READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D6.7
Table and Form Analysis Tool P1

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	31.12.2016
Actual date of delivery	28.12.2016
Date of last update	21.12.2016
Deliverable number	D6.7
Deliverable title	Table and Form Analysis Tool P1
Type	Report, Demonstrator
Status & version	in progress
Contributing WP(s)	WP6
Responsible beneficiary	CVL
Other contributors	CVL, UIBK, XRCE
Internal reviewers	XRCE,NCSR
Author(s)	Florian Kleber, Markus Diem and Stefan Fiel
EC project officer	Martin MAJEK
Keywords	table analysis, forms analysis

Contents

1 Executive Summary

Due to the presence of structured documents in archives (forms, tables) task 6.3 analyzis tables and forms. The line information as well as the basic layout will be used for form classification and matching, resulting in the layout definition of the current form document. This information will be used to extract pre-printed and filled-in data, since the syntactical knowledge allows to extract the semantic information of forms. Previously published work uses the shape context of the line information to assign document images to certain templates. The methodology has been published in Kleber et al. [1]. Since the proposed approach uses only the line information the classification can have errors for similar form types and additionally the result provides no information about the alignment of the template to the classified document. Thus, a combined approach using line information and the basic layout will be developed within T6.3. It will be one tool that can be applied to one entire image (a form document or a single table document). The first version will align a document image to a specified template (the template is selected by the user). The second version of the tool will automatically select the corresponding template. Thus, a classification of the document type will be developed. The task of table detection/analysis within a document is done in T6.5 Document Understanding. 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. The module is part of the CVL READ Framework. It is Open Source under LGPLv3 and available at github: `https://github.com/TUWien/ReadFramework`.

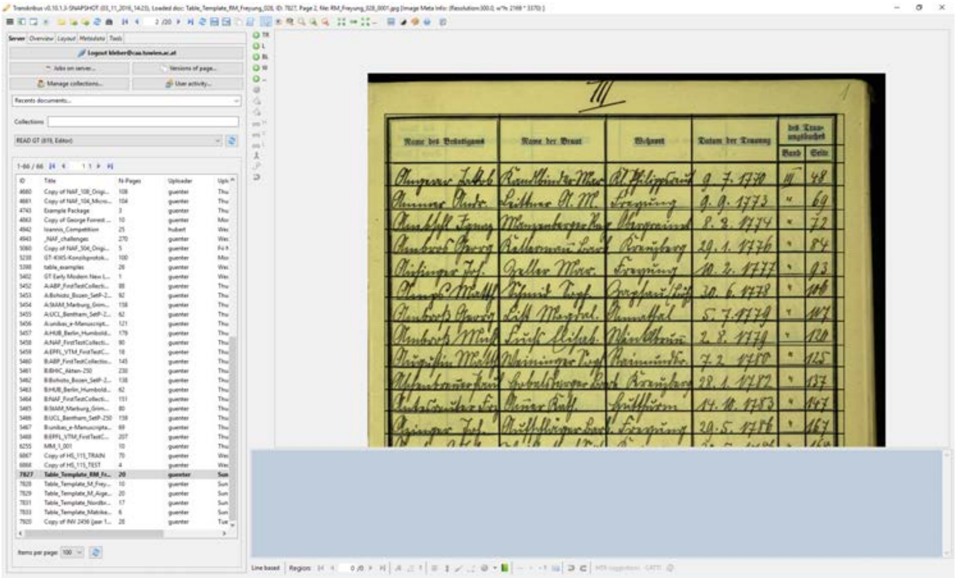
2 Form/Table Data and GT Definition

To allow the definition of the GT for document images containing forms/tables, the form/table structure must be specified. CVL and XRCE have defined the GT requirements/specification for form and table analysis (including requirements for document understanding tasks). Thus, rows, columns, single cells and the information of preprinted text in headers together with the line (cell border) information is marked as GT. The GT is stored in PAGE XMLs and UIBK has developed a form/table editor for the Transkribus software (see D4.1). Figure 1 shows an image of the form/table editor together with the GT information of a sample page. The form/table editor allows the creation of datasets together with their corresponding GT (see D4.1 for a description of the editor).

The requirements of the GT specification is also based on [2] and available datasets like *NIST Structured Forms Reference Set of Binary Images*[1] and UW3 and UNLV[2]. In D6.8 a dataset of form and table documents will be selected from different collections and the GT will be created using the table editor. The user interface allows to define the tabular structure of a document, including the labeling of the header information.

[1] see `http://www.nist.gov/srd/nistsd2.cfm`

[2] see `http://www.iapr-tc11.org/mediawiki/index.php/Table_Ground_Truth_for_the_UW3_and_UNLV_datasets`



Transkribus v0.10.1.3-SNAPSHOT (03_11_2016_14:23), Loaded doc: Table_Template_RM_Freyung_028, ID: 7827, Page 2, file: RM_Freyung_028_0001.jpg [Image Meta Info: (Resolution:300.0, w*h: 2169 * 3370)]

ID	Title	N-Pages	Uploader	Upl.
4680	Copy of NAF_108_Origi...	108	guenter	Thu
4681	Copy of NAF_104_Micro...	104	guenter	Thu
4743	Example Package	3	guenter	Thu
4863	Copy of George Forrest ...	10	guenter	Mo
4942	Ioannis_Competition	25	hubert	We
4943	_NAF_challenges	270	guenter	We
5060	Copy of NAF_504_Origi...	5	guenter	Fri
5238	GT-KWS-Konzilsprotok...	100	guenter	Mo
5398	table_examples	26	guenter	We
5402	GT Early Modern New L...	1	guenter	We
5452	A:ABP_FirstTestCollecti...	88	guenter	Thu
5453	A:Bohisto_Bozen_SetP-2...	92	guenter	Thu
5454	A:StAM_Marburg_Grim...	138	guenter	Thu
5455	A:UCL_Bentham_SetP-2...	62	guenter	Thu
5456	A:unibas_e-Manuscript...	121	guenter	Thu
5457	A:HUB_Berlin_Humbold...	179	guenter	Thu
5458	A:NAF_FirstTestCollecti...	90	guenter	Thu
5459	A:EPFL_VTM_FirstTestC...	18	guenter	Thu
5460	B:ABP_FirstTestCollectio...	145	guenter	Thu
5461	B:BHIC_Akten-250	230	guenter	Thu
5462	B:Bohisto_Bozen_SetP-2...	138	guenter	Thu
5463	B:HUB_Berlin_Humbold...	62	guenter	Thu
5464	B:NAF_FirstTestCollecti...	151	guenter	Thu
5465	B:StAM_Marburg_Grim...	80	guenter	Thu
5466	B:UCL_Bentham_SetP-250	138	guenter	Thu
5467	B:unibas_e-Manuscripta...	69	guenter	Thu
5468	B:EPFL_VTM_FirstTestC...	207	guenter	Thu
6255	MM_1_001	10	guenter	Thu
6867	Copy of HS_115_TRAIN	70	guenter	We
6868	Copy of HS_115_TEST	4	guenter	We
7827	**Table_Template_RM_Fr...**	**20**	**guenter**	**Sun**
7828	Table_Template_M_Frey...	10	guenter	Sun
7829	Table_Template_M_Auge...	20	guenter	Sun
7831	Table_Template_Nordbr...	17	guenter	Sun
7833	Table_Template_Matrike...	6	guenter	Sun
7920	Copy of INV 2456 (per 1...	28	guenter	Tue

Figure 1: Table Editor of the Transkribus Software (see D4.1)

3 Methodology

In the following Sections the current implementation of the form/table analysis is presented. Furthermore, the planned future work for D6.8 and D6.9 is presented. The previous form classification presented in Kleber et al. [1] has an overall accuracy of 87.11% for 8 different form templates (see [1]). It can be seen that a higher amount of form types will lead to a lower classification rate. Thus, a combined approach will be developed within D6.7-D6.9.

3.1 Document Matching based on Line Information

To be able to align tables/forms based on line models a method based on Beveridge and Riseman has been developed [3]. The previous approach (Kleber et al. [4]) extracts lines from a binary image. The line detection in [1] is based on Zheng et al. [5] and also described in Diem et al. [4]. The matching uses the line model of a template (defined with the table editor) and calculates a similarity based on minimizing the perpendicular distance (see [3]) of all lines. Compared to general line models, only horizontal and vertical lines are used to reduce the search. The matching approach is similar to Chamfer matching and the similarity value states how well all lines are aligned and correctly matched. The proposed approach allows also a slight misalignments between the line template and current model. This is necessary, since in manually drawn forms, the height of rows can vary dependent on the cell contents (see also Figure 2).

Figure 2 shows two form documents of the same form class (left and middle image). The left image is the form template and the line model is shown in green. The middle image is a different form document of the same class and the line model is shown in blue. The aligned image (right side) shows the line model (red and green lines) of the document image, aligned to the template using the proposed methodology.

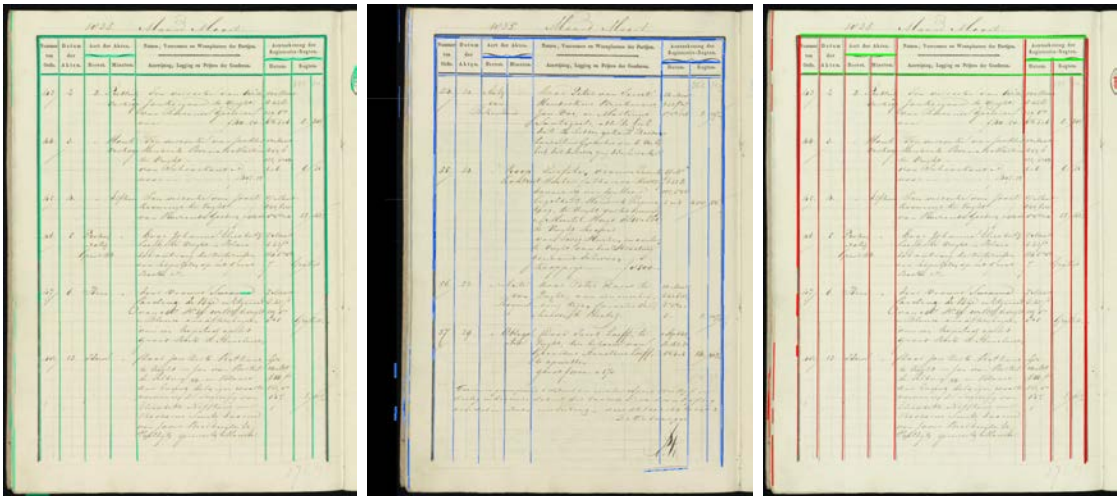


Figure 2: Example of the Form Matching based on a line model. The first image shows the detected lines of the template (green), the second image shows the detected lines of the form image to align (blue), and the last image shows the aligned horizontal and vertical lines of the template to the new document image.

3.2 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 [6] will be implemented in the READ framework for D6.8. Furthermore, the basic layout information (and also text classified in preprinted and handwritten) based on Task 6.2 will be used for an enhanced alignment of forms/tables. An evaluation of the implemented form analysis will be done on the planned table dataset. Furthermore, an interface to the Transkribus Software will be realised in D6.9.

References

[1] 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.

[2] M. Goebel, T. Hassan, E. Oro, and G. Orsi, "Icdar 2013 table competition," in *2013 12th International Conference on Document Analysis and Recognition*, Aug 2013, pp. 1449–1453.

[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. Diem, F. Kleber, and R. Sablatnig, "Document Analysis Applied to Fragments: Feature Set for the Reconstruction of Torn Documents," in *Proceedings of the Inter-*

national Workshop on Document Analysis Systems (DAS), D. Doermann, V. Govindaraju, D. Lopresti, and P. Natarajan, Eds., Boston, USA, June 2010, pp. 393–400.

[5] 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.

[6] 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.