This method provides a simple, repeatable way to fairly account for a wide variety of errors in table structure recognition (e.g. extra blank columns, split rows, undetected colspans, etc.) As no coordinate information is used, result files in HTML, text and other formats can also be easily evaluated using this method. In order to account for possible character encoding issues, each content string was normalized by removing whitespace, replacing all special characters with an underscore and converting all lowercase letters to uppercase.

C. *Alternative ground truths*

Although great care was taken in avoiding excerpts containing ambiguous tables when generating the dataset, some of these ambiguities only became apparent when analysing the participants' submissions. Therefore, "alternative" ground truth files were later generated for four of the excerpts in the dataset. Where there were discrepancies between the ground truths in generating the numerical results, the ground truth returning the better numerical result was always chosen.

D. *Combining results*

There are several ways to average the precision and recall scores over the complete dataset. For both region and structure results, we chose to first calculate these scores for each document separately and then calculate the average based on the document scores. This way, each document has equal weighting and the result is not skewed by the few documents containing tables with hundreds or thousands of cells.

Because of the relatively small number of tables in a single document, we chose not to do this for completeness and purity and simply totalled the number of complete and pure tables over the complete dataset.

IV. Participating methods

The following subsections describe the various systems that have participated in the competition. A summary of the main features is given in Table I.

Participant	Format	Internal model	Methodology	Sub-competitions
Fang et al.	PDF	Objects	Heuristics	LOC
Nurminen	PDF	Img. & obj.	Heuristics	LOC, STR, COM
Yildiz	PDF	Text lines	Heuristics	LOC, COM
Silva	TXT	Text lines	Heur. + ML	LOC, STR, COM
Stoffel	PDF	Text lines	Heur. + ML	LOC
Hsu et al.	Images	Objects	Heuristics	LOC, STR
Liu et al.	PDF	Objects	Heuristics	LOC

TABLE I
Summary of the main features of each participating method

A. *ICST-Table system, Fang et al.*

The ICST-Table system [5] was submitted by Jing Fang, Leipeng Hao, Liangcai Gao, Xin Tao and Zhi Tang from the Institute of Computer Science & Technology, Peking University, Beijing, China and is designed to recognize tables in born-digital PDFs, which are parsed using a commercial library. The heuristic approach locates tables by finding whitespace and line separators and filtering out regions containing paragraphs of text. It is worth noting that in [5] authors compared their evaluation results with those presented by Liu et al. in [6], obtaining better precision and recall. In this competition, we were able to compare the two systems directly, and this time Liu et al. obtained better results on our dataset.

B. *Tabler system, Nurminen*

Anssi Nurminen developed the *Tabler* system as part of his MSc degree at Tampere University of Technology, Finland. The system processes born-digital PDF documents using the Poppler library and combines raster image processing techniques with heuristics working on object-based text information obtained from Poppler in a series of processing steps.

C. *pdf2table system, Yildiz*

Burcu Yildiz developed the *pdf2table* system [7] at the Information Engineering Group, Technische Universität Wien, Austria. The system employs several heuristics to recognize tables in PDF files having a single column layout. For multi-column documents, the user can specify the number of columns in the document via a user interface; however, such user input was not allowed in the competition. The approach was able to handle most of the documents where the tables span the entire width of the page. However, the issue of false positives was not properly addressed, as in the original workflow these would have been discarded via user interaction.

D. *TABFIND algorithm, Silva*

Ana Costa e Silva, from the Laboratory of Artificial Intelligence and Decision Support (LIAAD-INESC), Porto, Portugal, used an algorithm that works on textual files line-by-line, and the PDF dataset was therefore converted into text format, resulting in loss of information. The method used in the competition differs somewhat from the one presented in her thesis [8] and was adapted specifically for the competition dataset by assuming, for example, that tables have at least one line where all cells are non-empty. Furthermore, the algorithm also incorporates a training procedure for parameter tuning.

E. *Stoffel's system*

Andreas Stoffel, from the Department of Computer and Information Science, University of Konstanz, Germany, participated with a trainable system [9], [10] for the analysis of PDF documents based on the PDFBox library. After initial column and reading-order detection, logical classification is performed on the line level. In order to detect tables, the system was trained on the practice dataset using a sequence of a decision-tree classifier and a conditional random field (CRF) classifier. Consecutive lines labelled as tabular content were then grouped together and output as a table region.

F. *KYTHE system, Hsu et al.*

The *Kansas Yielding Template Heuristic Extractor (KYTHE)* was submitted by William H. Hsu (group leader), Xinghuang Xu and Jake Ehrlich from the Department of Computing and Information Sciences, Kansas State University, in collaboration with Praveen Koduru of iQGateway LLC.