# *Matching Table Structures of Historical Register Books using Association Graphs*

Florian Kleber, Fabian Hollaus,
Herve Dejean, Jean-Luc Meunier and Eva Lang

## INTRODUCTION

A template-based table structure matching using association graphs for handwritten/printed historical documents is proposed.

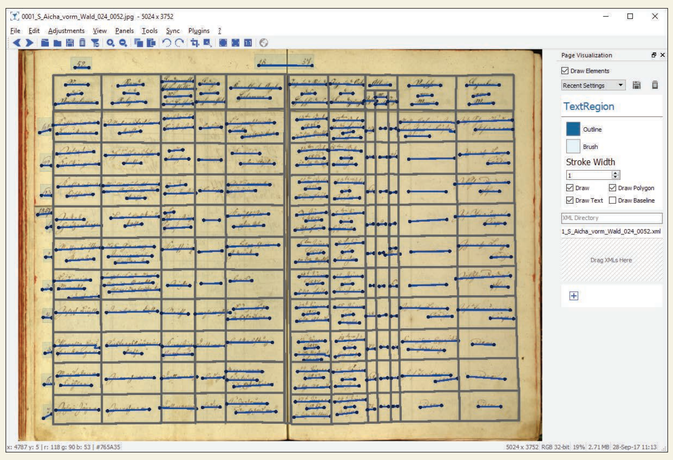
- Allows for variations of widths and heights of rows and columns
- Evaluation is done on historical register books (death records) of the Archive of the Diocese of Passau

### Motivation

Tables are prevalent means of representing and communicating structured data [Couasnon and Lemaitre].
Thus, for automated information extraction in document analysis, table detection and recognition is needed.

### Dataset

- Dataset is a subset of the AB_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.
- Scans originate from 212 pastoral districts.
- 142 documents with 5 different table layouts manually annotated.
- Exemplary document of the dataset with annotation of the table elements and baselines:

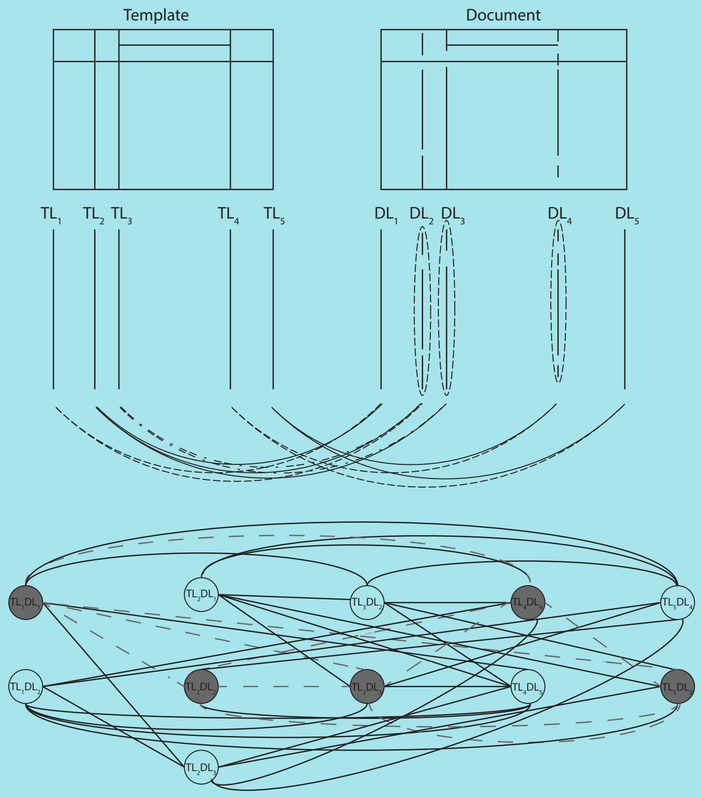



## METHODOLOGY

- Recognition of the table structure consisting of column and header information is done by detecting maximum clique in an association graph
- The association graph represents the matching of the line information (printed or hand-drawn table layout) of the template and the document of interest
- Table matching is a prerequisite for the subsequent row detection
- Row detection is based on Clinchant et al.

### Table Template Matching

- Matching of relational structures can by solved by transforming it into the equivalent problem of finding the maximum clique in an auxiliary graph structure, known as the association graph.
- The following image shows an 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.




### Table Row Detection [Clinchant et al.]

- Detects rows using baseline information
- Each text line is tagged as: Beginning, Inside, End, Singleton or Outside (BIESO)
- Based on the BIESO categorization, a pattern mining based step regroups text lines belonging to the same row.
- The following image shows the BIESO labels on text lines for a table image:

## OUTLOOK

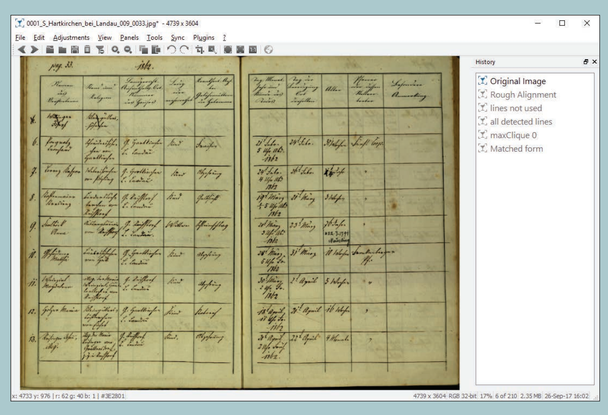
As future work, a weighted maximum clique method will be tested to achieve better results. Using a weigthed maximim clique method, each node can be assigned a weight according to the line match. Thus, errors for ambiguous cell borders will be avoided.

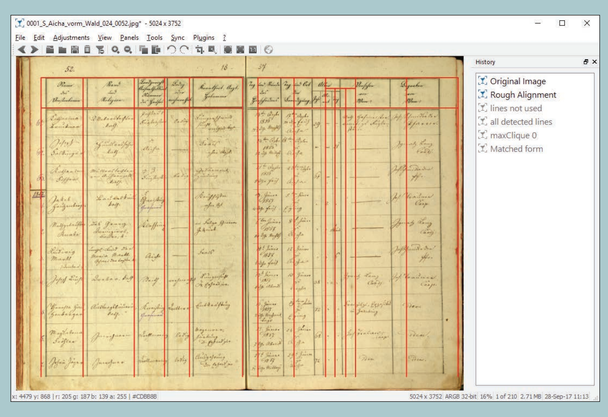
It is planned to publish the table dataset containing historical documents in conjunction with a competition. This will be done within the H2020 project READ ("Recognition and Enrichment of Archival Documents").

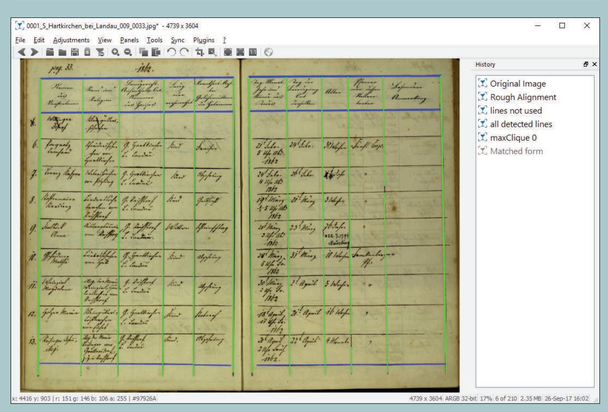
### Evaluation

- Dataset consists of 142 documents with 5 different table layouts.
- Mean Table Match (MTM), Mean Cell Match (MTM) and the Jaccard Index (JI) are used as evaluation measures.
- Results:

| | ABP_GT DATASET |
|---|---|
| MTM | 0.9875 |
| JI (TABLE) | 0.9305 |
| MCM | 0.8828 |
| JI (CELL) | 0.8374 |




First image shows the document image, the second image shows the rough alignment of the template and the last image shows the result of the maximum clique alignment.

https://read.transkribus.eu/

florian.kleber@tuwien.ac.at
This poster is presented at ICFHR 2018, Niagara Falls, USA.

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