Configurable Table Structure Recognition in Untagged PDF Documents[1]

Alexey Shigarov[1]
shigarov@icc.ru

Andrey Mikhailov[1]
mikhailov@icc.ru

Andrey Altaev[1]
altaev@icc.ru

[1]Matrosov Institute for System Dynamics and Control Theory,
Siberian Branch of the Russian Academy of Sciences

16th ACM Symposium on Document Engineering
September 15, 2016, Vienna, Austria

---

[1]This work was financially supported by the Russian Foundation for Basic Research (grant 15-37-20042)

Introduction

- Nganji[2] estimates that 95.5% of scientific articles published by four leading publishers are untagged PDF documents
- *"Untagged"* means no tables and cells, only printing instructions for text chunks and graphics
- So, *PDF Table Extraction* is the challenging task
- Today, some academic and commercial tools continue to appear and compete
- Motivation for our work consists in
  - defining a configurable part (parameters and ad-hoc heuristics) in the process of table structure recognition
  - examining features of appearance of text printing instruction in PDF files for recovering human reading order
  - reaching a high accuracy on the existing competition dataset

[2] J. Nganji, The Portable Document Format (PDF) accessibility practice of four journal publishers. Library & Information Science Research, 2015, 37(3), 254-262.

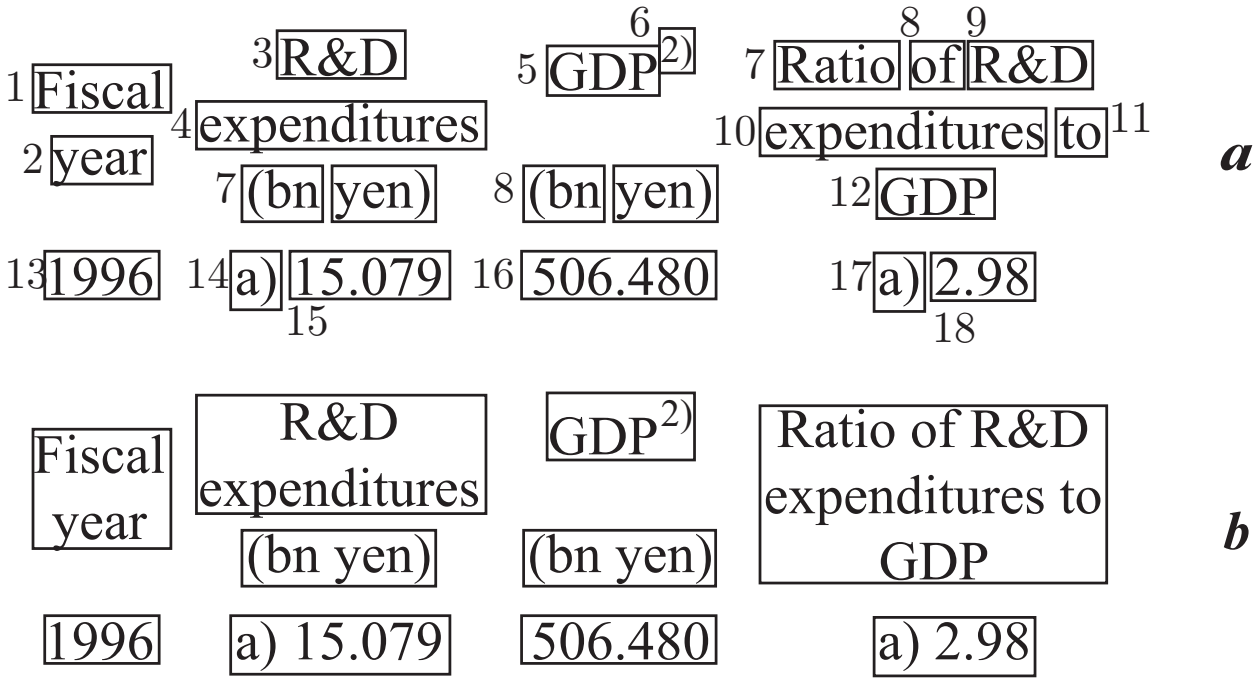
Table Structure Recognition

Preprocessing

- Splitting original text chunks (a) into one-character chunks (b)
- Merging one-character chunks into word chunks and reindexing the order of their appearance (c)
- Splitting each rectangle (d) into four rulings (e)
- Merging segments of one visual line (f) into one ruling (g)
- *Heuristics*: eliminating text chunks containing only itemization or padding characters

Table Structure Recognition

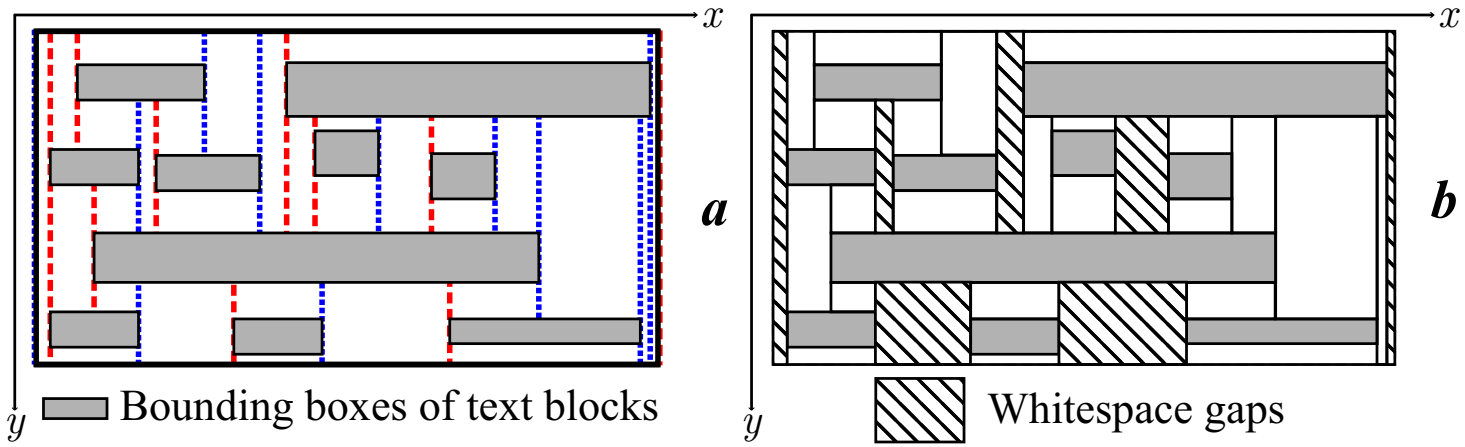
Text Block Recovering



- Merging word chunks (a) into text blocks (b)
- *In the best case*: each block is a textual content of a cell
- *Heuristics*: adjacency in the order of the appearance, no rullings, identical fonts, word and line spacing, vertical and horizontal projections

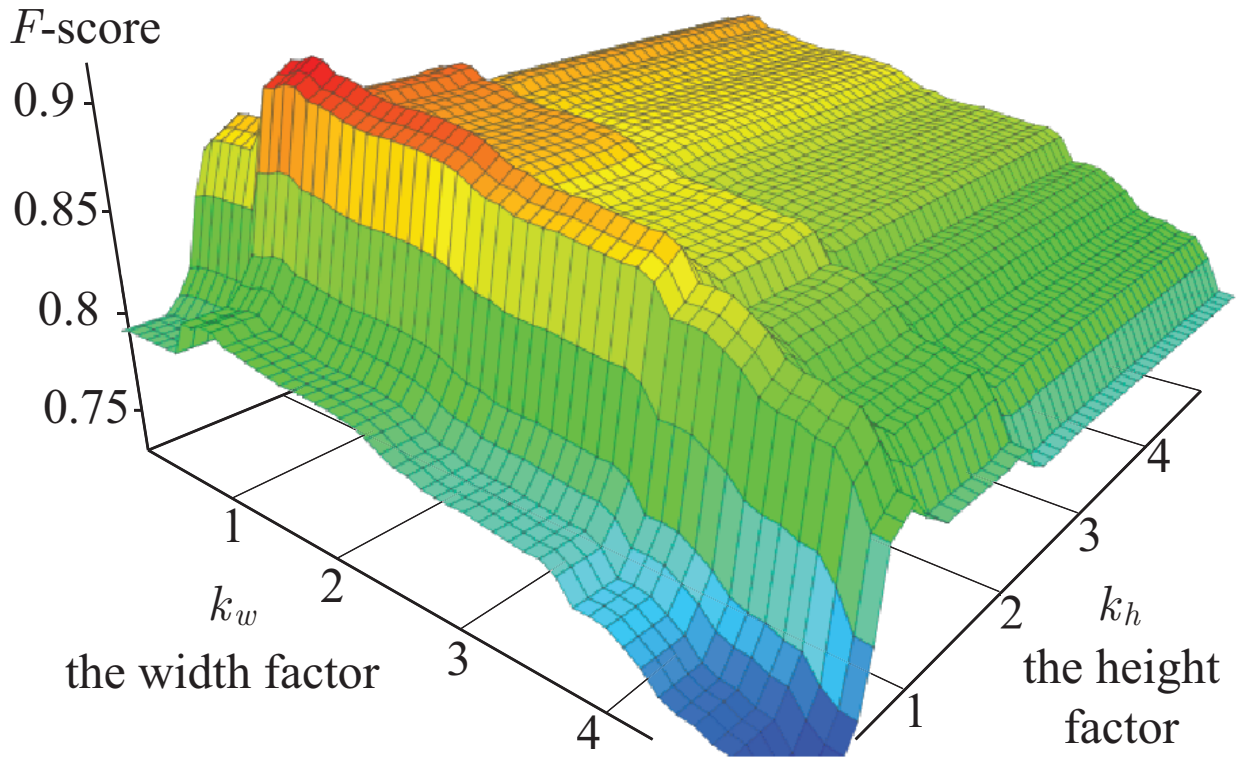
Table Structure Recognition

Cell Recovering



- There are two ways to arrange text blocks into cells
  - Analysis of witespace gaps between text block
  - Analysis of connected text blocks (bounding boxes)
- *Heuristics*: a column containing only one non-empty cell is merged with the nearest column to the left

Configuring



- Defining formulas to set up word and line spacing, as well vertical and horizontal projections
- Choosing predefined ad-hoc heuristics
- Searching “optimal” parametrs on a target dataset

Experimental Evaluation

- Two configuration were implemented. In the better case, we have:

Recall	0.9233
Precision	0.9499
F-score	0.9364

- The evaluation is based on
  - The methodology for algorithms for table understanding in PDF documents[3]
  - "ICDAR 2013 Table Competition" dataset[4]
  - Nurminen's Python scripts[5] for comparing ground-truth and results

[3] M. Göbel, T. Hassan, E. Oro, G. Orsi. A methodology for evaluating algorithms for table understanding in PDF documents. In Proc. of the DocEng'12. 2012, pp. 45-48.

[4] M. Göbel, T. Hassan, E. Oro, G. Orsi, ICDAR 2013 Table Competition. In Proc. of the 12th ICDAR, Washington, DC, 2013, pp. 1449-1453.

[5] `http://tamirhassan.com/competition/dataset-tools.html`

Web-Application for PDF Table Extraction

- Our experimental web-application is available at `http://cells.icc.ru/pdfte`
- Now, it enables only manual table selection, but automatic table structure recognition
- Extracted tables are accessible in HTML and Excel format
- Further they can be transformed into a relational form through `http://cells.icc.ru/ssdc`