, when combined; the parame-

ters for those restrictions are dynamically calculated for each document and group of components.

B. The Jouve Method

This method was submitted by Michaël Fontain of Jouve, France [9], a commercial organisation specializing in digitisation services. The main principle of the method is to identify and extract regions of text by analysing connected components constrained by black and white (background) separators – the rest is filtered out as non-text.

First, the image is binarised, any skew is corrected and black page borders are removed. Subsequently, connected components are extracted and filtered according to size (very small components are filtered out). By analysing the size and spacing of the components (using global and local information), characters and words are identified. Black horizontal and vertical lines (corresponding to separators) are also identified in the size filtering step. White separators corresponding to space between columns are then identified by aggregating white rectangles aligned at the end of words and filtering out non-viable separators. With the aid of white separators, words are grouped into text lines without risking merging words belonging to different columns.

Text lines of the same height and located at the same distances are grouped to reconstitute the paragraphs. Paragraphs are finally merged in order to obtain columns guided by both the black and the white separators detected. The reading order is determined by an iterative method using vertical white streams, horizontal and vertical black separators, and a heuristic to sort boxes.

C. The Fraunhofer Method

The Fraunhofer Newspaper Segmenter – Historical Archive Edition, was submitted by Iuliu Konya of the Fraunhofer Institute for Intelligent Analysis and Information Systems at Sankt Augustin, Germany. It is a specialization of the FhG Segmenter software (as described in [3]) focusing on the processing of scans of historical documents:

Pre-processing. A basic page border removal and the selection of a local or global binarisation algorithm are performed by employing several features computed from the document scan, such as the per-area averages and standard deviations of the black-to-white ratios for glyph candidates, densities for glyph candidate connected components vs. noise components, and the dominant character size is identified from the input grayscale image.

Black separator detection. First, the quality of the horizontal and vertical separators is improved [10] before being extracted [11]. A subsequent triage of the separators is performed by using information about the dominant character size on the page.

White separator detection. Maximally empty rectangles are detected [12], but they must also satisfy certain conditions, e.g. their height must be large enough in relation to the dominant character size.

Page segmentation. A hybrid approach is applied comprising a bottom-up process [13] 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).

Text line and region extraction. Exact text lines are detected again in the raw text regions detected in the previous step using a method similar to [13]. Font characteristics (e.g. stroke width, x-height, italics) are computed for each text line and used to derive the text regions with similar properties, with the aid of a dynamic programming approach minimizing the overall font distance.

D. The EPITA Method

This method [14] was submitted by Guillaume Lazzara of EPITA, France. It is a bottom-up approach based on connected-component aggregation. First, the document is binarised using a multiscale implementation of Sauvola's algorithm [15]. Vertical and horizontal separators are then identified, removed and the document is denoised.

The remaining components are labeled and from those similar component groups, component alignments and white spaces (on their sides) are determined. These virtual delimitors associated with separators provide a structure of the different blocks in the document. Using this information, component groups are merged to create text lines.

Subsequently, lines are linked into text regions. Text indentations, spaces between adjacent lines and text line features are then analysed in order to split regions into paragraphs. Paragraphs overlapping significantly are also merged together. Among the part of the documents where no text has been found, the components are retrieved and considered as images. Finally, some cleanup is performed: separators detected in images, in paragraph and in document borders are removed, false positive text areas are removed in images and borders and small images included in text areas are considered as drop capitals.

VI. Results

Evaluation results for the above methods are presented in this section in the form of graphs with corresponding tables. The Layout Analysis method of a leading product, ABBYY FineReader® Engine 9 (FRE9), is also included for comparison. It must be noted that FRE9 has been tested out of the box, with no training or knowledge of the dataset.

Two profiles have been defined for the competition. 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 Fig. 3.

The second profile is basically equal to the first one except that it also includes misclassification. As the main focus lies on text, misclassification of text is weighted highest. All other misclassification weights are set to 10%. Results for this profile are shown in Fig. 4.

Finally, a breakdown of the errors made by each method is given in Fig. 5.

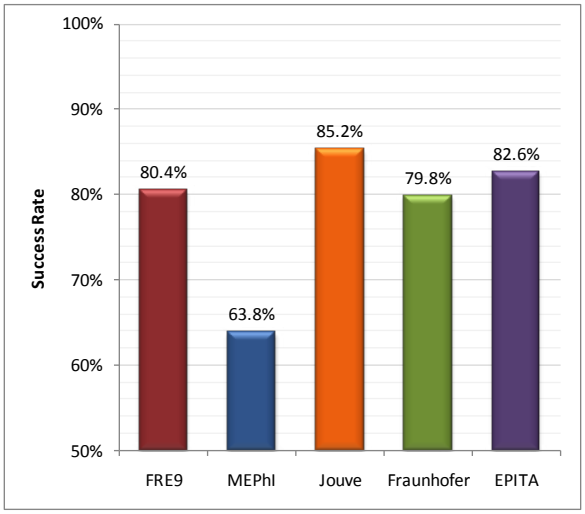

Figure 3. Results using the segmentation evaluation profile.

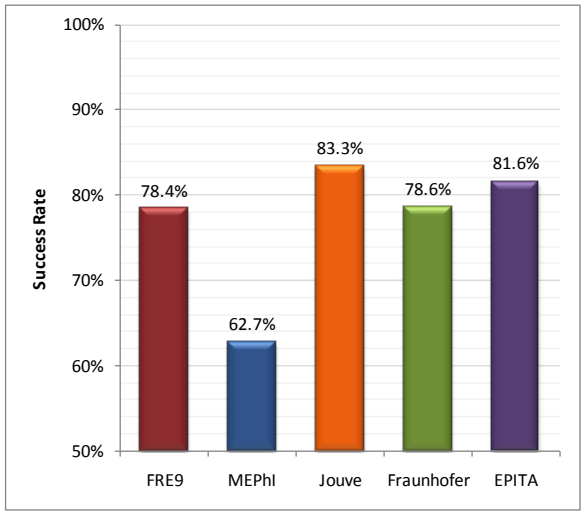

Figure 4. Results using the OCR-scenario evaluation profile.

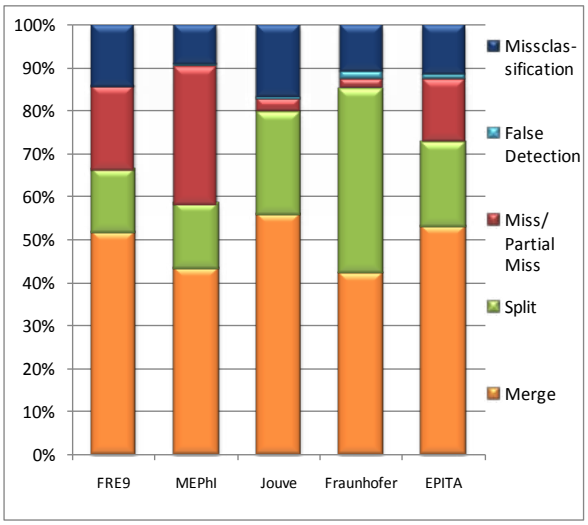

Figure 5. Breakdown of errors made by each method.

VII. Concluding Remarks

The aim of the Historical Document Layout Analysis competition was to evaluate the submitted Layout Analysis methods on a new comprehensive and extensive (in breadth and depth) printed historical document dataset, using a further refined objective performance analysis system. Two scenarios are reported in this paper, one evaluating the ability of methods to accurately segment regions and the other evaluating the whole pipeline of segmentation and region classification (with a text extraction goal). Four systems were evaluated and compared with a leading commercial product. The results show that the Jouve method has an overall advantage, although two other methods (based on similar methodology) are relatively close. It is also clear that there is still a considerable need to develop robust methods that deal with the idiosyncrasies of historical documents.

Acknowledgments

The Authors would like to acknowledge the funding of this work by the European Commission through the IMPACT project [8], and the help of Allen Fairchild in running the submitted method executables on the different hardware and software platform combinations required.

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

[6] 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.

[7] 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.

[8] IMPACT: Improving Access to Text, EU FP7 project http://www.impact-project.eu

[9] JOUVE – http://www.jouve.com

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

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

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

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

[14] Source available in our git Repository: git://git.lrde.epita.fr/olena - branch: icdar/hdlac2011 - location: scribo/src/contest/hdlac-2011

[15] Online demo: http://olena.lrde.epita.fr/SauvolaMs