Automatic Ground-truth Generation for Document Image Analysis and Understanding

Pierre Héroux, Eugen Barbu, Sébastien Adam, Éric Trupin

Automatic Ground-truth Generation for Document Image Analysis and Understanding

Pierre Héroux Eugen Barbu Sébastien Adam Éric Trupin
LITIS
Université de Rouen
76 800 Saint-Etienne du Rouvray, France
{Firstname.Lastname}@univ-rouen.fr

Abstract

Performance evaluation for document image analysis and understanding is a recurring problem. Many ground-truthed document image databases are now used to evaluate general algorithms, but these databases are less useful for the design of a complete system in a precise context. This paper proposes an approach for the automatic generation of ground-truth information using a derivation of publishing tools. An implementation of this approach illustrates the richness of the produced information.

1. Introduction

Most of operational document image analysis and understanding (DIAU) systems are used in a specific context. This context is used as an *a priori* knowledge and allows to adapt processing strategies and to tune its parameters. The resulting system is then inefficient in other contexts.

The design of a DIAU system should at least answer the following questions :

- Which tasks should compose the system ?
- Which architecture should be used to combine theses tasks ?
- Which algorithm is the most adapted to accomplish each task ?
- Which are the most adapted parameters of each parameters ?

An optimization strategy can be used to answer these questions. It requires a evaluation procedures. Evaluation procedures can involve human expertise, but automatic evaluation which compares processing results to ground-truth thanks to well defined metrics is often preferred since it allows to manage a large amount of data, and is then more representative. Finally, the design of a DIAU system turns into an optimization process requiring ground-truth information and metrics for performance evaluation.

This paper proposes an approach for the automatic generation of ground-truth information used in the design of a DIAU system. Section 2 presents the motivations for this work and discusses the options which have been chosen. Section 3 gives a general description of our approach. This is illustrated via an example of implementation presented in section 4. Finally, section 5 discusses possible extensions.

2. Motivations

Performance evaluation is a recurring problem in DIAU, which motivates several contests (ie. ICDAR Page segmentation or Arabic word recognition contests, GREC Graphic recognition contest). The scientific community is sensible to this problem since there is a need to compare different approaches. However the number of different databases is itself the sign that there is no consensus concerning either their coverage, representativity or the associated ground-truth information.

The most famous databases are proposed by University of Washington [10], MediaTeam [11], the US National Library of Medicine[9], and more recently, by PRImA.

In their article [2], the authors claim for a document image database having two principal aims to reach an objective and exhaustive performance evaluation. First, the database should include document categories such as every day life documents. Additionally, document category distributions should reflect a realistic usage, while keeping a representative number of various documents inside the same category. These objectives seem difficult to reach. How to limit the extend of such a base? On the other hand, the definition of ground-truth information by hand for such an amount of data seems too expensive. As an example, let us notice that

the delivering of the 3 CD-ROM database by University of Washington has costed 2 million dollars. Some omissions remain and the database contains only English and Japanese documents.

Our objective is not to provide a new database allowing an "exhaustive" performance evaluation, but rather, in a pragmatic way, our work is related to the design of a DIAU system for a specific use case. None of the document categories managed by the system under design might be included in any of the existing databases. On the other hand, ground-truth information format of these databases may be incomplete or not adapted for the evaluation of some tasks.

The solution consists in building a representative document image database for the specific use case in which the associated ground-truth information is adapted for the evaluation task.

A first approach would be to manually define the ground-truth for a representative amount of document images extracted from the use case. As we would like to prevent from the subjectivity of a human operator and possible omissions or mistakes, we propose an approach based on the automatic generation of synthesized document images and associated ground-truth.

3. General description

The proposed approach derives a document publishing software (Fig. 1. Theses softwares use an internal representation of documents and produce the associated formated form which might be rendered on the screen on the fly (ie. WYSIWYG word processor) or which can be produced later thanks to a compiling step (ie. LaTeX documents). This step consists in applying a set of formatting rules (fonts, size, alignment...) under disposal constraints (paper size, margin, column...). This processing produces and exploits internal informations. For example, the formatting of a paragraph generates its bounding box which is then used to locate the next one. Our approach proposes to generate a ground-truth information by extracting from these informations the one which can be useful in a DIAU evaluation task.

4. Example implementation

This section presents an example implementation to illustrate our approach. This implementation is based on the XML DocBook publishing framework.

4.1. The XML DocBook publishing framework

An XML publishing framework is generally structured as follow. XML documents are logically marked up. Their

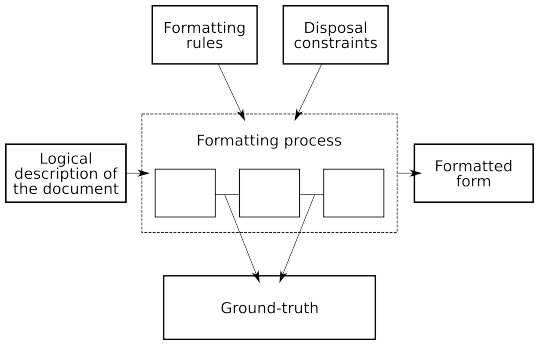


Figure 1. Formatting processing derivation

logical structure might then be directly available. These documents are then processed by a set of rules encoded in XSL stylesheets. A common usage is the transformation of XML documents into their (X)HTML equivalent, but our goal is to provide a formatted form of the document which might by printed on paper. With this intention, we use stylesheets which transform XML documents into XSL-FO. XSL-FO [1] is a W3C recommendation which aims at describing the formated form of documents. Then, XSL-FO documents can be converted into different usual formats (ie. RTF, SVG, PostScript, PDF) with the same rendering using free tools such as FOP, PassiveTex or commercial ones (AntennaHouse XSL Formatter, RenderX XEP...). This scheme is applicable to any XML Document. Only the stylesheet to transform XML into XSL-FO has to be defined.

DocBook [13] is a widely used DTD approved by OASIS (Organization for the Advancements of Structured Information Standards). As an example, O'Reilly uses DocBook as a standard for its publications. More and more softwares (ie. OpenOffice) integrate filters to handle XML DocBook Documents. DocBook is also the standard used for the Linux Documentation Project. These documents, published under the Free Documentation Licence, may then be used to build a large document database at a low cost. A DocBook document may represent a section, an article, a part, a book or a set of book. The DocBook DTD allows to logically mark up the textual content thanks to more than 300 tags. Among these tags, some express the logical hierarchy of the document: paragraph, section and subsection hierarchy, article, chapter, part, book, set. Others allow to declare elements such as tables, figures, bibliographies, internal or external links. Finally, XML DocBook documents may include fragments referring to other DTD: SVG fragments may be used to describe vectorial graphics and MathML fragments to mathematical formulas.

Many tools now interact with the XML DocBook format. Among them, we intensively use the DocBook XSL stylesheet [12] developed by Bob Stayton. These

stylesheets have been designed to be easily adapted via customization layers in which the users' rules overload the default ones. Then, several XML DocBook documents can be transformed using the same formatting rules. On the other hand, several customization layers can be applied to the same document to obtain as many formatted forms as shown on Fig. 2.

4.2. Ground-truth generation

In our XML DocBook publishing derivation, rather than producing an RTF, PostScript or PDF Format, a digital image is generated for each page of the document. The current implementation generates PNG, TIFF and JPEG output with several resolution and compression rates.

In parallel, our system outputs the internal information produced by the formatting process to generate a ground-truth information. A wide range of information are available:

Logical structure: It is directly available since the XML source document is logically marked up.

Formatting rules: The stylesheet used to transform the source document include the formatting rules applied to the logical elements. For each type of logical element, we can access to fonts informations (font family, size, style, color) and location attributes (left, right, center, justify).

Layout structure: It is a hierarchy of layout objects (page, column, block, line, word area, glyph area). Each of these elements is associated a set of information produced during the formatting step:

- its position and dimension
- type of content: text, graphic, array, formula, image...
- a logical label: title, author, paragraph, caption, bibliographic entry...
- Font attributes for textual elements : font family, size, style (bold, italic, underlined, strike), color...
- Location attributes: alignment, indentation, space before and after...
- Textual content
- A reference to the associated element in the logical structure

According to the evaluation task, the ground-truth format can include the whole or a part of this information. It can be encoded in several ways. For example, RDIFF or DAFS outputs could have illustrated that our system could be coupled with PSET [8] or PinkPanther [14], softwares designed for the page segmentation evaluation. But, we think that a TrueViz output better shows the richness of the information that can be extracted from the XML DocBook publishing framework. TrueViz [6, 7] is a DTD that has been initiated to encode ground-truth for the optical character recognition evaluation. TrueViz is extensible and has been adapted for the evaluation of other tasks of document image analysis and understanding. Moreover, a TrueViz comes with a Graphical User Interface allowing to visualize and edit ground-truth information. This GUI is an open source project which can then be extended or adapted to more specific needs as it has been done in the page segmentation evaluation project lead by the US National Library of Medicine.

Fig. 3 represents the TrueViz GUI. The regions of interest which constitute the ground-truth for page segmentation at deferent levels: blocks, lines, word areas, are shown on the left part together with the document image. The right part of the GUI shows the ground-truth as an expendable tree. These figures shows several levels of the automatically generated ground-truth. On the left part, we can notice that the reading order and the zones coordinates are kept in the generated ground-truth. The tree representation is more detailed and complete. It integrates the notion of zone inclusion (words in line, and lines in zones). One can also notice on Fig. 3(a) that the highlighted zone corresponds to a table of content (toc) element which contains only one line where the second word is "introduction". We could also have shown that it is possible to store font attributes.

5. Discussion

Our proposal states a procedure making it possible to automatically produced document images and ground-truth information for process comparison or their tuning in a specific context. This proposal is illustrated by the implementation presented in the previous section. It shows the variety of information that can be generated. This can at least be used for performance evaluation for the following tasks :

- Optical character recognition
- Optical font recognition
- Page segmentation
- Text/graphic/array/image/formula discrimination
- Layout structure analysis
- Logical labelling
- Logical structure understanding
- Document image classification

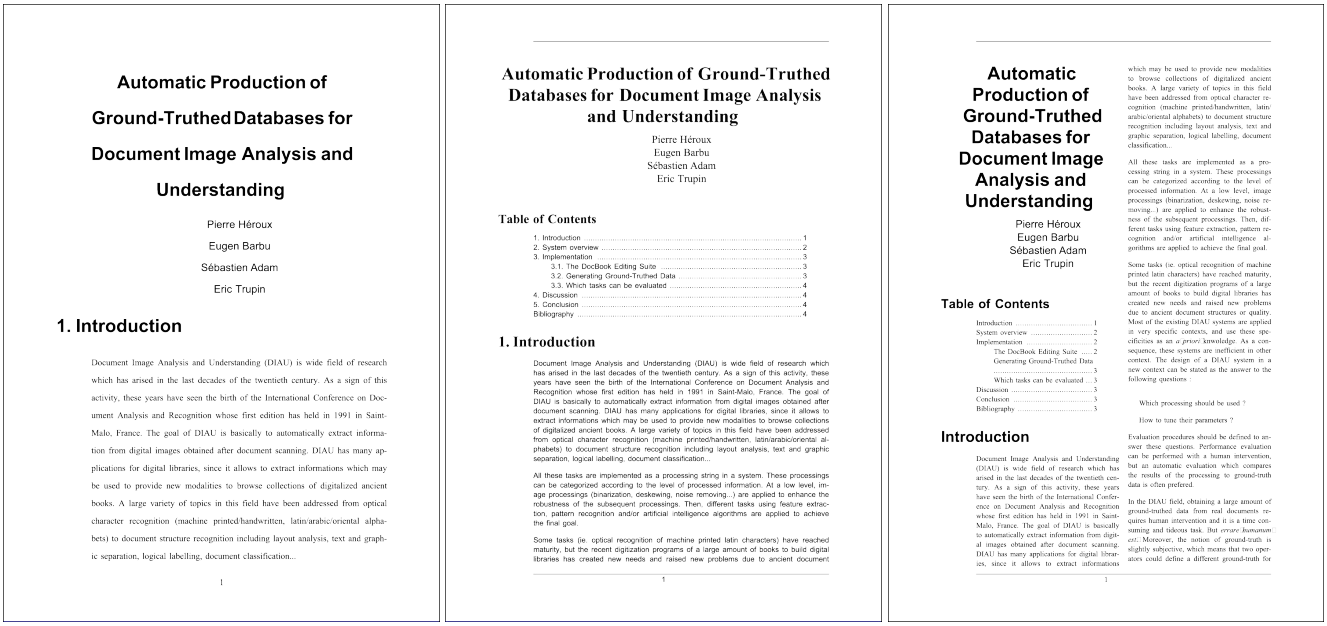


Figure 2. Three formatted forms of the same document

However, the ground-truth format can be restricted, extended or specialized to fit more specific needs such as locating particular zones corresponding to a logical element (figure, title, author...) or its font properties. The choice of generating synthesized image document leads to some limitations. First, the generated documents may only contain machine printed characters. This then excludes the evaluation of processings applied to handwritten text. In addition, current XSL-FO recommendations are limited to the description of isothetic rectangular text blocks, even if there are propositions [4] to extend its scope. Finally, one can argue the representativity of synthesized document images compared to real ones, regarding, first, the document content and its formatting, and, second, the image quality.

Many works have tried model document image deformation and degradations. The more significants have been lead by Baird [3] and Kanungo [5]. Other approaches [15] propose to mix the "perfect" image with the result of a white page digitization.

To answer to the question of representativity of synthesized document images, we propose to validate the stylesheet with the following procedure.

1. Choose some real documents
2. Write their XML Transcription
3. Process the transcription with the stylesheet
4. Compare the resulting document image with the initial real document
5. Modify the stylesheet and go back to step 3

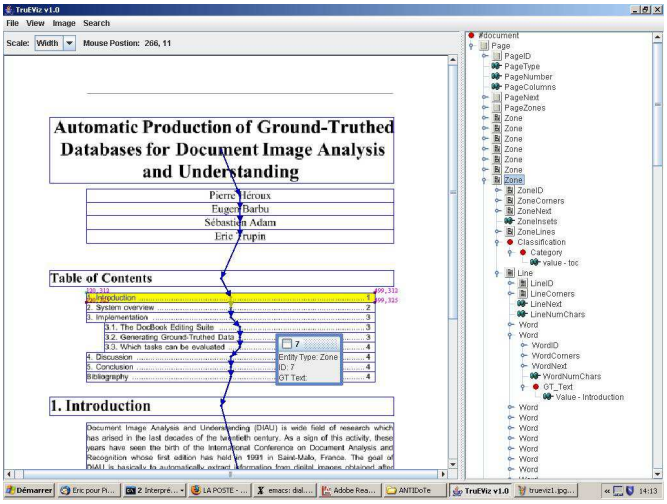
6. Conclusion

This article proposes an approach for the automatic generation of document images and the associated ground-truth for the performance evaluation in DIAU. This approach is an alternative to databases containing documents which are not representative for the specific use case or if the ground-truth is not adapted to the evaluation.

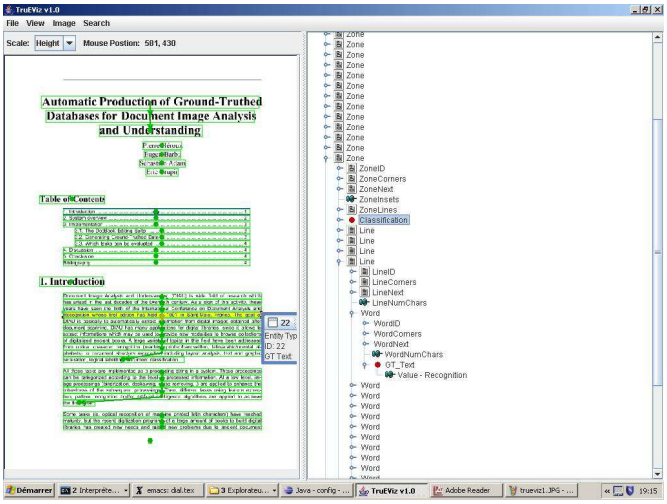
An implementation of this approach processing XML DocBook documents illustrates the richness of information that can be produced and the flexibility allowed for its encoding (DAFS, RDIFF, TrueViz or other for more specific usage). It allows performance evaluation of a wide range of processings. Indeed, if many existing databases give ground-truth information for OCR or page segmentation, our approach offers superior possibilities such as font recognition, table of content location, and logical structure understanding.

The automatic ground-truth generation prevents from omissions, mistakes and subjectivity.

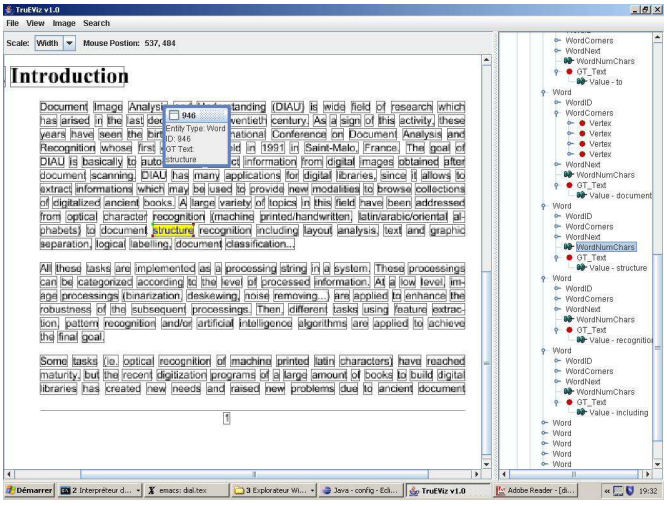
The use of XML technology make it easy to extend the system to generate other document types. Indeed, with only a little effort, we have succeeded in generating document images and associated ground-truth from XML TEI documents, thus producing documents which are closer to the ones currently processed in the digitization projects lead by the libraries.



(a) Zone view



(b) Line view



(c) Word view

Figure 3. Automatically generated ground-truth viewed with TrueViz GUI

In the future, we intend to complete our tool with implementation of synthetic image degradation algorithms and methods for the comparison of automatic processing results to ground-truth information. Then, we intend to give an public access to this tool for the scientific community and wish that its suggestions could allow to improve the tool and finally to converge to shared formats for ground-truth.

References

[1] S. Adler, A. Berglund, J. Caruso, S. Deach, T. Graham, P. Grosso, E. Gutentag, A. Milowski, S. Parnell, J. Richma, and S. Zilles. *Extensible Stylesheet Language (XSL)*. W3C, `http://www.w3.org/TR/xsl/`, 2001.

[2] A. Antonacopoulos, D. Karatzas, and D. Bridson. Ground-truth for layout analysis performance evaluation. In H. Bunke and A. L. Spitz, editors, *Document Analysis Systems VII, seventh International Workshop DAS 2006*, pages 303–311, 2006.

[3] H. S. Baird. State of the art of document image degradation modeling. In *Proceedings of the IAPR International Workshop on Document Analysis Systems*, pages 1–16, 2000.

[4] A. C. B. da Silva, J. B. S. de Oliveira, F. T. M. Mano, T. B. Silva, L. L. Meirelles, F. R. Meneguzzi, and F. Gianetti. Support for arbitrary regions in XSL-FO. In *Proceedings of the 2005 ACM Symposium on Document Engineering*, pages 64–73, 2005.

[5] T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuezle, and D. Madigan. Statistical, nonparametric methodology for document degradation model validation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1209–1223, 2000.

[6] T. Kanungo, C. H. Lee, J. Czorapinski, and I. Bella. TRUEVIZ: a groundtruth/metadata editing and vizualization toolkit for OCR. In *Proceedings of the SPIE Conference on Document Recognition and Retrieval*, pages 1–12, 2001.

[7] S. Mao and T. Kanungo. Empirical performance evaluation and its application to page segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Interlligence*, 23(3):242–256, 2001.

[8] S. Mao and T. Kanungo. Software architecture of PSET: a page segmentation evaluation toolkit. *International Journal on Document Analysis and Recognition*, 4(3):205–217, 2002.

[9] U. S. National Library of Medecine. Medical archive records groundtruth. `http://marg.nlm.nih.gov/index.swf`.

[10] I. T. Philips and R. M. Haralick. CD-ROM document database standard. In *Proceedings of the second International Conference on Document Analysis and Recognition*, pages 478–483, 1993.

[11] J. Sauvola and H. Kauniskangas. *MediaTeam Oulu Document Database*. MediaTeam, Oulu University, Finland, `http://www.mediateam.oulu.fi/MTDB/`, 1998.

[12] B. Stayton. *DocBook XSL: The Complete Guide*. Sagehill Enterprises, 2005.

[13] N. Walsh and L. Muellner. *DocBook: The Definitive Guide*. O'Reilly, 1999.

[14] B. A. Yanikoglu and L. Vincent. PinkPanther: A complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition*, 31:1191–1204, 1998.

[15] G. Zi and D. Doermann. Document image ground truth generation from electronic text. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 663–666, 2004.