A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents

Max Göbel,
Tamir Hassan
PRIP, Technische Universität Wien
goebel@prip.tuwien.ac.at
tam@prip.tuwien.ac.at

Ermelinda Oro
ICAR - CNR
Università della Calabria
oro@icar.cnr.it

Giorgio Orsi
Dept. of Computer Science
University of Oxford
giorgio.orsi@cs.ox.ac.uk

ABSTRACT

This paper presents a methodology for the evaluation of table understanding algorithms for PDF documents. The evaluation takes into account three major tasks: table detection, table structure recognition and functional analysis. We provide a general and flexible output model for each task along with corresponding evaluation metrics and methods. We also present a methodology for collecting and ground-truthing PDF documents based on consensus-reaching principles and provide a publicly available ground-truthed dataset.

Categories and Subject Descriptors: I.7.5 [**Document and Text Processing**]: Document Capture—*document analysis*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation*

Keywords: Table processing, metrics, ground-truth dataset, performance evaluation, document analysis, document understanding

1. INTRODUCTION

The problem of *table understanding* has attracted much interest in previous years from the database as well as the document engineering communities. On the Web, discovering structured data is a tremendous challenge [1] and PDF documents represent the most common document format after HTML. It is commonly recognized that table understanding consists of three tasks of increasing complexity:

- *table detection:* locating the regions of a document with tabular content;
- *table structure recognition:* reconstructing the cellular structure of a table;
- *table interpretation:* rediscovering the meaning of the tabular structure. This includes:
  (a) *functional analysis:* determining the function of cells and their abstract logical relationships;
  (b) *semantic interpretation:* understanding the semantics of the table in terms of the entities represented in the table, their attributes, and the mutual relationships between such entities.

DocEng'12, September 4–7, 2012, Paris, France.

The comparative evaluation of different table understanding algorithms is a non-trivial matter. Currently available datasets for table understanding algorithms suffer from the following limitations: *(i)* the documents in these datasets are scanned images, not natively digital PDF (*PDF Normal* or *Formatted Text and Graphics*) documents; *(ii)* in such cases where ground truth is provided, only the tabular regions are present; *(iii)* custom performance measures make it very difficult to appreciate fine differences between the algorithms being compared.

In order to overcome the limitations of existing datasets and evaluation approaches, we provide:

- a model and corresponding evaluation metric for each output of the three stages of table understanding. At the interpretation stage, we address only functional analysis because the semantic interpretation of a table is domain specific, and we believe it is premature to include it in a generic benchmarking dataset;
- a consensus-based methodology for collecting and ground-truthing natively digital PDF documents;
- an initial open-access and extensible ground-truthed dataset of PDF documents containing tables.

2. RELATED WORK

Common problems in the comparative evaluation of different table understanding algorithms are the lack of standardized datasets, benchmarking procedures and measures in experimental evaluation. This section discusses previous work in the: *(i)* creation of ground-truthed datasets, *(ii)* modelling of tabular information and *(iii)* definition of evaluation metrics.

Ground-truthed datasets. The lack of availability of ground truth datasets has proved to be a major hindrance to the comparative evaluation of table recognition algorithms.

Publicly available datasets containing tabular content in monochrome page images, such as the UW datasets [11], and UNLV[1], have long been available in the OCR community. The ground truth has been generated interactively by using visual tools [4, 12]. However, only the tabular regions, and no higher-level information, is provided.

The first publicly-available dataset containing natively digital PDF documents was used to test the PDF-TREX system [10]. However, it contains mostly Italian financial tables and does not include ground-truth information. As our goal was to create a multi-domain database in the En-

[1]UNLV dataset originally at `http://www.isri.unlv.edu/ISRI/OCRtk`. No longer available; accessed at `web.archive.org`

glish language, we decided to begin the document collection process from scratch.

Table models. There are a number of different levels at which table understanding can operate, a fact that is reflected in a variety of table models. In particular, we can distinguish between *structural models*, used for representing region and cell structures of tables, and *conceptual models*, enabling the abstraction of content from presentation.

Interesting structural models have been proposed in [5, 7, 12]. In particular Hu et al. [5] modelled a table as a directed acyclic attributed graph (table DAG) where columns, rows, cells and relations among them are represented. Hurst [7] presents an approach to deriving an abstract geometric model of a table from a physical representation based on spatial relations among cells named *proto-links*, which exist between immediate neighbouring cells. Shahab et al. [12] use an image-based representation to describe the cell structure, adopting different colour channels to represent different row and column positions. As discussed in Section 3.2, for comparing two cell structures of a table we use a model inspired by Hurst's proto-links, which enables an effective and simple evaluation measure to be defined.

Possibly the most well-known and cited conceptual model has been proposed by Wang [13] and extended by Hurst [6]. Wang defines a table divided into four main regions: *(i)* the *stub* that contains the row headings; *(ii)* the *boxhead* that contains the column headings; *(iii)* the *stub head* that contains the index sets in the stub and *(iv)* the *body* that contains entries (also named data cells). At the lowest level, a table can be seen as being composed of two types of cell: the *data cell*, and the *access cell* (or label). The data cells comprise the core of the table, whereas the access cells occur within headers and are further classified into *categories* that are organized hierarchically. In Section 3.3 we use many of these concepts in defining our functional model.

Evaluation metrics. In order to evaluate the results of table understanding algorithms, several metrics for table structure recognition have been proposed. However, well defined evaluation metrics do not yet exist for the results of table interpretation.

Performance measures from the information retrieval domain such as recall, precision [9] and combined F-measure have also found their way into evaluating table recognition algorithms [8, 10]. Results of the PDF-TREX system [10] were given using separate precision and recall values for table areas and cell structures. In [2] the concepts of *completeness* and *purity*, based on the definitions of recall and precision, were introduced as well-defined evaluation metrics for any segmentation task. Whereas these measures can intuitively be adopted for the table (region) recognition phase, they are not so applicable for table (cell) structure recognition. In table structure recognition, a variety of errors can occur that need to be considered separately (e.g. cells can be split in one direction, merged in another; entire blank columns can appear) and classifying these errors can lead to ambiguities [3]. An alternative approach, which uses several precision and recall measures at several levels, including cell, row, column and region, is proposed in [12].

Hurst [7] sidestepped these problems by evaluating precision and recall at the proto-link level.

3. MODELLING THE GROUND TRUTH

The ground-truth enables a fair comparison between different approaches to the table understanding problem. In order to be considered in our dataset, a table must have a meaningful representation in each of the output models of the three understanding tasks: (1) the *region model* for table detection, (2) the *cell structure model* for table structure recognition, (3) and the *functional model* for functional analysis. More precisely, a table in our dataset has the following characteristics:

(i) it consists of (rectangular) cells belonging to an unambiguous two-dimensional row-and-column structure. Cells may span more than one row or column;

(ii) the contents of the table must fit within a rectangular bounding box that must not contain any further textual content (titles, captions and footnotes are not considered to be part of the table);

(iii) it has a clear *functional model* based on clearly defined *access cells* and *data cells*. Each data cell must be accessed by at least two access dimensions.

3.1 Table regions

Region model. Table regions are defined as rectangular areas of a given page by their coordinates. Since a table can span more than one page, several regions can belong to the same table. For each region, we store the textual *operator* (and, if necessary, *operand*) *ID*s of their originating PDF text instructions (i.e. Tj and TJ), which point back to the particular point in the PDF file where the text was drawn. Each region in the ground truth is set to the minimal bounding box that bounds all textual objects within.

Comparing regions. In order to compare a table region against the ground truth, we can use two methods:

(i) if comparing algorithms that can be adapted to return the internal PDF operators, we can compare each character with reference to the particular operator responsible for drawing the text on the page;

(ii) for other (e.g. "black-box") algorithms, bounding boxes and content are used. A region is correct if it contains the minimal bounding box of the ground truth without intersecting additional content.

For comparing tabular regions, we use the measures *completeness* and *purity* [2] as they are well defined in the context of segmentation. In order to obtain the best mapping between two sets of regions, which may also differ from each other, a correspondence matrix [12] is used.

3.2 Cell structure

Cell structure model. The cell structure of a table is defined as a matrix of cells. The ground truth provides its textual content and its start and end column and row positions. Blank cells are not represented in the grid. A benefit of such a representation is that each cell is independent from what has previously occurred in the table definition.

Comparing cell structures. For comparing two cell structures, we use a method inspired by Hurst's proto-links [6]: for each table region we generate a list of *adjacency relations* between each content cell and its nearest neighbour in horizontal and vertical directions. No adjacency relations are generated between blank cells or a blank cell and a content cell. This 1-D list of adjacency relations can be compared to the ground truth by using precision and recall measures, as shown in Figure 1. If both cells are identical and the

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(a) Original table as in ground truth

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(b) Incorrectly recognized cell structure with split column

■ Correct adjacency relations □ Incorrect adjacency relations

$$\text{Recall} = \frac{\text{correct adjacency relations}}{\text{total adjacency relations}} = \frac{24}{31} = 77.4\%$$

$$\text{Precision} = \frac{\text{correct adjacency relations}}{\text{detected adjacency relations}} = \frac{24}{28} = 85.7\%$$

Figure 1: Comparison of an incorrectly detected cell structure with the ground truth

direction matches, then it is marked as correctly retrieved; otherwise it is marked as incorrect. Using neighbourhoods makes the comparison invariant to the absolute position of the table (e.g. if everything is shifted by one cell) and also avoids ambiguities arising with dealing with different types of errors (merged/split cells, inserted empty column, etc.).

3.3 Table interpretation

Functional model. Our functional model focuses on expressing the most important relations of a table, which reflect the way a naïve human reader would use the table to look up information. As in [13, 6], our functional model consists of a set of access relations defined as follows: Let $I = \{I_1, \ldots, I_n\}$ be a collection of *access dimensions* and E the set of *data cells*. An *access function* $f : \bigotimes I \to E$ maps the unordered cartesian product of access dimension sets to the set of entry values. Given a set of access cells as input, an access function returns a data cell.

A table's functional representation cannot usually be fully rediscovered from the layout alone. For example, in Figure 2 domain-specific knowledge is required to discover that the cell Nationality of parent: is a heading for the cells below it, and not the cells to its right. Dot notation is used to represent access cells arranged hierarchically. Although the *physical* structure of a table is 2-D, often more dimensions are projected into this 2-D space. For instance, in Figure 2 there are three dimensions that allows for describing a data cell: years, nations and the set given by the cells Activity, Passivity and Net position (which are repeated for each year).

It is not always clear which cells serve as access cells and which cells are the data cells in a table. For instance, in Figure 3 both the airline name and airline code could be used to look up the airline's turnover; thus both columns serve simultaneously as access cells to the figures. A further example is that of a conversion table between e.g. metric and imperial units, which could be read in either direction.

It is worth nothing that, in contrast to the cell structure model which is purely physical, in the functional model it is important to represent blank *data* cells. For instance, the table in Figure 3 includes a blank data cell that represents a null value.

Comparing functional representations. As with the cell structure model, we compute precision and recall measures for all the access relations within the functional representation of a table. An access relation is marked as correctly detected if it is identical to the ground truth, i.e. all levels of each access path are present. However, in cases where the heading structure has only been partly recovered but the lowest level access cells have all been correctly detected, the relation is marked as *partially detected*. If we consider an algorithm that analyses a table with multiple-level headings correctly, but misses some of the higher-level headings, the result is still likely to be useful. Thus, our evaluation measure better reflects the usefulness of the result.

Precision and recall can be calculated from the number of *correctly detected* access relations, the *total number of correct* access relations and the number of *incorrectly detected* (or false positive) access relations.

For an access relation that has had all of its lowest-level access cells and its data cell correctly recognized, the number of correctly detected access relations is incremented by the following fraction:

$$\frac{\text{number of correctly detected entities}}{\text{total number of entities}}$$

where *entity* refers to access or data cell. Access cells

INTERNATIONAL ASSETS AND LIABILITIES OF BIS REPORTING BANKS BY NATIONALITY OF PARENT
(outstanding amounts in billions of dollars)

Nationality of parent:	1997 Activity	1997 Passivity	1997 Net position	1998 Activity	1998 Passivity	1998 Net position
USA	**961.9**	**1 008.4**	**–46.5**	**1 105.3**	**1 173.0**	**–67.7**
Canada	**211.5**	**219.5**	**–8.0**	**239.1**	**237.3**	**1.8**
Japan	**2 045.1**	**1 598.0**	**447.1**	**1 758.2**	**1 312.1**	**446.1**
Europe	**5 025.7**	**5 218.3**	**–192.6**	**5 789.3**	**6 064.1**	**–184.8**
of which: Germany	1 346.9	1 345.1	1.8	1 630.3	1 638.2	–7.9
France	903.6	968.7	–65.1	1 021.7	1 060.1	–38.4
United Kingdom	478.8	539.1	–60.3	558.8	632.3	–73.5
Italy	419.0	416.8	2.2	443.1	434.0	9.1
Switzerland	709.4	706.0	3.4	836.5	836.5	9.6
Other regions	**539.0**	**522.7**	**16.3**	**626.7**	**515.1**	**111.6**
Total	**8 783.2**	**8 566.9**	**216.3**	**9 518.6**	**9 301.6**	**217.0**

Source: BIS

Source: Adapted from the PDF-TREX dataset [10]

Functional representation:

[Nationality of parent.USA],[1997],[Activity] → [961.9],
[Nationality of parent.USA],[1997],[Passivity] → [1 008.4],
[Nationality of parent.USA],[1997],[Net position] → [−46.5],
[Nationality of parent.USA],[1998],[Activity] → [1 105.3],
. . . ,

Figure 2: A financial table and its functional model

		Turnover ($bn)		
		2008	**2009**	**2010**
AA	American Airlines	17.5	18.1	17.2
AF	Air France	11.6	10.8	11.9
KL	KLM Royal Dutch Airlines	8.3	9.5	9.4
LH	Lufthansa	12.8	14.1	13.8
NA	New Airline		2.1	2.4

Functional representation:

[AA],[Turnover ($bn).2008] → [17.5],
[American Airlines],[Turnover ($bn).2008] → [17.5],
[AA],[Turnover ($bn).2009] → [18.1],
[American Airlines],[Turnover ($bn).2009] → [18.1],
. . . ,
[NA],[Turnover ($bn).2008] → [],
. . . ,

Figure 3: A table with two alternative access paths

are counted as having been correctly detected if the access path from the *lowest level upwards* is identical to the ground truth; otherwise they are considered as false positives, even if they are pointing to the correct cell.

Likewise, for an access relation that has had all of its lowest-level access cells and its data cell correctly recognized, the number of incorrectly detected (false positive) access relations is incremented by the following fraction:

$$\frac{\text{number of incorrectly detected entities}}{\text{total number of entities}}$$

Here, any incorrectly detected access cells above the lowest level are counted as incorrectly detected entities. If the data cell or any access cell at the lowest level is incorrect, the number of incorrectly detected access relations is incremented by 1.

4. THE DATASET

In order to build an objective dataset of freely distributable PDF documents from several domains, we performed a Google search and inspected each returned document in sequence. We used the following search terms in order to obtain documents from government sites whose publications are known to be in the public domain: (a) `filetype:pdf site:europa.eu` (b) `filetype:pdf site:*.gov`

The size of documents and the number of tables contained within the documents varied greatly. For longer documents (more than 5 pages), *excerpts* of pages containing tables were extracted, with approximately 2 pages of non-tabular content before and after the tables of interest. Thus, we also include non-tabular pages, giving the opportunity to also test each algorithm against its resistance to false positives.

Core dataset. Our core dataset, which is freely downloadable at `http://www.tamirhassan.com/dataset/`, contains 59 excerpts as individual PDF files, with a total of 117 tables, with ground truth information corresponding to all three tasks defined in Section 3. Each of these files has a domain-generic model, specified as an XML Schema Definition (XSD), enabling the output of existing systems to easily be converted for comparison with the ground truth. Table 1 gives overall statistics on the tables we have gathered.

The ground truth has been created interactively using a visual tool for annotating table regions and cells. Since the nature and the content of tables is often a subjective matter, the construction of the ground truth has followed a strict consensus-reaching methodology. Document excerpts have been collected and ground truth has been generated independently and then validated by a group of three experts. If it was not possible to reach consensus on any aspect of the ground truth or the representation in any of the models was considered ambiguous by at least one expert, the excerpt was excluded from our dataset. Because of the difficulties and increased ambiguity in the functional analysis of "one-dimensional" tables such as conversion tables, and tables with two "primary keys" (Figure 3), we decided not include to such typologies of tables in our dataset.

Data source	EU	US Gov.
Number of documents containing:	12	15
single-column layout	12	11
multi-column or complex layout	0	4
Number of excerpts	34	28
Number of pages	101	74
Number of tables of which:	74	38
are split across more than one page	0	5
contain indentations	9	1
are partly ruled	45	17
are fully ruled	19	34
are laid out using monospaced text	0	0

Table 1: Summary of the tables in the core dataset

5. CONCLUSION

In this paper we have presented an evaluation methodology for table understanding algorithms. Although we have focused on PDF documents, we believe that the same models can be easily adapted and applied to other formats such as scanned images and web documents (e.g. HTML). We invite researchers and practitioners from the web data management and document engineering communities to join our initiative and collaborate on the enrichment of the initial dataset that we provided. In addition, our models have been defined to be extensible and we expect them to be adapted to embrace more cases than those defined in this paper.

Acknowledgements: This work was funded in part by the DIADEM Project (EC FP7 Programme Grant No. 246858), the Oxford Martin School (Grant No. LC0910-019) and the Austrian Federal Ministry of Transport, Innovation and Technology (Grant No. 829602).

6. REFERENCES

[1] M. J. Cafarella, A. Halevy, and J. Madhavan. Structured data on the web. *Commun. ACM*, 54(2):72–79, 2011.

[2] A. C. e Silva. Metrics for evaluating performance in document analysis: application to tables. *IJDAR*, 14(1):101–109, 2011.

[3] T. Hassan. Towards a common evaluation strategy for table structure recognition algorithms. In *Proc. of DocEng*, 2010.

[4] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Evaluating the performance of table processing algorithms. *IJDAR*, 4(3):140–153, 2002.

[5] J. Hu, R. Kashi, D. Lopresti, G. Wilfong, and G. Nagy. Why table ground-truthing is hard. In *Proc. of ICDAR*, pages 129–133, 2001.

[6] M. Hurst. *The Interpretation of Tables in Texts*. PhD thesis, University of Edinburgh, 2000.

[7] M. Hurst. A constraint-based approach to table structure derivation. In *Proc. of ICDAR*, pages 911–915, 2003.

[8] T. Kieninger and A. Dengel. An approach towards benchmarking of table structure recognition results. In *Proc. of ICDAR*, pages 1232–1236, 2005.

[9] D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proc. of SIGIR*, pages 246–254, 1995.

[10] E. Oro and M. Ruffolo. PDF-TREX: An approach for recognizing and extracting tables from PDF documents. In *Proc. of ICDAR*, pages 906–910, 2009.

[11] I. T. Phillips. User's reference manual for the uw english/technical document image database III. Technical report, Seattle University, 1996.

[12] A. Shahab, F. Shafait, T. Kieninger, and A. Dengel. An open approach towards the benchmarking of table structure recognition systems. In *Proc. of DAS*, pages 113–120, 2010.

[13] X. Wang. *Tabular Abstraction, Editing and Formatting*. PhD thesis, University of Waterloo, 1996.