Matching Table Structures of Historical Register Books using Association Graphs

Florian Kleber
Markus Diem
Computer Vision Lab
TU Wien
Vienna, Austria
kleber@cvl.tuwien.ac.at

Herve Dejean
Jean-Luc Meunier
Naver Labs Europe
Meylan, France
firstname.lastname@naverlabs.com

Eva Lang
Archiv des Bistums Passau
Bischoefliches Ordinariat Passau
Passau, Germany
eva.lang@bistum-passau.de

***Abstract*—In this paper we present a template-based table structure matching using association graphs for handwritten/printed historical documents. The recognition of the table structure consisting of column and header information is the prerequisite for the subsequent row detection and handwritten text recognition used for information extraction. The table matching is done by detecting the maximum clique in an association graph, which represents the matching of the line information of the template and a document of interest. This allows for variations of widths and heights of rows and columns. The presented methodology is evaluated on historical register books (death records) of the Archive of the Diocese of Passau. The method shows a reliable detection of the structure of handwritten/printed tables with a mean cell match of 88.28%.**

***Index Terms*—document image analysis, table recognition, table matching**

I. Introduction

Tables and forms are structured documents, that is, the representation of structured data. Couasnon and Lemaitre define tables as "*prevalent means of representing and communicating structured data*" [1] with different content (words, numbers, formulae, . . .) existing of different formatting (e.g. handwritten vs. printed). For automated information extraction in document analysis, table detection and recognition is needed to obtain the physical and logical structure of tables. Based on the known structure of a table also the results of Document Image Analysis (DIA) tasks can be improved, e.g. baseline detection or automated text recognition [1].

Documents with information stored in tables can be found in archives and libraries providing data for research in biology (e.g. herbarium of the museum of natural history in Vienna), meteorology (e.g. weather and temperature information from weather stations of the last ˜100 years) and e.g. family search (e.g. death records of the Archive of the Diocese of Passau). The digitization of libraries and archives is an ongoing process. In Germany, the conference of the head of the archive administration of the German federation (*Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder (KLA)*[1] states in a recommendation letter [2] that archives plan a medium-term digitization of 5-10% of archive holdings, which shows the ongoing digitization of archives. Thus, there is a need for the automated information extraction in historical documents.

In this paper a template-based table structure matching is presented to detect the header and column information in archival table documents. The use of templates allows to define the logical table structure for a set of documents with a recurrent layout which is the case for archival documents used for data logging. Using a template allows variations in the physical layout of columns, rows or cells (variation of the height/width) but defines a fixed number of rows and columns and the logical layout. The matching is based on the information of the separators (visible lines) and the correspondence is modelled with an association graph [3], [4]. After the detection of the table structure, a baseline detection can be performed and the baselines are attributed to the according cells (also split if necessary). The row detection is based on Clinchant et al. [5] which detects rows using the baseline information. The presented methodology is evaluated on a historical dataset, which will be published as part of a competition dataset on table recognition planned in 2018 within the context of the READ project[2]. In comparison to the well-known UNLV and UW3 datasets, the proposed dataset focusses on historical handwritten documents.

Historical documents are challenging due to the handwritten content and even hand drawn tables with varying table layouts compared to printed tables. Couasnon and Lemaitre [1] present an overview of recognition of tables and forms. Martinat et al. [6] present a table recognition based on a table description language. For the table recognition intersections of separators are used. A matching of 2D line models is presented by Beveridge and Riseman [7] using local search. A similarity

978-1-5386-5875-8/18/$31.00 ©2018 IEEE.

[1] https://www.bundesarchiv.de/fachinformationen/kla/index.html.de, accessed 28.08.2017

[2] see https://read.transkribus.eu/

transformation allows rotations, translations and a scaling of the data. However, non-rigid transformations (i.e. variations of the column width and the row height) of the line model cannot be solved. The latest approaches use deep learning for table detection and recognition [8], [9]. These approaches do not use a priori knowledge of the table structure which may lead to different results, e.g. number of columns.

This paper is organized as follows: Section II shortly presents the historical dataset consisting of handwritten tables. The methodology as well as a short summary of the row detection is presented in Section III-A and Section III-B. The results are discussed in Section IV; then follows the conclusion.

II. Dataset

The evaluation is done on a subset of the ABP_S_1847-1878 dataset provided by the Passau Diocesan Archives, which 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 full ABP_S_1847-1878 dataset holds a total of 26,579 scanned pages. The scans originate from 212 pastoral districts (mainly parishes) with their own record keeping in the time between the uproar of 1848 and the beginning of the German Empire in 1871.

The register book pages show table format where according to the official order of the state, the parish scribes had to manually 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. Each entry mainly refers to one row in the table.

A thorough analysis of the dataset shows that for 22,001 images 88 different printed table schemes 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. On 4,578 pages, the requested information was recorded in manually drawn tables or manually extended table prints. The images are openly available through the matricula-online platform[3], records can be queried using a search engine supplied by the Diocese of Passau[4]. The data are used by family historians as well as by historian scholars interested in the age of those who died, the development or the spread of deadly diseases, etc.

The evaluation subset of data consists of 142 documents with 5 different table layouts (number of pages per layout in the evaluation set: 104, 13, 20, 2, 3), where each table is manually annotated. Additionally, each baseline is marked-up and available in the Ground Truth (GT) set.

Figure 1 shows one document of the dataset with the manually annotated table. Also the baseline information is available in the GT. Note that for the evaluation of the table matching only the column and header information is relevant. The table rows are detected and evaluated in a second step (see Section III-B). The dataset will be published as part of a competition dataset on table recognition planned in 2018 within the context of the READ project [5]. The GT is represented as extended PAGE XML (see https://github.com/Transkribus/TranskribusPageformat). For a description of the original PAGE XML see http://www.primaresearch.org/tools/PAGELibraries.

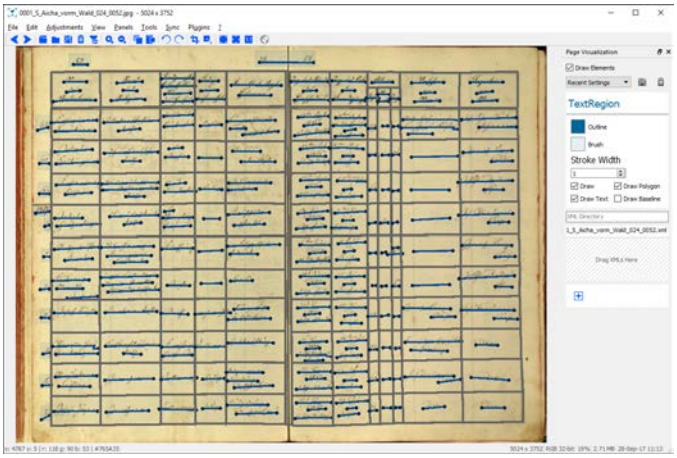

Fig. 1. Exemplary document of the ABP.S.1847-1878 dataset with annotations of table elements and baselines.

III. Table Recognition

In this section, the table structure matching using association graphs is presented. As a second part, the work of Clinchant et al. [5] is shortly summarized which allows for row detection based on detected columns of the presented table matching and detected baselines.

A. Table Structure Matching Using Association Graphs

The proposed methodology matches the table structure of a given table template based on the visible line information (separators). The table structure of the template is represented as extended PAGE XML (see Section II) defining the physical and logical layout of a representative sample document. The template information comprises the table columns and the table header.

The matching of relational structures can be solved by "*transforming it into the equivalent problem of finding the maximum clique in an auxiliary graph structure, known as the association graph*" [3]. A clique is a subset of vertices, where each vertex is adjacent to all others. A maximal clique is a clique, which cannot be subgraph of a larger clique, whereas a maximum clique is a clique with the largest number of vertices. It has been shown, that the maximum clique problem is NP-complete [3]. A basic algorithm for finding maximal cliques in an undirected graph is e.g. the Bron-Kerbosch algorithm. To find the maximum clique the MaxCliqueDyn algorithm of Konc and Janezic [10] has been used (see https://gitlab.com/janezkonc/mcqd). MaxCliqueDyn was

[3]http://data.matricula-online.eu/de/deutschland/passau/

[4]http://gendb.bistum-passau.de/

[5]see Scriptnet competitions https://scriptnet.iit.demokritos.gr/competitions/

originally developed with the purpose of quickly comparing protein structures and is faster than the algorithms of Tomita et al. and Oestergard on DIMACS and random graphs [10]. The result of Konc and Jenezic [10] is one maximum clique.

The proposed methodology is based on work of Ishitani [4]. As a first step, the rough alignment of the table is determined by a correlation of the template image and the document image. Then, 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 (see also Figure 2). 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 from/above TL_2 then also DL_1 must be left from/above DL_2
- if m=dist(TL_1,TL_2) and n=dist(DL_1,DL_2), then $m \times (1 - Th) \leq n < m \times (1 + Th)$ for a certain threshold Th.

The largest maximal clique on G defines the best matching of document D to a given template T. For the structural matching, a separate graph for the horizontal and vertical lines is modeled. The line detection in Kleber et al. [11], based on Zheng et al. [12] and also described in Diem et al. [13], is used here as well.

Figure 2 shows exemplarily the line information of a table template (left) and a document of interest (same table category). All vertical lines of the template and the document are used to model the *vertical* association graph. The maximum clique in this example is

$$\{(TL_1, DL_1), (TL_2, DL_2), (TL_3, DL_3), (TL_4, DL_4), (TL_5, DL_5)\} \quad (1)$$

In comparison to Ishitani [4] colinear lines are merged to avoid errors in the final table structure.

Matching each cell border separately leads on the one hand to larger graphs, which can be time-consuming to solve (NP-completeness), and on the other hand leads to non-aligned matches due to noise in the table images. This is illustrated in Figure 3.

Thus, compared to Ishitani [4], a line clustering of co-linear lines of neighboring cells is done to avoid alignment errors in the final table structure. Additionally, constraints based on the distance of parallel lines and the line length of the detected line compared to the matched line are introduced to minimize the number of edges in the association graph. Thus, the matching of co-linear cell borders in the graph and the line constraints minimize the matching errors. If a cell border defined in the template is not part of the association graph, an artificial line is introduced to retain the table structure defined in the template.

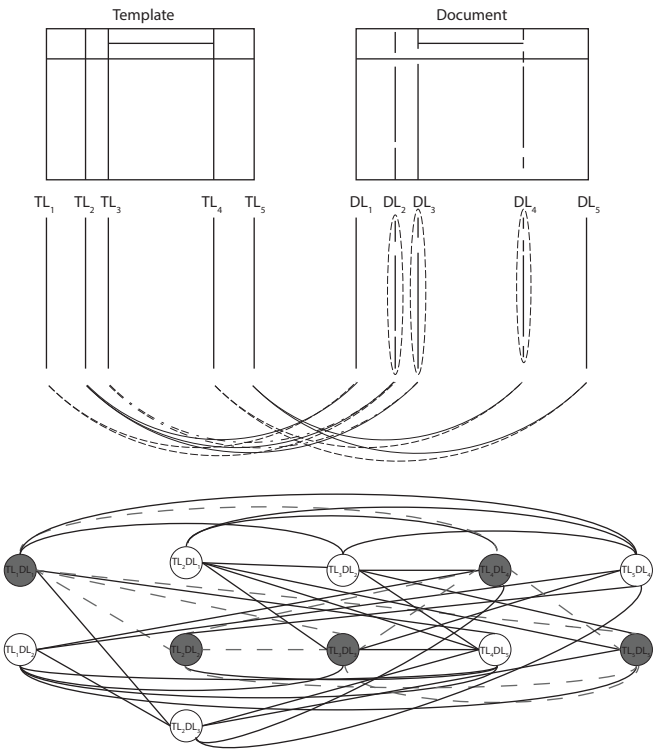

Fig. 2. Example of table structure matching based on a line model and the vertical association graph (maximum clique is represented by dotted edges and gray nodes). In the middle, the possible line associations (nodes) of the vertical lines are shown.

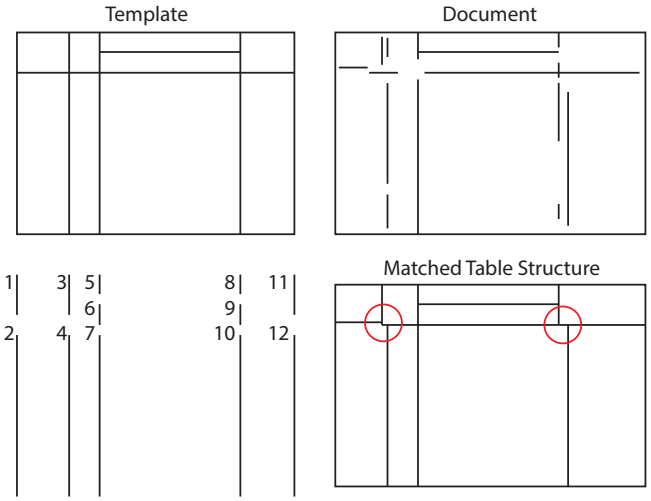

Fig. 3. Example of table structure matching based on a line model for each table cell and resulting alignment errors.

The artifical line in the detected document is inserted in order to keep the ratio of the distances of the cell borders as defined in the template.

Based on the columns and the header of the matched table, the table rows are detected using baseline information of the basic layout analysis [14]. The detection of the rows is

presented in [5] and summarized in the following section.

B. Table Row Detection

In this section a short summary of Clinchant et al. [5] is presented. The table matching and the baseline detection [14] is the main prerequisite of the row detection. This is a challenging task for this collection for various reasons: First, separators (hand-drawn lines) are not used systematically for delimiting the rows (and are anyway recognized imperfectly). Then, the row layout depends on each writer and can vary inside one single record book. Some writers minimized the space between two rows, and, using a thin handwriting, also scribbled each record onto a single-line table row, while other writers made use of more space, and preferred centered lines, with some cells far longer than the others.

Fig. 4. BIESO labels on text lines for a table image (orange, green, grey, yellow and purple), courtesy of [5]

The row detection problem is formulated as follows: Once the columns and the text lines have been identified, each textline will be tagged with one of the following categories: B(eginning), I(nside), E(nd), S(ingleton), O(utside of the table), which correspond of the position of the text line in the cell. Figure 4 shows an example of the annotation. This BIESO pattern is taken from the Natural Language Processing domain, and is used in order to recognize entities (sequence of words) in a sentence. Based on the BIESO categorization, a pattern-mining based step regroups text lines belonging to the same row. Associated with the Column segmentation, it allows us to segment a table into cells. Two graph-based approaches for categorizing text lines into BIESO categories are compared: Conditional Random Fields (CRF) and Graph Convolutional Networks (GCN). It has been shown that the CRF approach is slightly better than the GCN based approach (for details see [5]).

IV. Evaluation

The evaluation of the table structure matching is based on Shahab et al. [15] and Burie et al. [16]. 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) [15]. Note that the encoding of the table as an image can be generated by using the table description in the extended PAGE XML (see Section II). Based on this description established methods for evaluating image segmentation methods can be applied. Shahab et al. define the following measures (see [15] for a detailed description):

- Correct Detections
- Partial Detections
- Over-Segmentation
- Under-Segmentations
- Missed Segments
- False Positive Detections

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 ICDAR2015 Competition on Smartphone Document Capture and OCR (SmartDoc) [16] introduces the JI. Due to the proposed methodology (table structure matching based on a template, see III-A) the following measures of Shahab et al. and Burie et al. are used (e.g. due to the a-priori knowledge of the table structure, the template, no missing segments are possible). The table region is defined as the bounding rectangle of the matched table.

- 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 [15]).
- 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 $G_{j\neq i} > T$, $T = 0.2$.
- Missed Segments (Miss) defines the number of cells that do not have a major overlap with the corresponding detected segment (number of segments with $MCM < T$), $T = 0.2$.
- 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).

The methodology is evaluated on the dataset described in Section II. In overall 142 documents with 5 different table layouts have been used. The GT was annotated manually. The results of the proposed methodology are summarized in table I. It can be seen that the MCM is 88.28% which shows a reliable

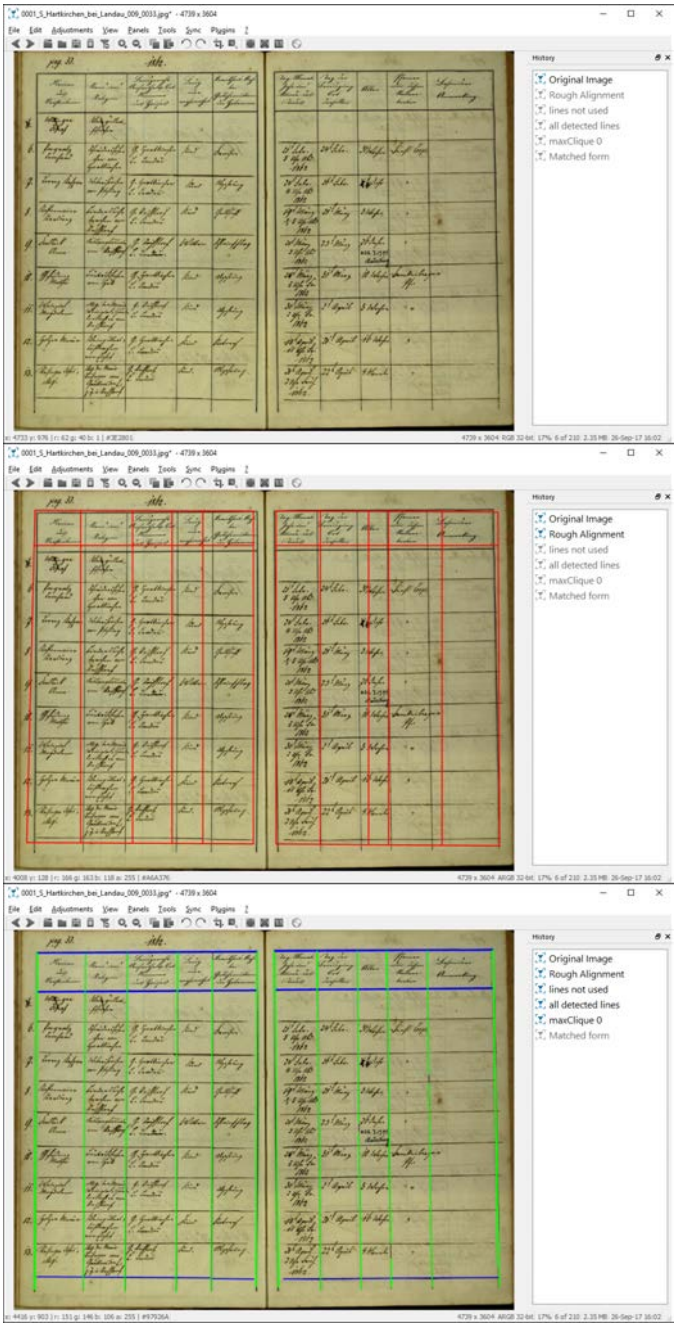

Fig. 5. Example of the table matching based on a line model. The first image shows the document image, the second image shows the rough alignment of the table template to the document (red lines), and the last image shows the resulting maximal clique (blue horizontal lines, green vertical lines).

table matching. The errors results from the outermost borders, which can be detected as the documents borders, which lead to a wider first and last column.

Figure 5 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 alignment is done by a correlation of

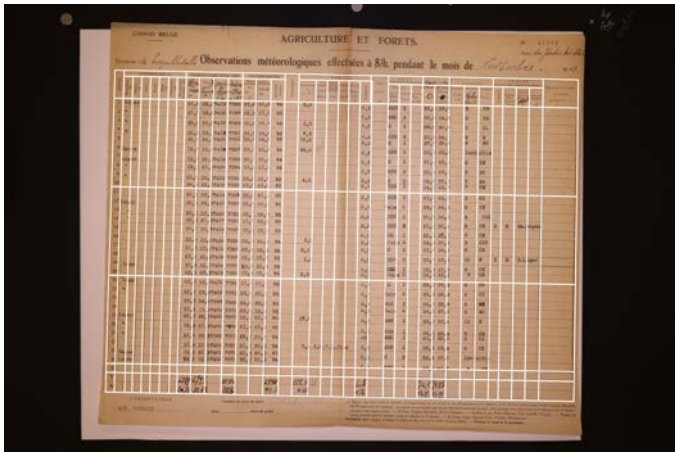

Fig. 6. Example of table structure matching for a different collection (not in the evaluation set) with a more complex table layout.

TABLE I
EVALUATION OF THE PROPOSED TABLE MATCHING.

	ABP_GT dataset
MTM	0.9785
JI (Table)	0.9305
MCM	0.8828
JI (Cell)	0.8374
USeg	0.0545
Miss	0.0761

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 7 shows a second example. Figure 6 shows the result of the proposed methodology for a different type of collection. It can be seen that a higher number of rows/columns is present compared to the documents in the Passau dataset (see Section II). The white lines show the matched table layout.

V. CONCLUSIONS

In this paper, a template-based matching of table structures has been presented. The use of a template allows modeling the layout of the table which avoids the merging or splitting of columns based on the content. The definition of a template is feasible, since archives and libraries hold collections with thousands of documents with the same style. The varying table layout and the handwritten content variations of the documents must be considered.

It has been shown that a reliable matching with a mean cell match of 88.28% can be performed by modelling the correspondences with an association graph and by finding the maximum clique. Based on the matching and the baseline detection the row detection of Clinchant et al. can be performed to detect the rows. The proposed method can be used for handwritten table documents with varying column/row widths.

As future work, a weighted maximum clique method will be tested to achieve better results. Using a weighted maximum

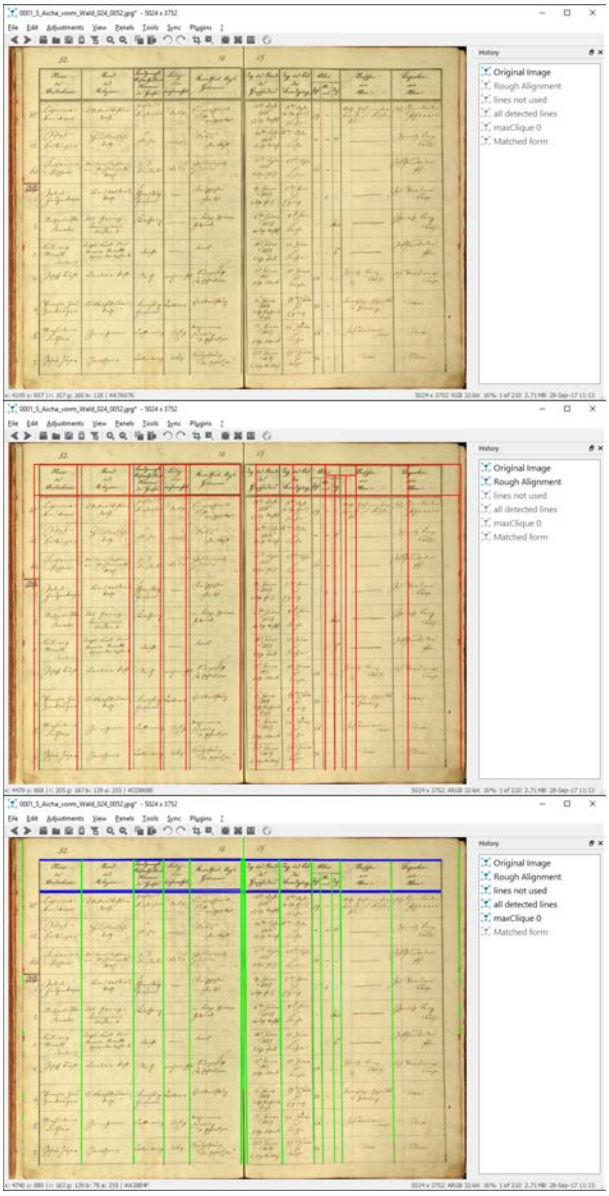

Fig. 7. Another example of the table matching based on a line model and an association graph.

clique method each node can be assigned a weight according to the line match. This will avoid errors for ambiguous cell borders (can occur if two lines are detected at the table borders). Additionally, it is planned to publish the table dataset containing historical documents in conjunction with a competition.

ACKNOWLEDGMENT

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

REFERENCES

[1] B. Couasnon and A. Lemaitre, "Recognition of Tables and Forms," *Handbook of Document Image Processing and Recognition*, 2014.

[2] Fototechnischer Ausschuss der KLA, "Wirtschaftliche Digitalisierung in Archiven, Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Laender (KLA)," 2016.

[3] 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.

[4] 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.

[5] S. Clinchant, H. Dejean, J.-L. Meunier, E. Lang, and F. Kleber, "Comparing machine learning approchaes for table recognition in historical register books," in *In Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS 2018)*, 2018.

[6] I. Martinat, B. Couasnon, and J. Camillerapp, "An adaptative recognition system using a table description language for hierarchical table structures in archival documents," in *Graphics recognition: Recent advances and new opportunities, LNCS 5046*, 2008.

[7] 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.

[8] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, "Deepdesrt: Deep learning for detection and structure recognition of tables in document images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov. 2017, pp. 1162–1167. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICDAR.2017.192

[9] S. F. Rashid, A. Akmal, M. Adnan, A. A. Aslam, and A. Dengel, "Table recognition in heterogeneous documents using machine learning," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov. 2017, pp. 777–782. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICDAR.2017.132

[10] 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.

[11] F. Kleber, M. Diem, and R. Sablatnig, "Form Classification and Retrieval using Bag of Words with Shape Features of Line Structures," in *Proceedings of Document Recognition and Retrieval XXI*, Bertrand Coüasnon and Eric K. Ringger, Eds. SPIE, 2014, pp. 902 107–1 – 902 107–9.

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

[13] 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.

[14] T. Grüning, G. Leifert, T. Strauss, and R. Labahn, "A robust and binarization-free approach for text line detection in historical documents," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov. 2017, pp. 236–241. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICDAR.2017.47

[15] 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

[16] 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.