2009 10th International Conference on Document Analysis and Recognition

A Realistic Dataset for Performance Evaluation of Document Layout Analysis[†]

A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher

Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford, Greater Manchester, United Kingdom
http://www.primaresearch.org

Abstract

There is a significant need for a realistic dataset on which to evaluate layout analysis methods and examine their performance in detail. This paper presents a new dataset (and the methodology used to create it) based on a wide range of contemporary documents. Strong emphasis is placed on comprehensive and detailed representation of both complex and simple layouts, and on colour originals. In-depth information is recorded both at the page and region level. Ground truth is efficiently created using a new semi-automated tool and stored in a new comprehensive XML representation, the PAGE format. The dataset can be browsed and searched via a web-based front end to the underlying database and suitable subsets (relevant to specific evaluation goals) can be selected and downloaded.

1 Introduction

The objective evaluation of the performance of layout analysis methods is invaluable for both document analysis system integrators and developers. The former need to compare existing methods and identify which method is best suited overall to a particular workflow and target document type(s). The developers on the other hand, need additional detailed information on how methods perform in specific circumstances, under certain challenging conditions and application scenarios. Particular errors and successes observed can then be taken into account in order to improve methods and publish them.

The two key prerequisites to achieving an objective evaluation are:

- a realistic and comprehensive *dataset* and
- an in-depth *evaluation methodology*.

The dataset is of fundamental importance and will be further discussed in the remainder of this paper. An in-depth evaluation framework has recently been described by the authors [1].

To highlight the importance of the dataset, it should be noted that a number of evaluation methodologies can be based on a good dataset but even the best evaluation method cannot compensate for an inadequate dataset.

There are several qualities that characterise a good dataset both in terms of content and usability. The three main desirable characteristics are:

(i) *Realistic*. The dataset must contain a representative breadth of real documents likely to be scanned in everyday situations.
(ii) *Comprehensive*. It must contain detailed information to enable in-depth evaluation.
(iii) *Flexibly structured*. It should be easy to browse, search and to select subsets with specific conditions.

Existing datasets that can be used for layout analysis performance evaluation do not fully capture the above requirements. The relatively widely-used University of Washington (UWASH) dataset [2] contains documents with simple layouts (where regions can be described by bounding rectangles), almost exclusively in the form of bilevel journal articles, most of them synthetically generated. The ISRI dataset [3] and the Medical Article Records Groundtruth (MARG) dataset [4] contain only bilevel images and similarly simple layouts, predominantly of journal articles. The MediaTeam Document Database [5] has a variety of documents in colour but concentrates on less complex layouts, also limited by the use of bounding rectangles for region representation. The UvA Document Dataset [5] contains more complex regions but concentrates predominantly on advertisement-type magazine pages which do not provide a representative sample of the documents most likely to be scanned in realistic every-day situations. Furthermore, the ground truth contained is not sufficient for full layout analysis as it is primarily oriented towards colour segmentation.

This paper presents a new dataset developed by the PRImA group which provides researchers with a wide selection of complex, contemporary documents together with accurate ground-truth and extensive metadata. This paper describes the content and format of the dataset, the ground-truthing workflow and tools which were developed specifically for this dataset as well as the web interface which allows researchers to access the dataset.

The following section describes the dataset in terms of its contents, structure, functionality and ground truth. Section 3 outlines the steps of the dataset creation workflow with a particular emphasis on ground truthing using a new semi-automated tool. Finally, the remarks of Section 4 conclude the paper.

[†] *Parts of the work presented in this paper have been supported by Google Inc. and through the EU 7th Framework Programme grant IMPACT (Ref: 215064).*

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.271

2 Dataset description

This section describes the dataset in terms of the type of documents contained, the detailed ground truth information provided and the facilities of the website which enable researchers to use the dataset according to their different evaluation goals.

2.1 Content

The dataset contains a wide variety of different document types, reflecting the various challenges in layout analysis. Particular emphasis is placed on magazines and technical journals which are likely to be the focus of digitisation efforts.

Magazine scans come from a variety of mainstream publications related to news, business and technology. The layouts of these pages contain a mixture of simple and complex layouts, including many instances of text wrapping tightly around images, varying font sizes and other characteristics which are useful to evaluate layout analysis methods on.

Figure 1. Magazine page examples.

Technical articles are sourced from several publications across a variety of disciplines, including journals and conference proceedings. Although the layouts are typically simpler than the magazine pages, with a significant proportion of layouts with purely rectangular regions, there exist a number of pages containing more complex characteristics.

Currently, the dataset contains 1240 ground-truthed images and the ratio of magazine pages to technical articles is 7 to 1. In the authors' experience this ratio is sufficiently representative of the distribution of existing publications. Figures 1-2 contain sample documents from the dataset.

It should be noted that, although the dataset concentrates largely on magazine and technical articles, it also includes examples of a large variety of other documents such as forms, bank statements and advertisements in order to provide a broader range of data to cater for researchers working in more specific application areas.

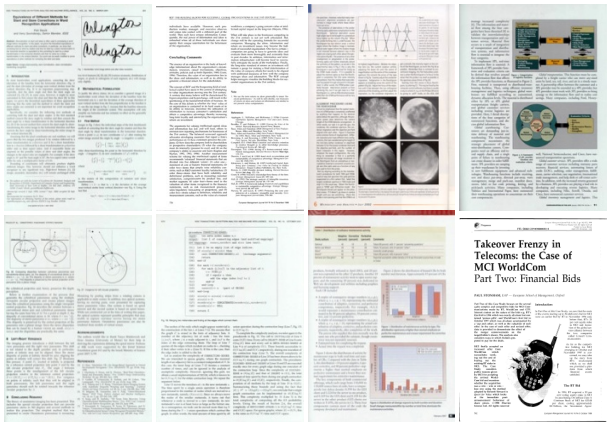

Fig. 2. Technical article examples.

2.2 Information structure

In order to enable users to efficiently search/access the dataset as well as ensuring a comprehensive representation of ground truth, the dataset is structured into two tiers. A *database* is used to store certain document-level metadata (in addition to the images and corresponding ground truth) while the *ground truth* itself is stored in a new XML-based format, first proposed in [7] and since then further extended [9].

The database contains metadata related to conditions found in documents which may be of interest to researchers in the field and also related to maintenance needs of the dataset. The following is a summary of the information stored in the database:

- *Bibliographic information* — title, publication, author, etc.
- *Imaging information* — scanner used, bit depth and resolution.
- *Layout features* — presence of images and/or graphical features, number of columns, variety of font sizes.
- *Administrative information* — copyright holder, etc.

For each page image, in addition to the above document-level metadata, there is a ground truth file which contains the layout ground truth for the page as well as a range of metadata related to layout elements (see Section 2.4 below).

2.3 Functionality

The main interface to the dataset is through the dataset website. Both browsing and search modes are supported to discovery or retrieval of documents with specific attributes. Browsing is facilitated through an overview of all documents in the dataset, divided into categories. This view displays thumbnails of all pages in each category, allowing researchers to select documents visually for the characteristics on which they are interested in evaluating a given layout analysis method.

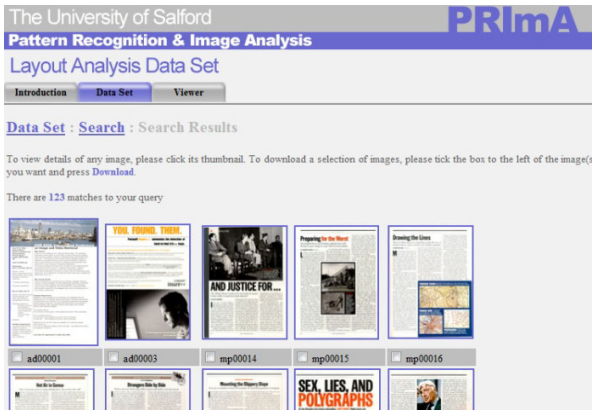

Figure 3. A sample search results page from the dataset website.

In addition to browsing, users can search through the dataset to identify sets of documents satisfying a number of different criteria. One can search by the colour depth of the document image (bi-level or true colour, say), the number of columns in the document (either a specific number or, in general, multi-column) as well as for a number of document features such as the presence of images, tables or varying fonts. This functionality allows researchers to efficiently select subsets of the dataset for focussed evaluation of methods (i.e. on the performance of a method on documents exhibiting the specific characteristics only). Figure 3 shows a sample search results page from the dataset website.

Browsing through document collections and examining individual images for certain characteristics can be time-consuming, especially due to the time it takes to transmit large original image file sizes. To facilitate more efficient examination of documents, two more levels of presentation are available, each offering progressively more information as necessary. When one clicks on a thumbnail (in search results or during browsing), an intermediate information page is displayed containing a larger thumbnail (for closer view) as well as document-level metadata (described in Section 2.2). When a user requires an even more detailed view of the document, a full-resolution preview of either the colour or bilevel images can be displayed (in high quality JPEG format for faster loading) by clicking on the document thumbnail.

The website allows researchers to download sets of one or more original – highest quality – document images (colour and/or bilevel) and/or their associated ground truth files by selecting a number of documents from a browsing list or a search results page. The selected items are bundled into a zip file on the fly for downloading.

The system also offers sufficient functionality for maintaining and expanding the dataset. A two-stage procedure is followed when adding new images and ground truth to enable external contributions as well as to ensure quality control. Researchers can upload prepared material to the server which marks it as tentative. Administrator users then check and verify the details and ground truth of all tentative material. If all information is found to be correct, images, metadata and corresponding ground truth are then committed to the dataset (released for public access).

Administrator users have also access to a variety of statistics on the dataset, such as the proportion of the different document types, which can be used for planning purposes, ensuring that the dataset remains representative.

2.4 Ground truth format

The PRImA ground-truth format for page content has evolved over a number of years (through experience gained from the ICDAR page segmentation competitions and initial versions of this dataset) and enables a most comprehensive description of digitised documents for the purpose of layout analysis. In general, it is compatible with other established page description formats such as ALTO [8] but it offers several additional (more specialised) features specifically designed to aid in the representation and evaluation of layout analysis results.

The format provides for the representation of several different region types, which may be subject to different processing in recognition systems. The most important types of region are *text*, *image*, *line drawing*, *graphic*, *table*, *chart*, *separator*, *maths*, *noise* and *frame*. In the ground truth, the highest-level textual regions correspond to paragraphs (a conscious choice as a paragraph is also a complete logical entity, as opposed to columns of text for instance). The format allows for the representation of further subdivisions of *text* regions into *text lines*, *words* and *glyphs*, in order to enable the evaluation of segmentation methods at those lower levels also.

For each region there is a description of its outline in the form of an isothetic polygon (i.e. a polygon having only horizontal and vertical edges). Such a representation enables a very accurate and efficient geometric description, especially for complex-shaped regions (there is also no efficiency loss in the case of rectangular regions – the corresponding isothetic polygon will be a rectangle). 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. Examples are groups which may contain ordered or unordered elements and may even be nested. This is important for documents with intricate logical structures like newspapers.

The format for page content ground truth itself is specified by means of a new XML Schema which is part of the PAGE (Page Analysis and Ground truth Elements) image representation framework [9]. The framework in general is capable of representing multiple types of ground truth in other stages also of the image enhancement, segmentation and OCR process and is currently being used in the creation of a new comprehensive dataset for the evaluation of all document analysis and recognition steps in historical documents.

3 Building the dataset

Efficiency and accuracy are crucial in creating a large dataset of complex documents which are ground truthed in detail. This section outlines the workflow developed for this dataset which enables data collection and ground truthing in reasonable volumes while ensuring that the quality of the result is maintained.

3.1 Workflow

To ensure that the dataset includes a sufficient number of documents with both simple and complex layouts in each of the content categories, the first step of document selection is an off-line expert-driven activity.

Once the documents have been selected, digitisation is initiated. This step can be carried out by relatively unskilled people, following strict protocols. First, all selected documents are scanned at 300 dpi and in 24-bit colour. Reasonable care is taken to align each document when placing it on the scanner so that no skew is introduced. A black card is inserted behind the page being scanned to minimise show-through. However, for efficiency reasons, the resulting images are not checked just after scanning each document but only at the end of the scanning process. While scanning a page, document-level administrative metadata is collected.

Subsequently, each image is checked using a viewer for any scanning defects. Such defects may be: missing parts of a page, excessive skew (digitally correcting it may adversely affect the quality of the image) and show-through. Any such cases are flagged and the corresponding pages are scheduled for later individual rescanning, this time using a preview facility to ensure better scanning results than in the initial attempt. If the scanning result even after this is not of a satisfying quality, the entry for this page is flagged and excluded from subsequent processing, until a suitably qualified person either repairs the problem or decides to exclude the image from the dataset.

Deskewing is performed next automatically (using the method included in the ABBYY FineReader engine) for all the images passing the scanning quality control step. This step is necessary to correct even the smallest amount of skew that may have been introduced despite careful scanning. Once this is complete, each deskewed image is opened in an editor to be manually checked (deskewing quality control) and then cropped to the size of the actual page to remove any excess page borders introduced during scanning.

All successfully scanned, deskewed and cropped images are then automatically binarised. Such bilevel images are produced for the dataset as the majority of layout analysis methods do not work on colour/greyscale images. The binarisation method used is that of ABBYY FineReader engine (chosen as a baseline quality method which aims at improved OCR results).

Images that pass the relevant binarisation quality control step are scheduled for layout ground-truthing. This step must be carried out by a more knowledgeable (in document analysis) trained operator and involves marking the region boundaries, region labelling and adding the region-level metadata. All this is performed with the ground-truthing tool described in the next section.

Once the ground truth has been produced for a given document, a final quality control step (carried out by a different operator) ensures that the final result is correct before the document is committed to the dataset.

3.2 Ground truthing

Since the ground truth is by definition the baseline layout description against which all methods will be compared, it cannot be created automatically. Its production must be of the highest possible quality and, correspondingly, the process is very labour-intensive.

In order to maximise the efficiency of the ground truthing process, a semi-automated ground-truthing tool, *Aletheia*, has been developed. For instance, in creating region outline descriptions, one of the most time-consuming aspects of ground truthing, it enables the user to quickly specify *imprecise* region boundaries using a variety of drawing modes such as using rectangles or arbitrary polygons) rather than performing detailed selection and editing operations. Such an imprecise outline can be drawn very loosely around any distinct region as long as no parts of any other regions are included. The programme then improves upon this rough outline, automatically fitting it to the region contents more pre-

cisely (the process can be likened to shrink-wrapping a plastic sheet around an object of any shape).

In the authors' considerable ground-truthing experience over the years, the above approach has proven significantly more efficient than automatically obtaining a full page segmentation first (using any of the best region segmentation methods available) and then manually correcting it.

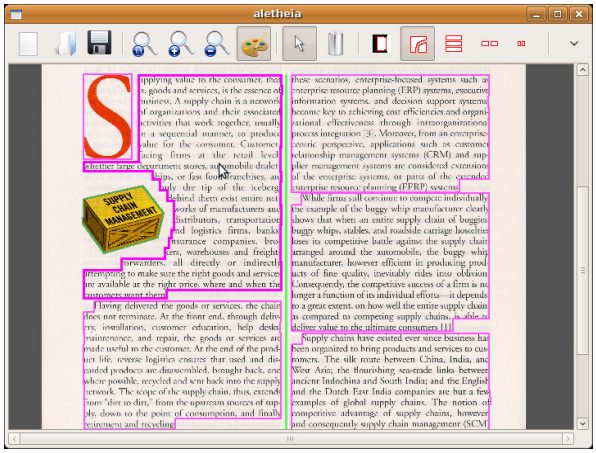

Fig. 4. Complex-shaped region ground truthing using *Aletheia*.

Outlines of page boundaries, regions (see an illustration in Fig. 4), text lines, words and glyphs can all be marked using *Aletheia*. In addition, it enables properties of each defined region to be entered as part of the metadata stored in the ground truth (see Section 2.4). Once regions (at all desired levels) have been marked on a given image and the corresponding region-level metadata has been entered, *Aletheia* produces the ground truth file in the specified format.

4 Concluding Remarks

This paper has presented a new dataset for layout analysis research as well as the process of its creation. In contrast to existing datasets, it provides comprehensive and very accurate ground truth description and associated metadata for a wide variety of layouts that reflect documents that are likely to be of wide interest to be digitised. Also in contrast to previous datasets that are distributed as concatenation of files, the new dataset can be browsed and searched through a database with a web-front end in order to examine document conditions and identify suitable subsets for more focussed evaluation.

Initial parts of previous versions of the dataset have been used in ICDAR Page Segmentation Competitions and a subset of the current dataset will be used for the ICDAR2009 Page Segmentation Competition.

The ground truth in the dataset is structured in a new format that can support multiple types of ground truth at different levels of description. The page content ground truth that is relevant to layout analysis evaluation has been created using a new semi-automated tool, *Aletheia*, that has been designed to maximise the efficiency of the region marking process while ensuring a high level of description accuracy.

Also described in this paper is an established workflow (with appropriate quality control points) for layout analysis dataset population and ground truthing. This workflow can be followed, together with the use of *Aletheia*, by other colleagues to prepare documents and ground truth to expand the dataset. The existing infrastructure allows for such expansion through additional quality control procedures.

The location is: http://dataset.primaresearch.org/

Acknowledgement

The authors would like to thank Dimosthenis Karatzas for his significant contributions to previous versions of the *Aletheia* ground-truthing tool.

References

[1] 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, pp. 1258–1262.

[2] I.T. Phillips, S. Chen, J. Ha, and R.M. Haralick, "English document database design and implementation methodology," in *Proceedings of the 2nd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, United States, April 1993, pp. 65–104.

[3] T.A. Nartker, S.V. Rice and S.E. Lumos, "Software tools and test data for research and testing of page-reading ocr systems", *Proceedings of the IS&T/SPIE 2005 International Symposium on Electronic Imaging Science and Technology*, San Jose, USA, January 2005, pp. 37–47.

[4] The National Library of Medicine, "Medical article records groundtruth," http://marg.nlm.nih.gov/, Bethesda, USA.

[5] J.J. Sauvola and H. Kauniskangas, "Mediateam document database," http://www.mediateam.oulu.fi/downloads/MTDB/, Finland, 1998.

[6] L. Todoran, M. Worring, and M. Smeulders, "The UvA color document dataset," *Int. J. of Document Analysis and Recognition*, vol. 7, no. 4, pp. 228–240, 2005.

[7] A. Antonacopoulos, D. Karatzas, and D. Bridson, "Ground truth for layout analysis performance evaluation," *Proceedings of the 7th IAPR International Workshop on Document Analysis Systems*, H. Bunke and L. Spitz, Eds. Nelson, New Zealand: Springer, February 2006, pp. 302–311.

[8] "ALTO: Analyzed Layout and Text Object, CCS Content Conversion Specialists", http://www.ccs-gmbh.com/alto/

[9] http://schema.primaresearch.org/PAGE/