READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D6.8
Table and Form Analysis Tool P2

Florian Kleber, Markus Diem and Stefan Fiel
CVL

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	31.12.2017
Actual date of delivery	28.12.2017
Date of last update	16.11.2017
Deliverable number	D6.8
Deliverable title	Table and Form Analysis Tool P2
Type	Report, Demonstrator
Status & version	in progress
Contributing WP(s)	WP4, WP6
Responsible beneficiary	CVL
Other contributors	CVL, ABP, NaverLabs
Internal reviewers	Naverlabs, NCSR
Author(s)	Florian Kleber, Markus Diem and Stefan Fiel
EC project officer	Martin MAJEK
Keywords	table analysis, table matching, forms analysis

Contents

1 Executive Summary

Due to the presence of structured documents in archives (forms, tables) task 6.3 analyzes tables and forms. Based on the GT definition presented in Deliverable D6.7 a dataset consisting of the documents of the Passau Diocesan Archives has been created, which is also used for the evaluation and is also a basis for the Large Scale Demonstrator (LSD). Since the dataset mainly consists of hand-drawn tables a high variation of the column width and rows height is present. To be capable to deal with this variation a new approach based on D6.7 has been developed. The method detects the table region, the table columns and the header based on the line information using a specified template. Currently, the template is selected manually. This method is combined with the approach of Naverlabs to detect the rows (since separators are manually drawn or are missing, the baselines of the text are analyzed to detect the rows). A metric has also been defined for the evaluation of the detected table structure. The input is a page xml defining the table/form structure and the output is the alignment of the template to the current document image.

Section 2 describes the GT dataset, while the evaluation metric is presented in Section 3. Section 4 gives an overview of the methodology and the results of the table matching. The future work is presented in Section 5. All modules are part of the CVL READ Framework. It is Open Source under LGPLv3 and available at github: `https://github.com/TUWien/ReadFramework`.

2 Ground Truth Dataset of Handwritten Tables

The structure of tables/forms is represented as extended PAGE XML. The extension of the PAGE XML was defined in Deliverable D6.7 by CVL and Naverlabs and is also used by the table editor of Transkribus developed by UIBK. To evaluate the table analysis and for the Large Scale Demonstrator (LSD) a GT dataset has been created together with ABP and Naverlabs. The following paragraph describes the new dataset of the Passau Diocesan Archives.

The $ABP\_S\_1847 - 1878$ dataset contains information about the parishioners who died within the geographic boundaries of the various parishes of the Diocese of Passau between the years 1847 and 1878. The dataset holds a total of 26,579 scanned pages. According to the official order of the Catholic Church, the parish scribes had to record name, profession, religion, court, address, marital status, reason of death, dates of death and burial, age, names of doctor and priest as well as additional information in written form. The images display the records mainly in tabular format referring to one person per row. A thorough analysis of the dataset shows that for 22,001 images 88 different table prints were used. These unique layouts were further categorized into eleven template categories. The vast majority of scans (15,147 images) even fall into one single template category. Based on this collection a manual annotation of the table information has been done for 210 pages which is used as GT dataset.

Figure 1 shows an annotated table of the ABP dataset in Transkribus. The annotation has been done with the table editor.

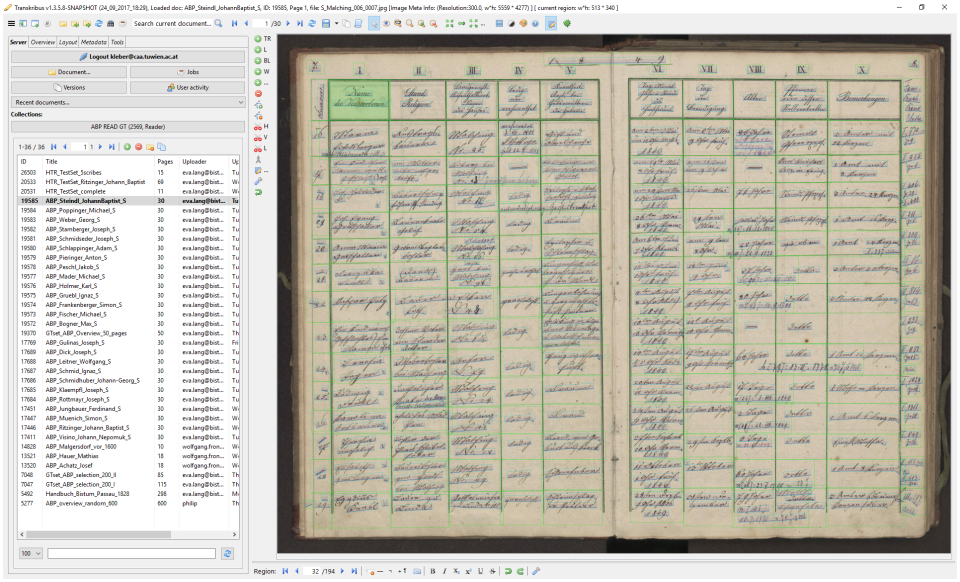

Figure 1: Table editor of the Transkribus software showing an annotated image of the ABP GT dataset.

3 Evaluation

The evaluation of the table matching is based on Shahab et al. [1] and Burie et al. [2]. Shahab et al. encode the table information directly in the image format which is similar to document image segmentation (each pixel belongs to a certain cell/row/column/table) [1]. Note that the encoding of the table as an image can be generated by using the table description in the PAGE xml. Based on these description established methods for evaluating image segmentation methods can be applied. Shahab et al. define the following measures: Correct Detections, Partial Detections, Over-Segmentations, Under-Segmentations, Missed Segments, and False Positive Detections (see [1] for a detailed description). Additionally, the Jaccard Index (JI) to measure the overlapping of a detected document region (quadrilateral) with the annotated document region in the image is used. The JI is introduced by the ICDAR2015 Competition on Smartphone Document Capture and OCR (SmartDoc) [2]. Due to the proposed methodology (table structure matching based on a *template*, see 4) the following measures of Shahab et al. and Burie et al. are used.

- Mean Cell Match (MCM): $\frac{1}{N}\sum_{i=1:N}\frac{|G_i \cap S_i|}{|G_i|}$ where G_i corresponds to the area of each cell of the GT segmentation and S_i is the corresponding cell area detected by the proposed methodology (one-to-one correspondence). N is the number of cells of the table. This corresponds to the value *Correct Detections* (the value is thresholded in [1]).
- Mean Table Match (MTM): same as MCM but only for the entire table region.
- Under-Segmentation (USeg) defines the number of cells that have a major overlap with more than one GT segment: overlap of the corresponding cell S_i with all

$G_{j \neq i} > T$.

- Missed Segments (Miss) defines the number of cells that do not have a major overlap with the corresponding detected segment (number of segments with $MCM < T$.
- Jaccard Index JI $= \frac{1}{N} \sum_{i=1:N} \frac{area(G_i \cap S_i)}{area(G_i \cup S_i)}$. The JI has a range from 0 to 1, where 1 is the best segmentation possible. For the table matching, the JI is calculated for the detected table region as well as for all table cells (mean value for all cells for one table is calculated).

4 Table Matching Methodology and Results

In the following Sections the current implementation of the form/table analysis is shortly summarized. Furthermore, the planned future work for D6.9 is presented. Additionally, the first evaluation on the presented dataset in Section 2 is shown.

4.1 Matching Table Structure Using Association Graphs

In D6.7 a method based on Beveridge and Riseman [3] has been developed to align tables based on line models to a corresponding template. Experiments have shown that especially for hand-drawn tables (e.g. documents of ABP, see Section 2) variations of the column width and the rows height (header) are present, which cannot be solved with the method of D6.7. Thus, a new table template matching has been developed. It matches the hierarchical structure of the table using an association graph (see Pelillo et al. [4] and Ishitano [5]) by finding a maximum clique [6]. Consider the matching of the table template T to a given document D. The template defines all table cells, and thus all visible (horizontal and vertical) cell borders TL_i where i is the number of cell borders. A line detection in D gives all (horizontal and vertical) lines DL_j where j is the number of detected lines in D. Each node N_i in the association graph G corresponds to a pair of horizontal/vertical lines $N_k = (TL_i, DL_j)$. To create edges between nodes, the compatibility of every combination of two pairs of nodes is examined. Two nodes $N_1 = (TL_1, DL_1)$ and $N_2 = (TL_2, DL_2)$ are compatible if they fulfill the following criteria:

- if TL_1 is left/above from TL_2 then also DL_1 must be left/above from DL_2
- if m=dist(TL_1,TL_2) and n=dist(DL_1,DL_2), then $m \times (1-T) \leq n < m \times (1+T)$ for a certain threshold T.

The largest maximal clique on G defines the best matching of document D to a given template T. The line detection in D6.7 and [7] is based on Zheng et al. [8] and also described in Diem et al. [9] is also used here.

Figure 2 shows a table image of a document from the Passau Diocesan archive. The second image shows the rough alignment of the table template with the current document. It can be seen that there are variations of the width of the columns. The rough

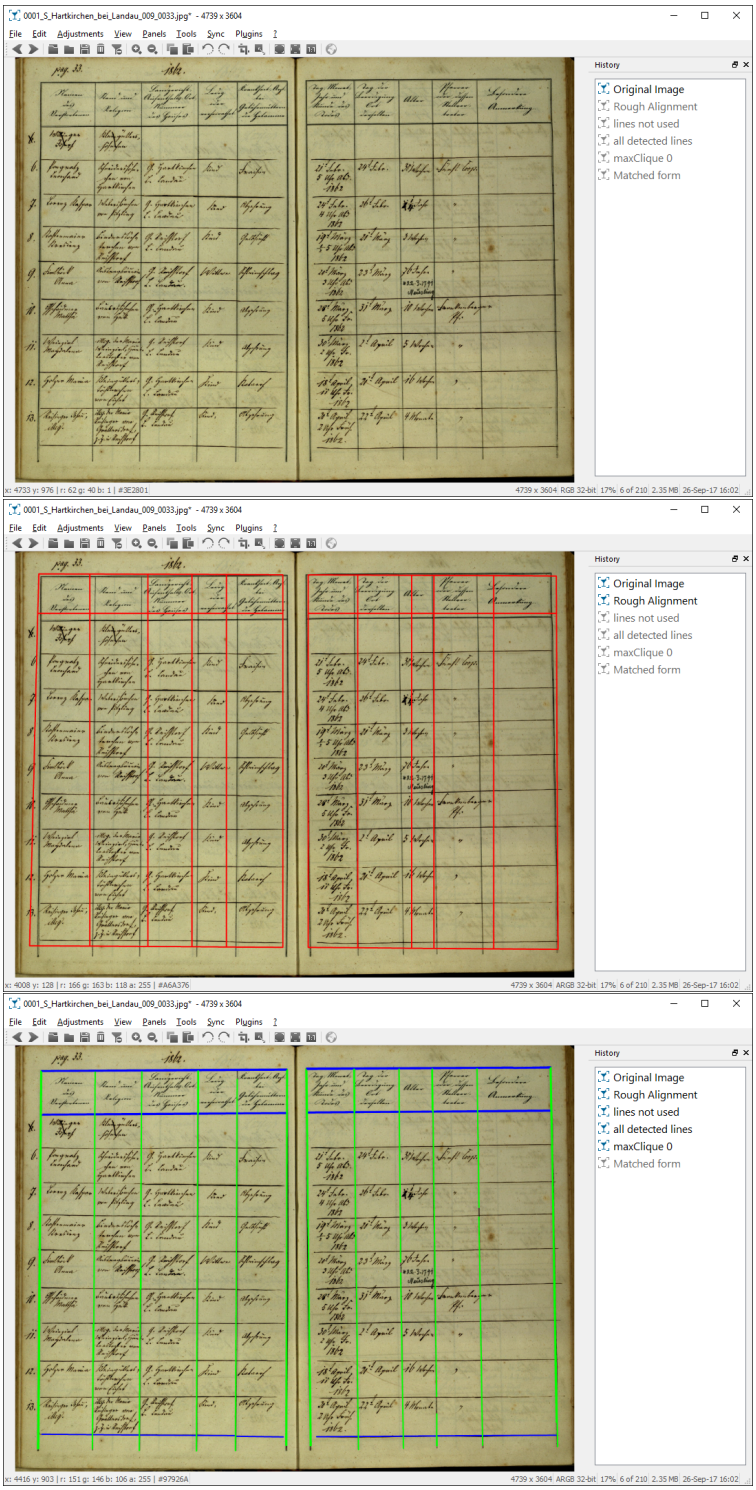

Figure 2: Example of the table matching based on a line model. The first image shows the document image, the second image shows the rough alignment of the table template to the document (red lines), and the last image shows the resulting maximal clique (blue horizontal lines, green vertical lines.

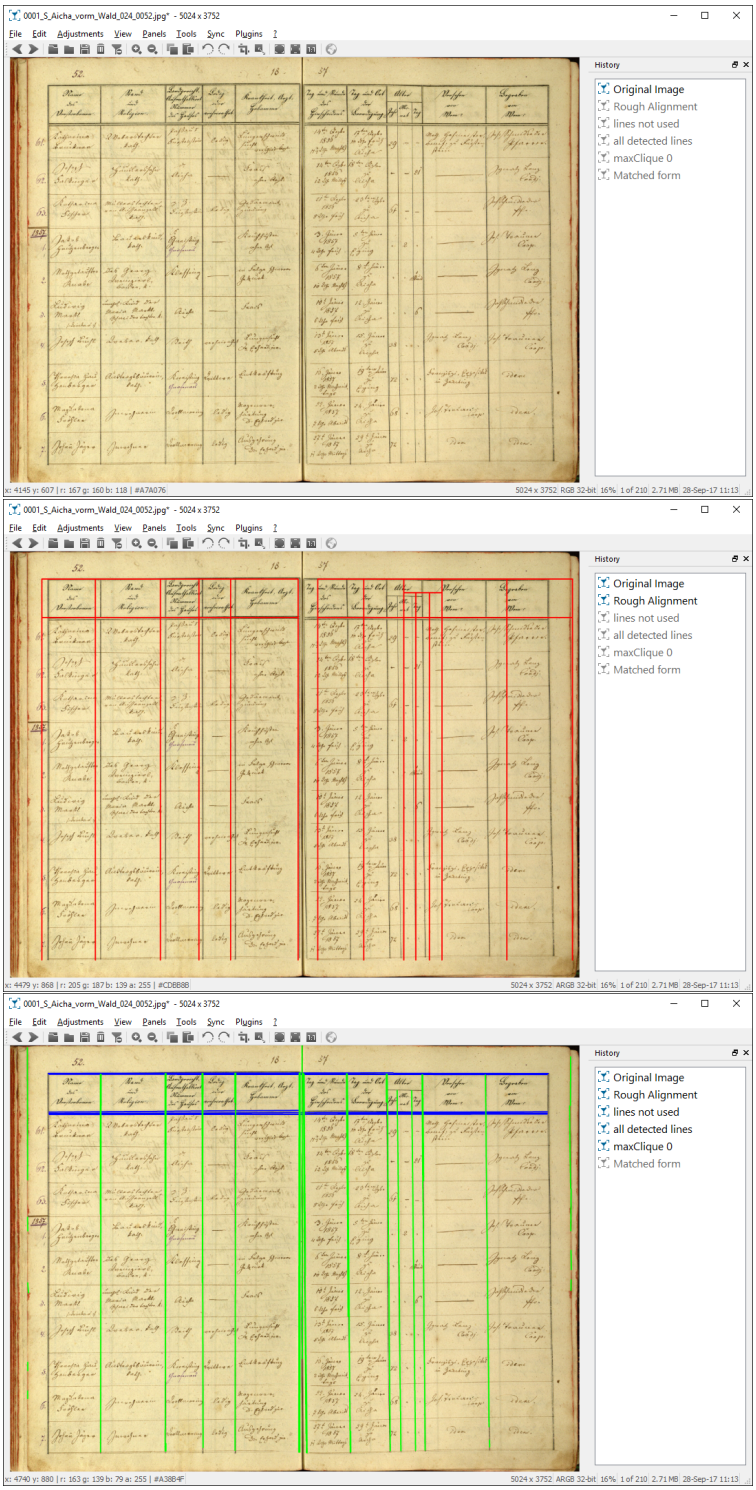

Figure 3: Another example of the table matching based on a line model and an association graph.

alignment is done by a correlation of the template image and the document image. Afterwards, the association graph is calculated and the maximum clique of the graph is

visualized in the last image. Figure 3 shows a second example.

Based on the columns and the table region, the table rows are detected based on the detected baselines of the Basic Layout Analysis (Task 6.2). The detection of the rows is done in Task 6.5 (Document Understanding, D6.14).

4.2 Results

The evaluation of the table matching is done on the dataset described in Section 2. The metric is explained in Section 3 and the results are shown in Table 1.

Table 1: Evaluation of the proposed table matching.

	ABP_GT dataset
MTM	0.9785
JI (Table)	0.9305
MCM	0.8936
JI (Cell)	0.8754
USeg	0.0796
Miss	0.0566

5 Future Work

To avoid a binarization of the image and the resulting problems, a line detector using the gradient information of gray value images will be used. Thus, a state of the art line detection based on von Gioi [10] will be implemented in the READ framework for D6.9. Additionally, the current methodology will be enhanced to avoid errors for ambigous cell borders (can occur if two lines are detected at the table borders). A weighted maximum clique method will be tested to achieve better results. Also an automated template selection will be tested in D6.9.

Currently, also a command line interface for the table module exists, which will be the basis for the integration task into Transkribus. The integration to Transkribus will be realized in the next deliverable.

References

[1] A. Shahab, F. Shafait, T. Kieninger, and A. Dengel, "An open approach towards the benchmarking of table structure recognition systems," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. New York, NY, USA: ACM, 2010, pp. 113–120. [Online]. Available: http://doi.acm.org/10.1145/1815330.1815345

[2] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. M. Luqman, M. Mehri, N. Nayef, J. Ogier, S. Prum, and M. Rusinol, "Icdar2015 competition on smartphone document capture and ocr (smartdoc)," in *13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug 2015, pp. 1161–1165.

[3] J. R. Beveridge and E. M. Riseman, "How easy is matching 2d line models using local search?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 564–579, Jun 1997.

[4] M. Pelillo, K. Siddiqi, and S. W. Zucker, "Matching hierarchical structures using association graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1105–1120, Nov 1999.

[5] Y. Ishitani, "Model matching based on association graph for form image understanding," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, Aug 1995, pp. 287–292 vol.1.

[6] J. Konc and D. Janezic, "An improved branch and bound algorithm for the maximum clique problem," *MATCH Commun. Math. Comput. Chem.*, vol. 58, pp. 569–590, 2007.

[7] F. Kleber, M. Diem, and R. Sablatnig, "Form Classification and Retrieval using Bag of Words with Shape Features of Line Structures," in *Document Recognition and Retrieval XXI*, 2014.

[8] Y. Zheng, C. Liu, X. Ding, and S. Pan, "Form frame line detection with directional single-connected chain," in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2001, pp. 699–703.

[9] M. Diem, F. Kleber, and R. Sablatnig, "Document Analysis Applied to Fragments: Feature Set for the Reconstruction of Torn Documents," in *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, D. Doermann, V. Govindaraju, D. Lopresti, and P. Natarajan, Eds., Boston, USA, June 2010, pp. 393–400.

[10] R. G. von Gioi, J. Jakubowicz, J. M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 722–732, April 2010.