UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CENTRE FOR THE STUDY OF MANUSCRIPT CULTURES

DFG Deutsche Forschungsgemeinschaft

OpenX for Interdisciplinary Computational Manuscript Research

June 12-13, 2018

A Pre-Conference* Workshop Hosted by

The Centre for the Study of Manuscript Cultures (CSMC), University of Hamburg, Warburgstraße 26, 20354 Hamburg, Germany

and Funded by DFG via SFB 950 "Manuscript Cultures in Asia, Africa and Europe", Faculty of Humanities, Universität Hamburg **with Support by** DIVA, Université de Fribourg, Switzerland and Department of Informatics, Universität Hamburg

Joint Scientific Organization Committee:
H. Siegfried Stiehl (SFB 950, Scientific Service Project Z03 and Department of Informatics, Universität Hamburg) & Andreas Fischer (DIVA, Université de Fribourg)

Joint Technical Organization Committee:
Vinodh Rajan Sampath (Department of Informatics, Universität Hamburg) & Marcel Würsch (DIVA, Université de Fribourg)

Local Organization Committee:
Christina Kaminski (SFB 950) & Kirsten Tatam (Department of Informatics)

Tuesday 12th June 2018

08:30 **Arrival**

Tea, Coffee and Soft Drinks

09:00-09:15 **Welcome:** Oliver Hahn (SFB 950, Coordinator of Scientific Service Projects)

Introduction: H. Siegfried Stiehl & Andreas Fischer

09:15 **Morning Session: *Methodic Cores*** (Chair: Andreas Fischer)

09:15-10:00 **Basilis Gatos** (Computational Intelligence Laboratory, National Center for Scientific Research „Demokritos“, Athens, Greece)
Word Spotting Techniques for Historical Document Images

10:00-10:45 **Marçal Rossinyol** (Computer Vision Center, Universitat Autònoma, Barcelona, Spain)
Segmentation-Free Word Spotting

10:45-11:30 **Stefan Fiel** (Institute of Visual Computing & Human-Centered Technology, TU Wien, Vienna, Austria)
Writer Retrieval and Identification in Historical and Modern Manuscripts

11:30-12:15 **Lambert R. B. Schomaker** (Artificial Intelligence & Cognitive Engineering, Rijks Universiteit Groningen, The Netherlands)
One Method Fits Everything? New Developments in Deep and Regular Machine Learning for Historical Document Analysis?

12:15 **Burning Questions and Group Discussion**

12:30-14:00 **Lunch**

14:00 **Afternoon Session: *Use Cases in Humanities*** (Chair: Lambert R.B. Schomaker)

14:00-14:45 **Dominique Stutzmann** (Institut de Recherche et d'Histoire des Textes, Centre National de la Recherche Scientifique, Paris, France)
Writer Identification and Script Classification: Two Tasks for a Common Understanding of Cultural Heritage

14:45-15:30 **Daniel Stökl Ben Ezra** (École Pratique des Hautes Études (EPHE), Section des Sciences Historiques et Philologiques, Paris, France)
Ocropy and the Holistic Approach: Applied to the Automatic Transcription of Manuscripts with Classical Rabbinic Hebrew Texts

15:30-16:00 **Tea, Coffee and Soft Drinks**

16:00-16:45 **Peter Stokes** (*Keynote*, École Pratique des Hautes Études (EPHE), Section des Sciences Historiques et Philologiques, Paris, France)
On Digital and Computational Humanities for Manuscript Studies. Where Have we Been, Where are we Going?

16:45-17:30 **Vanessa Hannesschläger (**Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Vienna, Austria)
Legally Open: Copyright, Licensing and Data Privacy Issues in the Context

17:30 **Burning Questions and Group Discussion**

Wednesday 13th June 2018

09:00	**Arrival**
	Tea, Coffee and Soft Drinks
09:30	**Morning Session: *Bridging the (So-Called) Gap*** (Chair: H. Siegfried Stiehl)
09:30-10:15	**Marcel Würsch** (DIVA, Université de Fribourg, Switzerland) *DIVAServices? How WebServices Can Bridge the Gap between Computer Science and the Humanities*
10:15-11:00	**Joseph Chazalon** (Laboratoire Informatique, Image et Interaction (L3i), Université de La Rochelle, La Rochelle, France) *Building an Evaluation Framework Researchers Will (Want to) Use*
11:00-11:45	**Vinodh Rajan Sampath (**Department of Informatics, Universität Hamburg, Germany) *Interactive Exploration of Digitized Manuscripts: Introducing the iXMan_Lab*
11:45-12:30	**Lessons Learned and Next Steps** (Chairs: H. Siegfried Stiehl & Andreas Fischer)
	Farewell

***... and Immediately Following:**

3rd International Conference on Natural Sciences and Technology in Manuscript Analysis, CSMC, June 13-14, 2018 (for details, e.g. program and registration, please visit the CSMC website https://www.manuscript-cultures.uni-hamburg.de/natural_sciences_2018.html)

Abstracts

Basilis Gatos (Computational Intelligence Laboratory, National Center for Scientific Research „Demokritos“, Athens, Greece)

Word Spotting Techniques for Historical Document Images

Word spotting is an alternative solution to optical character recognition in order to provide access to historical document images that suffer from several problems such as typesetting imperfections, writing style variations, document degradations and low print quality. The goal of word spotting is to retrieve all instances of user queries in a set of document images. The user formulates a query and the system evaluates its similarity with the stored documents and returns a ranked list of word results that are similar to the query. Word spotting is usually based on matching between common representations of features (e.g. color, texture, geometric shape or textual features) without involving a recognition step in order to convert the documents to a machine-readable format. Depending on how the query input is specified by the user, we have the query-by-example (QBE) and the query-by-string (QBS) word spotting approaches. At the QBE approaches, the user selects an image of the word to be searched in the document collection, while at the QBS approaches, the user provides a text string as input to the system. Depending on whether training data are used offline either to learn character and word models or tune the parameters of the system, word spotting approaches can be categorized as learning-based or learning-free. Depending on the possible involved segmentation phase, word spotting approaches may be segmentation-based or segmentation-free. Segmentation-based approaches involve a segmentation step at line or word level during pre-processing while the segmentation-free approaches are directly applied to the entire document pages without any segmentation.

Word spotting applications for document indexing and retrieval include searching online in cultural heritage collections stored in libraries all over the world, word spotting in graphical documents such as maps, retrieval of cuneiform structures from ancient clay tablets, assisting human transcribers in identifying words in degraded documents (Giotis et al. 2017). Several challenges which are related to the nature of the original documents have to be addressed by the word spotting methods. Historical documents typically contain text written in a language, an alphabet and a style that maybe no longer in use. They may also suffer from degradations such as stained paper, faded ink or ink bleed through.

(Fig. 1).

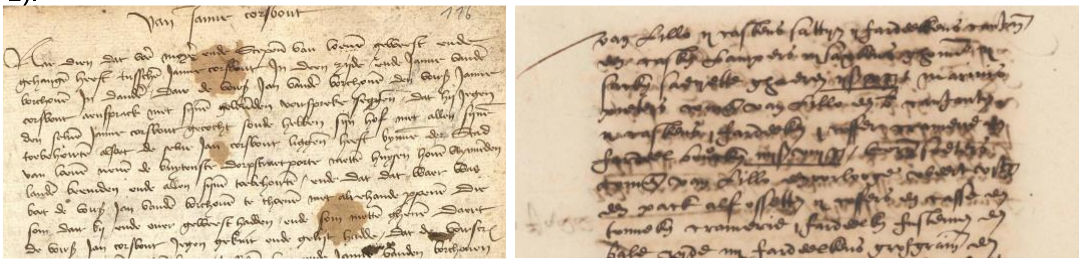


Fig.1. Examples of degradations in historical document images

Pre-processing steps of a word spotting workflow usually include document image binarization, enhancement, segmentation and normalization. Binarization refers to the conversion of the original grey-scale or color image to a black-and-white (binary) image. For the case of degraded historical document collections which usually suffer from non-uniform illumination, image contrast variation, bleeding-through or smear effects, efficient local thresholding techniques have been proposed (Khurshid et al. 2012; Moghaddam and Cheriet 2009). Image enhancement techniques are used mainly to improve the overall contrast between the script and the document background (Fink et al. 2014). Segmentation-based word spotting methods involve a segmentation pre-processing stage in order to segment the document pages at word or line level. Segmentation of historical documents is still an open research problem due to the significant challenges that are involved. These include variations in inter-line or inter-word gaps, overlapping and touching text parts, existence of accents, punctuation marks and decorative letters, local text skew and slant. The segmentation is usually followed by a normalization step in which several variabilities such text skew, slant and warping are removed.

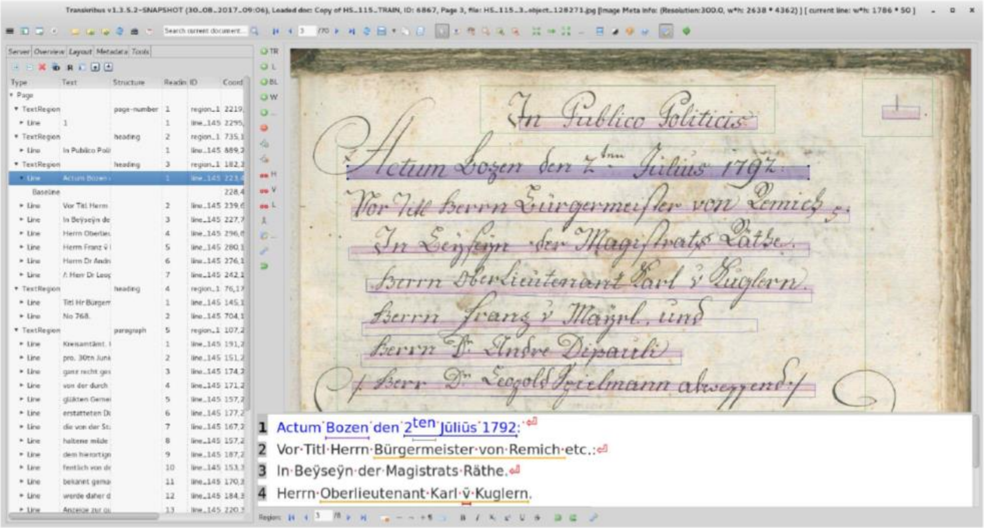
The appropriate selection of features has a great impact on the performance of a word spotting system. Global features can be extracted from the object of interest which can be either a word image or a document region as a whole. Examples of such features are the width, the height or the aspect ratio of the word image, the number of foreground pixels and the moments of background pixels. On the contrary, local features may be detected independently at different regions of the input image, which may be a text line, word or primitive word parts. The pixel densities, the position or the number of holes, valleys, dots and crosses at key points or regions are some examples of local features. Different feature types for word spotting applications are evaluated in (Rodríguez-Serrano and Perronnin 2009). After a set of features has been extracted, a suitable representation of their values has to be defined in order to allow efficient comparison between the query image and the documents at a specific level. Variable-length representations describe word images or text lines as a time series, usually using a window that slides over the image in the writing direction. In contrast, fixed-length representations extract a single feature vector of fixed size which characterizes the document region as a whole (Giotis et al. 2017).

The matching task is composed of the similarity computation between the feature representations of the query, which may be a feature vector, a graph, or a statistical model and the document image at word, line or page level. There are also Neural Network (NN) - based model approaches where a convolutional neural network accepts pairs of word images as inputs and returns a similarity score in the output. In that case, there is no image descriptor in the classical sense and images are processed and represented internally throughout the NN layer pipeline (Zhong et al. 2016).

Several methods have been proposed to improve the retrieved results of a word spotting system in terms of incorporating the information of the ranked lists obtained from user queries. This is done either by involving the user to select positive query instances in a supervised process (Ntzios et al. 2007), or in a purely unsupervised manner (Shekhar and Jawahar 2012). Some word spotting systems may result into several ranked lists which need to be combined into a final ranked list using a data fusion method (Rusiñol and Lladós 2014).

The ranked list of results obtained from a word spotting system for a number of different queries is finally used to evaluate its accuracy. Several publicly available datasets can be found for evaluation purposes. These include the IAM (Marti and Bunke 2002), the George Washington (Lavrenko et al. 2004) and the H-KWS competition (Pratikakis et al. 2016) datasets.

An example of a QBS word spotting engine for historical document images can be found at the Transkribus platform (https://transkribus.eu/). This is a comprehensive platform for the computer-aided recognition, transcription and searching of handwritten historical documents and is part of the EU-funded Recognition and Enrichment of Archival Documents (READ) project (https://read.transkribus.eu/) (Fig.2,3).

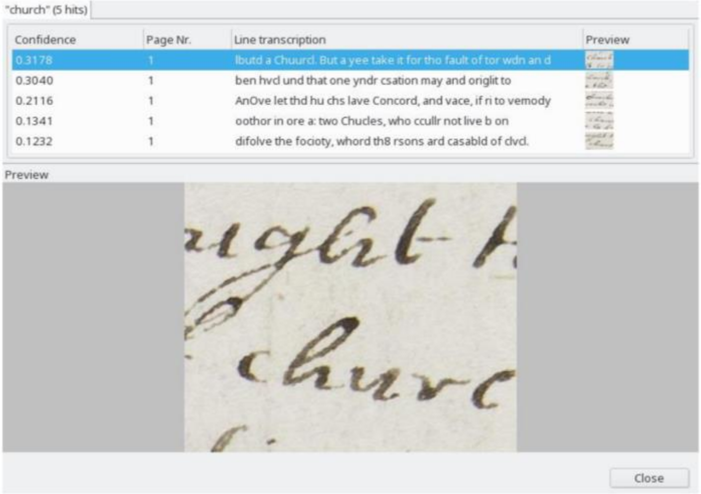


Transkribus v1.3.5.2-SNAPSHOT (30_08_2017_09:06), Loaded doc: Copy of HS_115_TRAIN, ID: 6867, Page 3, file: HS_115_3_object_128271.jpg [Image Meta info: (Resolution:300.0, w*h: 2638 * 4362)] [current line: w*h: 1786 * 50]

Type	Text	Structure	Readin	ID	Coord
Page					
TextRegion		page-number	1	region_1	2219,
Line	1		1	line_145	2295,
TextRegion		heading	2	region_1	735,1
Line	In Publico Poli		1	line_145	889,2
TextRegion		heading	3	region_1	182,1
Line	Actum Bozen		1	line_145	223,4
Baseline					228,4
Line	Vor Titl Herrn		2	line_145	239,6
Line	In Beÿseÿn de		3	line_145	227,7
Line	Herrn Oberlie		4	line_145	296,6
Line	Herrn Franz v		5	line_145	280,1
Line	Herrn Dr Andr		6	line_145	276,1
Line	/: Herr Dr Leop		7	line_145	242,1
TextRegion		heading	4	region_1	76,17
Line	Titl Hr Bürger		1	line_145	145,1
Line	No 768.		2	line_145	704,1
TextRegion		paragraph	5	region_1	107,2
Line	Kreisamtämt.		1	line_145	191,2
Line	pro. 30tn Juni		2	line_145	151,2
Line	ganz recht ge		3	line_145	174,2
Line	von der durch		4	line_145	171,2
Line	glükten Geme		5	line_145	157,2
Line	erstatteten D		6	line_145	177,2
Line	die von der St		7	line_145	167,2
Line	haltene milde		8	line_145	157,2
Line	dem hierortig		9	line_145	187,2
Line	fentlich von d		10	line_145	153,1
Line	bekannt gema		11	line_145	170,1
Line	werde daher d		12	line_145	184,1
Line	Anzeige zur		13	line_145	220,1

1 Actum Bozen den 2ten Jūliūs 1792:
2 Vor·Titl·Herrn·Bürgermeister·von·Remich·etc.:
3 In·Beÿseÿn·der·Magistrats·Räthe.
4 Herrn·Oberlieutenant·Karl·v̄·Kuglern.

Fig.2. The Transkribus desktop application



"church" (5 hits)

Confidence	Page Nr.	Line transcription	Preview
0.3178	1	lbutd a Chuurd. But a yee take it for tho fault of tor wdn an d	
0.3040	1	ben hvd und that one yndr csation may and origlit to	
0.2116	1	AnOve let thd hu chs lave Concord, and vace, if ri to vemody	
0.1341	1	oothor in ore a: two Chucles, who ccullr not live b on	
0.1232	1	difolve the focioty, whord th8 rsons ard casabld of dvd.	

Fig.3. The word spotting engine of the Transkribus platform. Results window of an exemplary search for the keyword “church”

Using the word spotting (keyword spotting - KWS) capabilities of Transkribus, there is no need to correctly transcribe documents before searching. Simply, a Handwritten Text Recognition (HTR) model is first employed to automatically produce an approximate transcript and then word searching is enabled. Even if the automatically generated transcript contains errors, KWS will reliably find words, phrases and even parts of words and regular expressions in the documents. At the example of Fig.3, the word “church” has been spotted correctly although the HTR results have lot of errors.

References

[1] A. Giotis, G. Sfikas, B. Gatos and C. Nikou (2017), A survey of document image word spotting techniques. Pattern Recognition, 68: 310-332.

[2] K. Khurshid, C. Faure, N. Vincent (2012). Word spotting in historical printed documents using shape and sequence comparisons. Pattern Recognition, 45 (7): 2598–2609.

[3] R.F. Moghaddam, M. Cheriet (2009). Application of multi-level classifiers and clustering for automatic word spotting in historical document images. 10th International Conference on Document Analysis and Recognition (ICDAR’09), pp. 511–515.

[4] G. Fink, L. Rothacker, R. Grzeszick (2014). Grouping historical postcards using query-by-example word spotting. 14th International Conference on Frontiers in Handwriting Recognition (ICFHR’14), pp. 470–475.

[5] J.A. Rodríguez-Serrano, F. Perronnin (2009). Handwritten word-spotting using hidden Markov models and universal vocabularies. Pattern Recognition, 42 (9): 2106–2116.

[6] Z. Zhong, W. Pan, L. Jin, H. Mouchère and C. Viard-Gaudin (2016). SpottingNet: Learning the Similarity of Word Images with Convolutional Neural Network for Word Spotting in Handwritten Historical Documents. 15th International Conference on Frontiers in Handwriting Recognition (ICFHR’16), pp. 295-300.

[7] K. Ntzios, B. Gatos, I. Pratikakis, T. Konidaris and S.J. Perantonis (2007). An Old Greek Handwritten OCR System based on an Efficient Segmentation-free Approach. International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, 9(2-4): 179-192.

[8] R. Shekhar and C. Jawahar (2012). Word image retrieval using bag of visual words. 10th IAPR International Workshop on Document Analysis Systems (DAS’12), pp. 297–301.

[9] M. Rusiñol and J. Lladós (2014). Boosting the handwritten word spotting experience by including the user in the loop. Pattern Recognition, 47 (3): 1063–1072.

[10] U.V. Marti and H. Bunke (2002). The IAM-database: an English sentence database for offline handwriting recognition. Int. J. Doc. Anal. Recognit. 5 (1): 39–46.

[11] V. Lavrenko, T.M. Rath, R. Manmatha (2004). Holistic word recognition for handwritten historical documents. 1st International Workshop on Document Image Analysis for Libraries, pp. 278–287.

[12] I. Pratikakis, K. Zagoris, B. Gatos, J. Puigcerver, A.H. Toselli and E. Vidal (2016). ICFHR2016 Handwritten Keyword Spotting Competition (H-KWS 2016). 15th International Conference on Frontiers in Handwriting Recognition (ICFHR’16), pp. 613-618.

Marçal Rossiñol (Computer Vision Center, Univ. Autònoma de Barcelona, Spain)

Segmentation-free Handwritten Keyword Spotting

Nowadays, in order to grant access to the contents of digital document collections, their texts are transcribed into electronic format so users can perform textual searches. When dealing with large collections, automatic transcription processes are used since a manual transcription is not a feasible solution.

In the context of digital collections of historical documents, handwriting recognition strategies are applied to achieve an automatic transcription since most of those documents are manuscripts. However, handwriting recognition often does not perform satisfactorily enough in the context of historical documents. Documents presenting severe degradations or using ancient glyphs might difficult the task of recognizing individual characters, and the lexicon definition and language modelling steps are not straightforwardly solved in such context. Keyword spotting has become a crucial tool to provide accessibility to historical collection's contents.

Keyword spotting can be defined as the pattern recognition task aimed at locating and retrieving a particular keyword within a document image collection without explicitly transcribing the whole corpus. Its use is particularly interesting when applied in scenarios where Optical Character Recognition (OCR) performs poorly or cannot be used at all, such as in historical document collections, handwritten documents, etc.

Handwritten Keyword Spotting is a mature research problem. The term was introduced in the mid 90's by the seminal paper by Manmatha et al. [1] since then, many different keyword spotting approaches have been proposed through the years. Until quite recently, all the proposals followed the same processing pipeline:

- A layout analysis step aimed at segmenting text-lines and words individually;
- An extraction of robust visual features that represent the character shapes;
- An accurate word matching strategy that cluster similar words together.

However, such pipeline presents some important drawbacks. The end-to-end performance can be seriously affected by the errors introduced by the segmentation step. Such methods are also hardly scalable because we have to compute the distances to all the words in the corpus in the retrieval stage. And, finally, the user needs to crop an example of the word he searches. In order to address such challenges, the late trends in keyword spotting research is focused on:

- Segmentation-free methods,
- indexable word features,
- query-by-string methods

We will present our latest research focused on query-by-string and segmentation-free keyword spotting methods using bleeding edge deep learning methods that are able to compute embedding’s between text and image information and perform single-shot detection and recognition.

[1] R. Manmatha, C. Han, E.M. Riseman and W.B Croft, “Indexing Handwriting Using Word Matching” Int. Conf. on Digital libraries, 1996

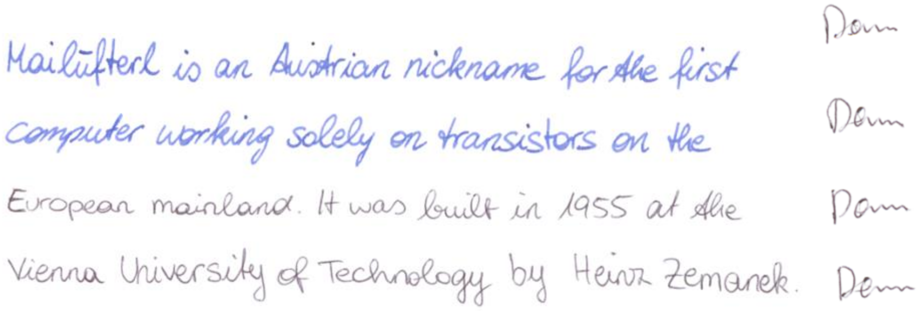
Stefan Fiel (Institute of Visual Computing & Human-Centered Technology, TU Wien, Vienna, Austria)

Writer Retrieval and Identification in Historical and Modern Manuscripts

In this paper a short overview about the evolution of methods for writer identification and writer retrieval is given. Mostly my work is references, but in the community, many people have used similar methods for this task. Results show, that methods for writer identification achieves good performance on scientific datasets. Two datasets are used to present some results.

1. Introduction

Writer retrieval is the task of retrieving document images from a dataset according to the similarity of handwriting. Experts can then analyze this ranking and thus new documents of the same writer can be found in an archive. This also allows drawing new connection between different manuscripts if they have written by the same writer. In modern context writer retrieval can be used for forensics to analyze ransom or threat letters. It can link different letters and thus the chance to find the author of the letters is increased. In contrast to this writer, identification is the task of determining the writer of a document. Thus, a dataset of writers has to be created beforehand and the system tells which of these writers have written a specific document. This can be used to identify the writer of an unknown document in case possible authors are known. The handwriting style of people depends on different parameters like which pen is used or outside influences such as distractions by something or someone. Figure 1 left shows a sample page of the CVL Database [1] where the writer changed the pen during writing. For humans the handwriting looks different at the first sight, but taking a detailed look at for example the word “the”, it can be seen that the same person wrote all four text lines. Figure 1 right shows another sample of the CVL Database. The text has four times the German word “dann” written in a column and a crop of this region is shown. The word has never been written exactly the same; small variations in different characters are occurring. Methods for writer identification and retrieval have to deal with variations like this when applied to real world samples. Another challenge, which is not covered by any scientific database so far, is that the handwriting changes with the age of the writer. Especially when these methods are applied to historic data, these variations must be investigated.



Mailüfterl is an Austrian nickname for the first
computer working solely on transistors on the
European mainland. It was built in 1955 at the
Vienna University of Technology by Heinz Zemanek.

Dann
Dann
Dann
Dann

Figure 1: Left: Sample image of the CVL dataset. The writer used two different pens, thus the handwriting looks different, Right: Again a sample of the CVL dataset, where the same writer has written the word “Dann” for times different

2. Methods for Writer identification

This chapter gives an overview over recent advances in the field of writer identification. It starts with features which are calculated directly, followed by methods using local features are presented and then also two methods using deep learning are described.

Features on characters

Bulacu et al. [2] presented in 2007 a method, which uses different features, which are extracted on the binarized version of the document image. They introduce also contour-hinge features, and grapheme emission. The angles of the writing are determined (shown in Fig.2) and they encode the roundness of the characters, which is a suitable feature for writer identification. For the grapheme emission, they extract a small patch of the characters and search for the most similar in a dataset and they count how often each patch in the dataset is used. On the IAM dataset an identification rate of 89% is reached.

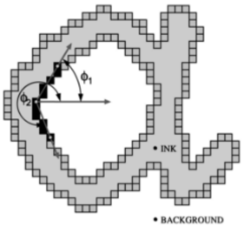


Figure 2: Contour hinge feature, which was introduced by Bulacu. Image Taken from [2]

Local Features

In Fiel and Sablatnig [3] local features, namely SIFT features, are used for writer identification and writer retrieval. SIFT features automatically detect so called interest points in the image and describe their neighborhood. Figure 3 left show some of these points on a handwriting sample. The size represents the neighborhood they are describing. The Points mainly lie on the end of the edges, at crossings or in the middle of circles and thus they are able to capture the characteristics of the respective handwriting. To generate one feature vector for each document image first SIFT features are calculated on a training set. These features are grouped into a given number of clusters. For each new document image, the SIFT features are calculated and for each feature the nearest cluster center is determined. As feature vector for the document image the occurrences of the cluster centers is taken.

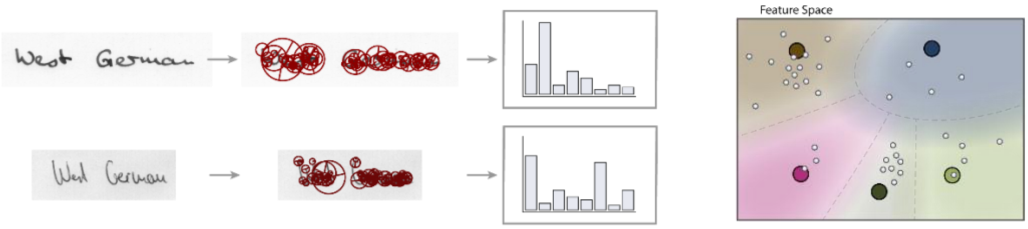


Figure 3 left: SIFT features are calculated on the document image and a occurrence histogram is generated which is then used for writer identification; right: Instead of the strict borders (dashed lines) of k-Means clustering probability functions are fitted.

In Fiel and Sablatnig [4] this method is further improved. Instead of grouping the features on the trainings set using k-Means, a Gaussian Mixture Model (GMM) is used. Figure 3 right shows the difference. The small white dots represent the location of the SIFT features in the feature space, whereas the large dots represent the cluster centers. Instead of taking the k-Means clustering (dashed-lines), which forms strict borders in the feature space, Gaussians are used. This allows to locate a SIFT feature more precisely in the feature space since the influence of all cluster centers to this particular feature can be calculated. For all features, the Fisher vector of the GMM can be calculated and is used as feature vector for the complete document image.

Deep Learning

The deep learning methods, which have arisen from digit recognition, have been proposed for various computer vision problems in the last years. These methods have found their way back to the field of document image analysis, e.g. handwritten text recognition, and recently also methods using deep learning are proposed for writer identification and retrieval for example by Fiel and Sablatnig [5] and Christlein et al. [6]. In Fiel and Sablatnig [5] image patches are extracted along the text lines, which have been size normalized and deslanted. The image patches extracted on the training set are then fed into a Convolutional Neural Network (CNN) with the task of classifying the specific writer. The CNN is able to capture the structure of the image patch and thus it learns the characteristics of the writer. For an identification task the patches of the document image are again fed to the CNN, but this time the activation functions of the second last layer are takes as feature vector for each patch. To generate a feature vector of the document image, these patch feature vectors have to be combined. In Fiel and Sablatnig [10] the naive approach of just taking the mean of the patch vectors is used.

In Christlein et al.[6] a different workflow is used, which is presented in Figure 4. To extract patches on the document image, the location of the SIFT keypoints, which has also been used in Section II-B, are taken. But also the description of the neighborhood is taken into account. For training the features are clustered into a predefined number of clusters. The CNN is now trained to classify these clusters instead of the writer. This has the advantage, first, that the writers of the training set has not to be known in advance and second, since the clusters form groups of patches with similar structure the CNN learns to identify these patches and thus give a descent description encoded in the feature vector. Again, the penultimate layer is takes as feature vector for the patches. For the combination of the feature vector of the patches VLAD encoding is used. It encodes first order statistics by aggregating the residuals of local descriptors to their corresponding nearest cluster center.

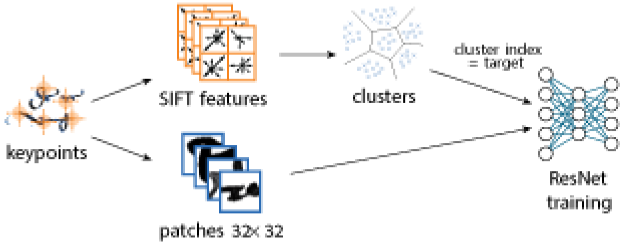


Figure 4 Workflow of the [6ö The patches are extracted on the location of the keypoints. The CNN is told to classify these patches according to a clustering of their descriptors.

3. Results

For the evaluation of writer identification and retrieval algorithms a competition is carried out at the International Conference on Document Analysis and Recognition. In 2013 the dataset consists of 1000 pages of 250 writers, who has contributed 2 pages in English and 2 pages in Greek. The results for the local feature method [4] is an identification rate of 94.5% and for the first deep learning method 88.5%. In 25.7% respectively 15.8% of the cases, the algorithm is able to find all other 3 pages of the same writer. These numbers are low because of the change of the alphabet in this dataset. Other methods achieve a performance of 61% on this dataset.

In 2017 this competition [7] was enlarged and used 3600 pages from 720 different writers. This time the document images were real world images from the Universitätsbibliothek Basel and were written from the 13th to the 20th centuries. They contain more text lines but also more noise, like stroke-through, underlined text and also some remarks, which may originate from a different writer. The text region of the document images are cropped out manually and a binarized version is available. Even though the dataset is quite large, the identification rate of [5] is 81.4 percent whereas the identification rate of the second deep learning approach is 88.6%. When all 4 other pages of a writer are searched, the algorithms have a success rate of 27.7 respectively 46.8%.

4. Conclusion

This paper presents the development of writer identification methods over the last decade. First, the feature were calculated on the character itself, later local descriptors, which describe the neighborhood of keypoints, are used. Currently, deep learning methods are used for the retrieval of similar images, which lead to a very good performance, but have the drawback that pre- and post-processing methods are needed. Currently the state of the art methods for writer identification have an identification rate of nearly 90%, which means that the datasets have to be increased dramatically, so that significant improvements are possible and the methods more generic and not limited to a specific dataset.

References

[1] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, “CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting,” in 2013 12th ICDAR 2013, pp. 560–564.

[2] M. Bulacu and L. Schomaker, “Text-Independent Writer Identification and Verification Using Textural and Allographic Features,” PAMI vol. 29, no. 4, pp. 701–717, apr 2007.

[3] S. Fiel and R. Sablatnig, “Writer Retrieval and Writer Identification Using Local Features,” in 2012 10th IAPR International Workshop on Document Analysis Systems (DAS), March 2012, pp. 145 – 149.

[4] S. Fiel and R. Sablatnig, “Writer Identification and Writer Retrieval Using the Fisher Vector on Visual Vocabularies,” in 2013 12th ICDAR, 2013, pp. 545–549.

[5] S. Fiel and R. Sablatnig, “Writer Identification and Retrieval using a Convolutional Neural Network,” in 16th CAIP 2015, 2015, pp. 26–37.

[6] V. Christlein, M. Gropp, S. Fiel, and A. Maier, “Unsupervised feature learning for writer identification and writer retrieval,” in 2017 14th IAPR ICDAR, vol. 01, Nov 2017, pp. 991–997.
[7] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, N. Stamatopoulos, and B. Gatos, “ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI),” in 2017 14th ICDAR, 2017, pp. 1377–1382.

Lambert R. B. Schomaker (Artificial Intelligence & Cognitive Engineering, Rijks Universiteit Groningen, The Netherlands)

One method fits everything? New developments in deep & regular machine learning for historical document analysis

Recent advances in deep learning by means of convolutional neural networks are very impressive in many application domains. Are these methods also suitable for the recognition of -hitherto unseen-handwritten documents in a rare script and language? What to do if the amount of training data is severely limited? What are the options for characterizing documents in terms of writer identity, general style or 'estimated date of production'? The presentation will give an overview of developments around the Monk system, i.e., the three projects: Himanis, ' Making Sense' and the Dead Sea Scrolls. Results indicate that both modern deep learning and regular pattern recognition need to live side by side peacefully in order to realize usable results, under conditions of sparse labelling. Whereas deep learning allows for new paradigms such as multi-modal task diversification and trainable image-to-image transforms, the advantage of explicit probabilistic modelling is located in the explainability of results.

Dominique Stutzmann (Institut de Recherche et d'Histoire des Textes, Centre National de la Recherche Scientifique, Paris, France)

Writer identification and script classification: two tasks for a common understanding of cultural heritage

This contribution addresses the divide between two tasks that have been separated as well in computer science (as evidenced by the different competitions, e.g. Historical WI and CLAMM) as in palaeography (levels 2 and 3 of "palaeography of expertise"). From a modelling perspective however, it is only one and the same question (distinguish what belongs together and what does not belong, which can be addressed by the same means) and a discretization of a body of evidence which can be seen as an historical continuum.

Daniel Stökl Ben Ezra (École Pratique des Hautes Études (EPHE), Section des Sciences Historiques et Philologiques, Paris, France)

Ocropy and the Holistic Approach :
Applied to the Automatic Transcription of Manuscripts with Classical Rabbinic Hebrew Texts

While considerable high quality textual and lexical data is openly available for many languages, such as Greek or Latin (even though the situation is improvable), classical Hebrew and Aramaic together with many other important European languages such as Armenian or Georgian are still groping in the dark.

Among the most important classical Hebrew texts are those redacted during the tannaitic Rabbinic period, i.e. around the 3rd century CE: The Mishnah, the Tosefta and so called Halakhic Midrashim. All these sources are juridical texts concerning Jewish life in Palestine, an Eastern province of the Roman Empire. Better known to the outside world is perhaps the Talmud, which is in fact a commentary to the Mishnah. The length of these texts is substantial, e.g. about 200k tokens for the Mishnah, about 300k tokens for the Tosefta, and they probably represent the most extensive sources from the pre-Christian Roman Empire that are still extant and not written in Greek or Latin. Their importance for our understanding of the development not only of classic rabbinic Judaism, but also of Roman provincial law and social history cannot be overstated.

Among the tannaitic sources mentioned above, the first two, Mishnah and Tosefta, are structured around topics, while the Halakhic Midrashim are commentaries to the Bible and therefore follow the order of their Biblical base texts (Exodus-Deuteronomy). The Tosefta is a text closely related to the Mishnah with the same structure and a complex intertextual relationship. In an oversimplified way one could compare it to that between two synoptic Gospels. While there are several low quality online open source texts, there is no high quality transcription openly available. An excellent linguistically annotated transcription of one manuscript of the Hebrew part of almost all texts can be accessed via the website of the Israel Academy of the Hebrew Language. While access to its resources is free, it is not open source.

Three years ago, Hayim Lapin from the University of Maryland and Daniel Stökl Ben Ezra from the EPHE/PSL in Paris have joined their relevant projects on these texts in the eRabbinica project to start closing this gap. They have secured funding from different sources for different subprojects. A pilot edition of 3 treatises of the Mishnah with transcription, automatic textual criticism, French and English translation and linguistic annotation, based on TEI/XML and eXist is planned to go online in June 2018 and will be presented at the conference. This extended abstract briefly introduces a selection of the machine learning algorithms applied in our project following a chronological sequence to perhaps elucidate others with similar projects.

1. A tailor made OCR

One of the most important manuscripts of the Mishnah is the Cambridge MS. Add 470.1 from 15th century Byzantium. In 1883, W.H. Lowe published an extremely precise transcription that represented faithfully not only the text of the manuscript but also changes in writing style using various fonts and special disposition of characters above the line for interlinear additions and at the end of lines or of paragraphs in the margins for marginal additions. Dots above letters indicate abbreviations and corrections.

At first, we envisaged training a commercial OCR of this 19th century transcription, yet the multiplicity of fonts and the use of the less common “Rashi” font did not give very good results. Furthermore, all the precious semantic information conveyed in the letter positions would be lost. Therefore, in 2016 Stökl developed a tailor made simple but very effective OCR engine. Its center consists of a K-clustering of 335 classes based on HOG-features of connected components. In a first step, horizontal projections were used to locate headers and footers subsequently excluded from further analysis. The next step was the creation of a huge database of all connected components on the main part of all pages. The vector for the euclidean distance K- clustering consisted of a concatenation of HOG features of 3 resized representations of each single connected component: 64x64 square and a flat 32x128 and a tall 128x32 rectangle with a cell size of 4x4 and 8x8, plus the height, width and the height-width proportion of each connected component. The 335 clusters were named manually.

Paragraph segmentation and recognition of marginal additions was done with vertical profiles. Row segmentation was based on horizontal profiles. All connected components could then be assigned to rows. A combination of the clustering result and the centroid position plus the top and bottom boundaries vis-à-vis the row base-line served to evaluate whether a letter was superposed or not. Tags served to add all layout and font information to the transcription of the letters. Subsequently, the automatic transcription was corrected manually. We estimate that the precision was probably higher than 99.5%.

2. Holistic Manuscript layout analysis for writing block detection

Originally, we had developed this system for automatically transcribing manuscripts. The system attained a precision of about 85%, which was not high enough. Transcription-glyph alignment based on synthetic “manuscriptization” of the transcription was more successful, but still not precise enough for production. Some of these steps were presented as a poster in the CSMC conference in February/March 2016. The main challenge consisted in letter segmentation in connected script. With the help of morphological transformation this worked quite well, but was completely manuscript and scribe dependent. Then, in a lecture in the e-philologie lecture series at PSL Université Paris, Marcus Liwicki mentioned ocropy. Jean-Baptiste Camps from the École Nationale de Chartes at PSL reported quite encouraging results on medieval Latin manuscripts. Despite the note of Thomas Breuel that ocropy does not suit for handwritten text recognition, the biLSTM of ocropy is in fact powerful at least for certain medieval manuscripts. More problematic is its layout analysis, which suits the needs of printed documents but not the small and larger irregularities of manuscripts. The solution was to do develop the binarization and column/writing block and line segmentation ourselves and to subsequently feed the results into ocropy.

The **binarization** employed is based on a sequence of well known morphological transformations (closing with a circle of a size depending on resolution and script size to calculate background; deducing background from image to create foreground; adjusting image intensity values of the foreground; Otsu binarization of the resulting image). It is simple but the results were good enough on the material.

With regard to layout analysis, Stökl's approach was to better exploit the regularity of our manuscripts. It seemed absurd to deal with pages of a manuscript one by one as if they were unrelated to the others. Exploiting this previous knowledge can improve existing algorithms, probably even in the age of convolutional neural network layout analysis. Even if lines can be slightly oblique or curved, or paragraphs can be oblique, or there are frequent marginal additions, the basic concept of these manuscripts of literary texts is regularity. Columns have a relative constant position, width and height. They can be interrupted by intermediary titles or empty space, but in principle they are quite regular.

Most frequently, documents are considered as two dimensional objects (even if pages are warped). What we would term **'holistic approach'**, however, takes into serious consideration the overlooked third dimension of manuscripts, the z-axis in addition to y and x. Instead of a horizontal or a vertical profile of single pages, the method consists of calculating a two dimensional z-profile, as if an X-ray was looking through the manuscript and then applying horizontal and vertical profiles. This provides precious information about the regularity of writing block disposition in a manuscript or printed book. While this idea may appear extremely simple, it has proven very efficacious in practice. Areas that are more frequently part of a writing block have higher z-profile values than those that are not. This also makes it possible to calculate the variability of writing block width and height and the distance to marginal additions. Manuscripts with a very regular layout have very sharp z-profile, while manuscripts with less regular layout have a more blurred z-profile. Especially marginal additions that are difficult to detect in a one-page-a-time approach, become discernible with the z-profile that distinguishes the normative basis from the addition.

Distinguishing between the z-profile of even and odd pages further sharpens the z-profile since many if not most manuscripts have a mirrored layout. Calculating the distance from the z-profile for each page can subsequently help to establish different z-profiles for different parts of the manuscript, i.e. for the material in the beginning and the end of the book, or for pages that commence a new chapter, pages with illustrations or tables etc.

3. Line segmentation with the heartbeat seamcarve

In our manuscripts, lines have a relatively constant distance but they can be empty, they can end early or start in the middle of the column, writing direction can slope be slightly curved. All these features pose problems to a horizontal profile approach, but also to a seam carving algorithm. The success of the seam carving algorithm depends to a large extent on the correctness of the detection of median lines. If, however, a line ends early and the subsequent line starts late, the simple median line approach will consider the second a continuation of the first which will result in the two lines being seamcarved as a single line.

At a joint workshop in Kaiserslautern of the Fribourg DivaDia team, the DFKI and the EPHE, Mathias Seuret, Marcus Liwicki and Stökl started to add an exploitation of the assumption of regularity to the seam carving algorithm, which has been published in the recent HIP@ICDAR. In the age of deep learning, this simple combination may perhaps seem less interesting from a computer science perspective, but the results are most useful for our daily work as philologists, and in the end, this is what counts in our real life at the École *pratique* des hautes études. With a Fourier transformation of the horizontal projection of each of n slices of a writing block, the procedure first calculates the median line distance. Wherever the line is too short or empty and therefore the horizontal profile misses a peak, the algorithm adds one or several artificial peaks according to the regular line distance with regard to the lines above and below. The algorithm is now implemented in the DIVAServices. Even without the seam carving, the line segmentation was already extremely efficacious and served in the pipeline for manuscript transcription and transcription-glyph alignment with ocropy.

4. Manuscript transcription and transcription-glyph alignment

Once a pipeline for the production of relatively clean manuscript-line-image and transcription established, models trained with Ocropy showed very useful results. A preliminary step is data augmentation with noisified data. We used the well known methods of salt and pepper as well as inclination of the manuscript line image in different dosages, angles and combinations to multiply input pairs by the factor ten.

The most challenging stage was the production of transcription text lines that correspond to the visible signs in the main text block. All marginal or interlinear additions had to be deleted. On the other hand, all deletions of the main text by simple strikethrough had to be kept. Numbering and paratext which uses regular Hebrew letters had to be kept. Letters functioning as simple line fillers without importance for the linguistic text, a frequent habitus in Hebrew manuscripts, had to be kept. Abbreviations had to remain unresolved. Ligatures had to be represented by special marks. Luckily our transcription markup distinguished between the various forms of addition and deletion and it was mainly a question of the order of transformation steps.

For the preparation of the most complex manuscript (Kaufmann), we used Microsoft Word with numerous styles to emulate XML tagging because XML editors like Oxygene are still difficult to manage with Right-to-Left scripts whose writing direction counters that of the tags. Hayim Lapin wrote a converter to XML/TEI in visual basic.

So far, we have applied our pipeline to the following Hebrew manuscripts:

- Mishnah: ms Kaufmann A50 in the Library of the Hungarian Academy of the Sciences. Written in Italian script from the 11th or 12th century. 256 folios.
- Mishnah: Cambridge University Library MS Add. 470.1 written in Byzantine script from the fifteenth century. 250 folios.
- Mishnah: Cod. Ebr. 95 of the Bayerische Staatsbibliothek in Munich (on the part of the Mishnah only because the resolution is very low for the tiny script of the Talmud itself). The manuscript was written in 1342 probably in France. 576 folios. Semi-manual manuscript layout segmentation of the complex Talmudic layout was very kindly provided with by the Larex team around Christian Reul.
- Tosefta: Austrian National Library at Vienna, Cod. Hebr. 20. Written around the 14th century in square Sephardic script. 327 folios.
- Tosefta: London British Library Add. 27296 with 73 folios written in 15th century Sephardic script.

For most manuscripts, we could achieve a CER <5%, sometimes <3%. We tried different sizes for the hidden layer, different learning rates, different sizes of training data. We also mixed training data from different manuscripts with encouraging results. 5 columns (171 lines, 200 neurons) achieved a CER <10% for the Vienna manuscript, while 19 columns (645 lines) sufficed for 2.1% CER (43k iterations). Due to limited manpower and calculation power (ocropy runs on CPUs only), we did not apply all tests on all materials. The main aim was not to improve the LSTM but to arrive at exploitable results. Of course, in a manuscript of 1M characters, 3% CER still means 30k errors to spot and to correct. However, where a vulgate text is available or one manuscript of a text is already transcribed, we can automatically align both versions with the recent Shmidman-Koppel-Porat algorithm. While we have made use of the transcription-glyph alignment of the OFTA algorithm, the neural network demands less human work.

In June, we will begin two new projects called *Sofer Mahir* (=*tachygraph* in Hebrew) and *Tikkoun Sofrim* (“scribal error correction”) with Dicta and with Haifa University in collaboration with the Hebrew manuscript portal Ktiv at the National Library in Israel to create a pipeline and produce further open source manuscript transcriptions of tannaitic texts that via IIIF link coordinates for words, and where possible glyphs to the manuscript images. With Haifa University we will work on correction of automatic transcription with crowdsourcing and gamification. In collaboration with Dicta, the texts will be automatically analyzed linguistically. We hope to be able to integrate the linguistic analysis directly into the transcription pipeline to further reduce error ratio. In the LAKME project, we have already annotated 25k tokens lexically and morphologically and created the corresponding lexicon in French, English and German in order to apply a neural network architecture developed by Dicta on all of our transcriptions. The resulting text will be presented according to the CTS system developed at the Chair of Digital Humanities in Leipzig that permits the selection of whole texts or parts thereof (chapters, “verses” or words) via a canonical URL system. All this will be of more general applicability in the Scripta-PSL project dealing with most written types of artefacts from most historical cultures, where we are looking for further collaborations.

Peter Stokes (École Pratique des Hautes Études (EPHE), Section des Sciences Historiques et Philologiques, Paris, France)

On Digital and Computational Humanities for Manuscript Studies. Where Have we Been, Where are we Going?

In the last decade or so, the application of digital methods to historical writing has grown enormously and shown very significant advances in many areas. Technological developments have been applied to the study of manuscripts for centuries, and debate has been ongoing for many years about quantitative methods and the nature of palaeography as an ‘art’ or a ‘science’, but recent years have nevertheless seen an enormous and rapid increase in this work. This is demonstrated among other things by increasing institutional recognition such as the attention to historical materials at conferences such as ICFHR and ICDAR, by centres like the CSMC, and by the recent creation of a chair in digital and computational humanities applied to the study of historical writing in France. The objective of this paper is therefore to review these developments, focussing not on technical advances *per se* but rather on the ‘view from the Humanities’, making comparison with previous reviews such as the two at Dagstuhl in 2014 and 2016 and a previous lecture delivered in Hamburg in 2016, to suggest some points around where we have been and where we might be going.

‘Digital’ and ‘Computational’ Palaeography. In general, it still seems the case that one can divide approaches to manuscripts into two groups. One includes more ‘computational’ or ‘statistical’ methods which generally rely on a greater degree of automation, less direct human intervention, and based more on latest developments in machine vision, image processing, pattern recognition and so on. The other more ‘symbolic’ approach focusses on structured descriptions generated more or less directly by domain experts or ‘in-betweeners’ for the purpose of knowledge creation through experimentation and exploration, as well as the communication of evidence to support the resulting argument.

‘Computational’ or ‘statistical’ approaches. Significant work in the last decade or more has been done on topics such as automatic analysis of handwriting for dating, localising and (especially) writer identification, script classification, and layout analysis. Recently, substantial progress has been made in fields such as line detection, HTR, wordspotting in handwritten documents, automatic layout analysis, and so on. Success has also been achieved in aligning images of text to pre-existing transcripts, as well as the identification of fragments of manuscripts from the same original document, and the identification of specimens of script likely to have been written by the same individual. More specifically aligned to palaeographical research is work on the characterisation or dating of script. All this is extremely promising and could transform manuscript studies, provided that the results can be trusted by those who will use them. Another important development is the degree to which these methods and techniques are becoming freely available for use by other projects, through free or Open Source software but also through web APIs. This is potentially a significant boon to palaeographers with some understanding of digital methods but without the resources or expertise to implement their own code. It may also go some way towards addressing the need for benchmarking and standard algorithms, a need which is also being addressed directly through benchmarking datasets that are becoming increasingly available for historical material.

‘Digital’ or ‘symbolic’ approaches. Partly in direct response to challenges of ‘algorithmic accountability’ in computational methods, the ‘symbolic’ approach relies less on statistics and computation and more on representing palaeographical knowledge in transparent but tractable ways. This approach emphasises data representation, interface design and UI/UX, visualisation, and so on, rather than the more ‘computational’ approaches listed above. This emphasis on discovery, analysis and communication relates directly to larger questions in Digital Humanities and beyond about how one represents expert knowledge in systems that are tractable to the computer, connecting to areas and technologies such as ontologies, formal modelling, Linked Open Data and the Semantic Web. Perhaps the most important work here at present for Humanities scholars is IIIF which is transforming the way in which people are working with manuscripts today. This provides (among other things) stable protocols for addressing and manipulating images of manuscripts and other cultural heritage over the Web, in a system which is becoming increasingly widely used by libraries, archives and other cultural heritage institutions. This means that we are now becoming able to refer unambiguously to images of manuscript pages and to regions in those images, and to access the images directly from many different repositories. For the Humanities researcher this means access to material and – really for the first time – the ability to easily compare images from different institutions in the same software. It also responds very directly to the need for ready and open access to data which has been noted in the Dagstuhl events and elsewhere.

Some Continuing or Future Directions?

Without claiming completeness or indeed originality, the following issues are relevant to the Humanities and seem are likely to become increasingly important in the near future

Algorithmic Accountability. In terms of future developments, an important question that has often been raised with regards to computational methods and has long been discussed in Digital Humanities is the need to be able to interpret and understand algorithms and their approaches. This problem of the ‘black box’ and of inherent bias in algorithmic approaches has become increasingly prominent in recent years and is recognised also in computer science, particularly under the rubric of ‘algorithmic accountability’. The ACM US Public Policy Council, for instance, published a ‘Statement on Algorithmic Transparency and Accountability’ (2017) which notes the ‘growing evidence that some algorithms and analytics can be opaque, making it impossible to determine when their outputs may be biased or erroneous’, noting technical, economic, and/or social reasons why this is the case. Although work on historical documents does not have the societal implications of the cases discussed by the ACM, nevertheless machine-generated features and highly computational methods are often not meaningful to humans and particularly to those in the Humanities. Recent work is seeking increasingly to change this.

Combining ‘digital’ and ‘computational’. As mentioned above, computational approaches are extremely promising particularly for large-scale questions, and for potentially allowing Humanities scholars to automate the routine parts of their work, freeing them to focus on the questions that interest them. These methods do have limitations, like any other method, and are only applicable to certain types of question. In contrast, the more symbolic approaches have demonstrated their value on a smaller scale with a more ‘close’ analysis, but they are laborious and are limited in different ways. What therefore seems very promising but is much less attested in practice is a combination of the two, simultaneously applying the ‘close’ and ‘distant’ approaches, using each to add to the other to produce entirely new results. Such work has been discussed at workshops and conferences but has been done relatively little in practice.

Multigraphism. Researchers in both palaeography and computer science are now recognising the difficulties of multigraphism, namely cases where individuals or cultures simultaneously use entirely different scripts, alphabets, or even writing systems. Most computational methods have been developed at least theoretically independently of any given script or writing system, and examples in practice include the same software successfully applied to different writing-systems, but generally to only one script at a time. However, in many – perhaps almost all – cultures, people used (and use) different scripts or writing systems together, and this raises challenges to ‘digital’, ‘computational’ and ‘traditional’ palaeography, the significance of which is demonstrated (for instance) by a competition on the subject that will be run at ICFHR 2018.

Selected References

The following list makes no claim to completeness but instead seeks to list a small selection of some of the more important publications from the Digital Humanities or Humanities.

[1] Andrist, Patrick, Paul Canart and Marilena Maniaci (2013). *La syntaxe du codex: Essai de codicologie structurale*. Turnout: Brepols.

[2] Canart, Paul (2006). 'La Paléographie est-elle un art ou une science?' *Scriptorium* 60: 159–85.
[3] Ciula, Arianna (2005). 'Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis'. *Digital Medievalist* 1 (2005). Available at http://www.digitalmedievalist.org/journal/1.1/ciula/
[4] Costamagna, Giorgio *et al*. (1995 and 1996). 'Commentare Bischoff'. *Scrittura e Civiltà* 19, pp. 325–48, and 20, pp. 401–7.
[5] Davis, Tom (2007). 'The Practice of Handwriting Identification'. *The Library*, 7th series, 8: 251–76.
[6] Hassner, Tal, Malte Rehbein, Peter Stokes and Lior Wolf, eds. (2013). 'Computation and Palaeography: Potentials and Limits'. *Dagstuhl Manifestos* 2: 14–35. doi:10.4230/DagMan.2.1.14
[7] Hassner, T., R. Sablatnig, D. Stutzmann and S. Tarte (2015). 'Digital Palaeography: New Machines and Old Texts'. *Dagstuhl Reports* 4:7 (2014): 127–8. doi:10.4230/DagRep.4.7.112
[8] Kestemont, Mike, Vincent Christlein and Dominique Stutzmann (2017). 'Artificial Palaeography: Computational Approaches to Identifying Script Types in Medieval Manuscripts'. *Speculum: A Journal of Medieval Studies* 92: S86–S109.
[9] Sculley, D. and Bradley M. Pasanek (2008). 'Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities'. *Literary and Linguistic Computing* 23: 409–24.
[10] Stokes, Peter A. (2009). 'Computer-Aided Palaeography: Present and Future'. In *Kodikologie und Paläographie im Digitalen Zeitalter — Codicology and Palaeography in the Digital Age*, ed. by M. Rehbein, P. Sahle and T. Schaßan. Norderstedt: Books on Demand, pp. 313–42. Available at urn:nbn:de:hbz:38-29782.

Vanessa Hannesschläger (Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Vienna, Austria)

Legally Open: Copyright, Licensing and Data Privacy Issues

In the field of digital research, methods and approaches of open science are gaining momentum. A vital precondition of applying these methods is knowledge about various aspects of the legal landscape, which this paper aims to address. Specifically, the topics of copyright in an international context, possibilities and pitfalls of open licensing, and legal restrictions brought about by the EU's new data privacy legislation will be discussed.

Copyright

The legal frameworks we are embedded in define if, how, and how long texts, material, meta-/data, and software can be made (and kept) available. The most relevant area of legislation that digital research, and especially manuscript research, is affected by is copyright. In this context, researchers are always in a Janus-faced position: As creators of content on the one hand, and (re-)users of content on the other. It has even been argued that, due to "the requirement for researchers to make their publicly funded work available to the public"[1], "copyright is an unsuitable legal structure for scientific works. Scientific norms guide scientists to reproduce and build on others' research, and default copyright law by its very purpose runs counter to these goals."[2] Still, copyright law is a reality contemporary research communities have to face. Especially for researchers working with computational methods, the "increasingly rapid development of new media continuously leads to new and unanticipated ways of distributing copyrighted works"[3] - which affects researchers both as creators and as users. Therefore, the first topic to be addressed will be both sides of the copyright coin that researchers need to consider when opening up their work.

A second topic will be the main principles of national and international copyright legislation(s). In a European context, employment of the term "copyright" itself is already problematic, as it refers to a concept of the Anglo-Saxon legal tradition: *copy* right primarily aims at regulating the right to replicate and reproduce. However, in most European countries the Germanic legal tradition, which puts a stronger focus on the persona of the creator (" *Urheber* recht") has shaped "copyright" legislations.

Within (most of) Europe, it is therefore more accurate to speak of “intellectual property” (IP) rather than “copyright” law. However: “As IP law in the European Union is merely harmonized and not unified, the exact scope of copyright and similar rights may differ between Member States (e.g. some Member States recognize an exclusive right for ‘scientific and critical editions’, while others don’t).”[4] This paper will aim to raise awareness of this fact, address the most crucial differences between national legislations, and point out their most vital (and most likely) consequences for digital manuscript researchers.

Open licensing

Due to territorial limitations of copyright, the digital space that transcends national borders calls for new legal arrangements that are able to protect the researchers’ rights on the one hand and ensure the reusability of their work on the other. Open licensing models enable long term preservation as well as international research on data collected in local research projects, thus greatly supporting emerging open approaches in manuscript research. However, scholars often lack an overview of the various possibilities to license their findings. The most established model which has gained great popularity for creative content and is increasingly also applied to research data is Creative Commons[5] licensing (CC). In spite of the fact that CC has become a de facto standard for licensing research data, scholars are often unaware of the details of the different CC modules and their consequences; choosing appropriate licenses for software is an even more complex task. The second focus of this paper will therefore be available options of ready-made open licenses and their benefits, potential pitfalls of open licensing and license selection (such as license compatibility issues, copyright preconditions, and other legal commitments such as work contracts), and license selection tools that allow to avoid them.

The Public License Selector

Creative Commons offers a basic license selection tool which is helpful for researchers who are already sure that a) their content is licenseable under Creative Commons and b) they have made a conscious decision to use Creative Commons. However, in some cases, Creative Commons licenses might not be the best choice, for example in the case of code. A very nifty tool that helps select appropriate open licenses for both data and(/or) code is the Public License Selector[6] developed by the European research infrastructure CLARIN-ERIC[7]. Users start with a total selection of 22 open, publicly available ready-made licenses and have to answer a sequence of questions. Each answer narrows down the licenses compatible with the respective preconditions, leaving the user with a final choice of open licenses suited to their specific situation (as well as further information about the individual qualities of all available licenses) at the end of the process. During the presentation, the Public License Selector will be demonstrated.

As a third main topic, this paper will address the EU's General Data Protection Regulation[8] (GDPR), which comes into effect on 25 May 2018. As the GDPR is a *regulation*, it will take legal effect in all EU member states immediately on the day of implementation (in contrast to a mere *directive* , which has to be ported to national legislations before becoming the law). The GDPR will replace the EU’s Personal Data Directive[9] (1995). Although the GDPR does not differ from the Personal Data Directive in terms of fundamental concepts, it does establish a few new requirements, as well as tangible punishments (penalties) in case of infringement. While the main aim of the GDPR is to protect citizens and individuals from abuse of their personal information by international corporations, it affects everyone working in digital space (despite several “research exceptions” such as archiving in the public interest). Hence, this paper will outline the main concepts of the GDPR and explain the most vital points to be considered in the context of digital manuscript research. In addition, the GDPR’s encouragement of bottom-up standardisation (e.g. by developing codes of conduct or data security certificates) will be briefly explained, as this could motivate the development and formalization of de facto community standards and thus create new opportunities for the digital manuscript research community.

By covering these three crucial areas of the legal landscape, this paper will offer a basic toolkit for computational manuscript researchers to conduct their projects as openly as legally possible.

[1] Stodden, Victoria. “The legal framework for reproducible scientific research: Licensing and copyright.” IEEE Computing in Science and Engineering 11/1 (2009): 35–40, 40. https://web.stanford.edu/~vcs/papers/Legal-STODDEN2009.pdf

[2] Stodden 2009, 35.

[3] Darling, Kate. “Contracting About the Future: Copyright and New Media.” Northwestern Journal of Technology and Intellectual Property 10/7 (2012): 485–530, 485. http://scholarlycommons.law.northwestern.edu/njtip/vol10/iss7/3

[4] Kamocki, Paweł, Pavel Straňák, and Michal Sedlák. “The Public License Selector: Making Open Licensing Easier.” *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia* , edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, et al., 2533–2538; 2534. Paris: European Language Resources Association (ELRA), 2016. http://www.lrec-conf.org/proceedings/lrec2016/pdf/880_Paper.pdf

[5] Creative Commons: https://creativecommons.org/

[6] Available at https://ufal.github.io/public-license-selector/ , and described in detail in Kamocki et al. 2016.

[7] Common Language Resources and Technology Infrastructure - European Research Infrastructure Consortium. https://www.clarin.eu/

[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[9] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A31995L0046

Marcel Würsch (DIVA, Université de Fribourg, Switzerland)

DIVAServices? How WebServices Can Bridge the Gap between Computer Science and the Humanities

Creating and executing workflows for historical document imaging and processing is not a simple task. One has to find the suitable Document Image Analysis (DIA) methods, typically fine-tune the parameters of such methods, and then execute them on large datasets. With DIVAServices [1] Würsch *et al.* have already introduced a solution for one part of this problem, by introducing a framework for providing access to DIA methods as RESTful Web Services. This solution helps users to find methods suitable for their tasks. So far it is not possible though to plan and execute workflows on large datasets.
In recent years various Workflow Management Systems (WMS) have been introduced: Pegasus [2], a workflow solution for scientific experiments with a focus on exploiting distributed computing infrastructures, or Taverna [3], a domain-independent WMS that is mainly used in the life sciences. None of these tools found relevant adaptation in the DIA community. We believe that this is due to the special nature of the domain. Most of the WMS are designed to execute workflows with zero interaction. In DIA, however, it is often the aim to keep the *human in the loop*, meaning that the user gets the possibility to interact with the workflow as it is running.
For historical documents, in particular, fully automatic systems are prone to errors for DIA tasks such as Optical Character Recognition (OCR), handwriting recognition, writer identification, or manuscript dating. Considering the sheer amount of different historical scripts and languages, the goal is instead to provide human experts with interactive DIA tools that support them in their work. Examples include the CATTI system for computer-assisted transcription of historical documents [4], Aletheia for annotating historical prints [5], PhaseGT for binarization of historical manuscripts [6], and GraphManuscribble for intuitive interaction with digital facsimiles [7], to name just a few. Over time, annotations provided by humans can be used to train methods based on machine learning, which in turn can improve the suggestions made by DIA tools, thus closing the loop between human experts and

the semi-automatic systems. In this abstract, we introduce DIVA-DIP,[1] a novel WMS that puts the user at the center. The application allows to design and execute workflows. Furthermore, the results of individual steps are available to the user, to perform investigations, make adaptations and re-run only certain steps of the workflow. In comparison to existing WMS, processes on DIVA-DIP can have an explicit state where they wait for user input, thus allowing it to keep the user engaged in the workflow. DIVA-DIP is fully compatible with DIVAServices, *i.e.* it can be used to create complex workflows based on methods provided by DIVAServices.
Figure 1 shows an overview of the User Interface. In the center the user can see the image that he or she is currently working on, and on the right, the user can see all the different steps of the workflow and can execute each of them. In this case we aim at comparing two different binarization methods.

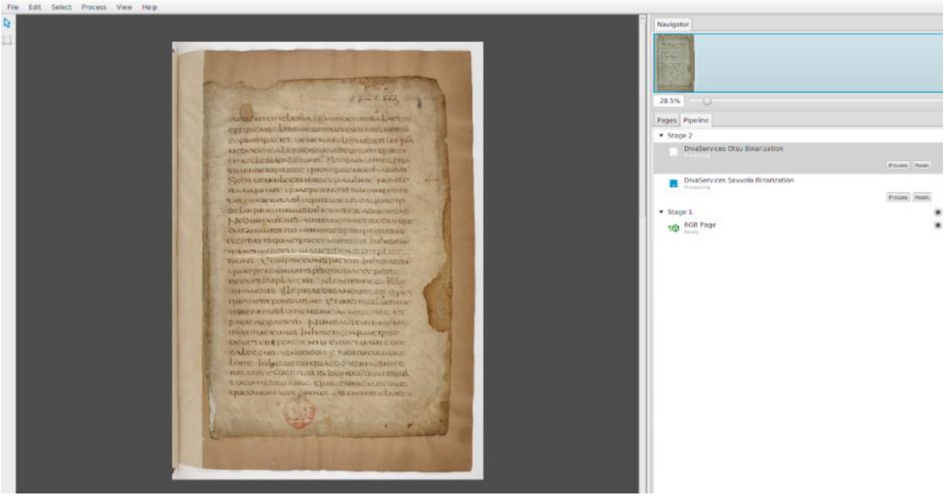

File Edit Select Process View Help
Navigator
28.5%
Pages Pipeline
Stage 2
DivaServices Otsu Binarization
Process Reset
DivaServices Sauvola Binarization
Process Reset
Stage 1
RGB Page

Figure 1 The User Interface of DIVA-DIP. The focus is on providing much space to the actual document. All workflow related information is available on the right.

Once a computation is performed, the user gets the possibility to see the results of this computation. This is visualized in Figure 2. When a computation is successfully executed, the icon of the method turns green, and using the radio button, the user gets the possibility to look at this specific result.

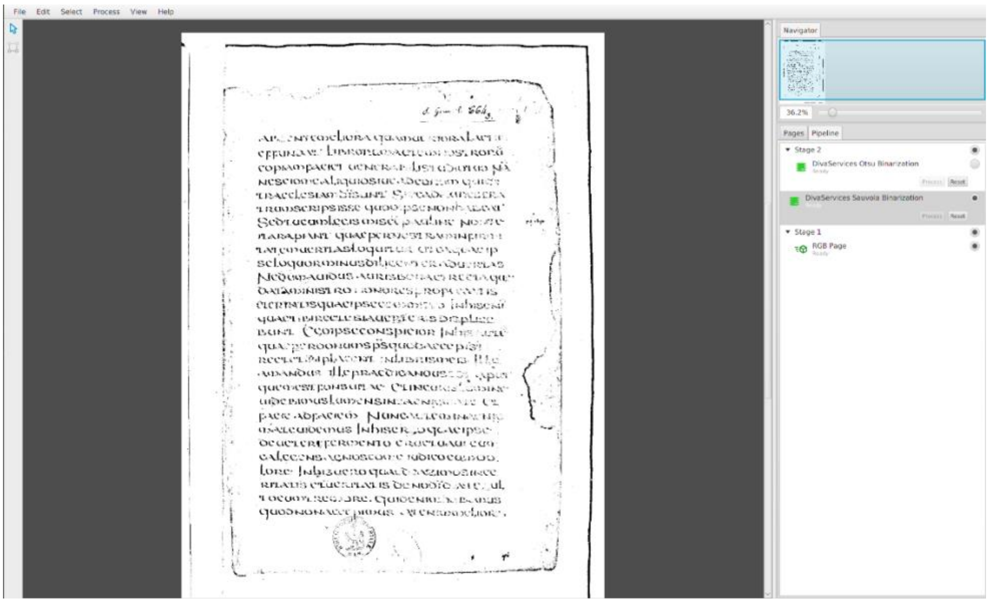
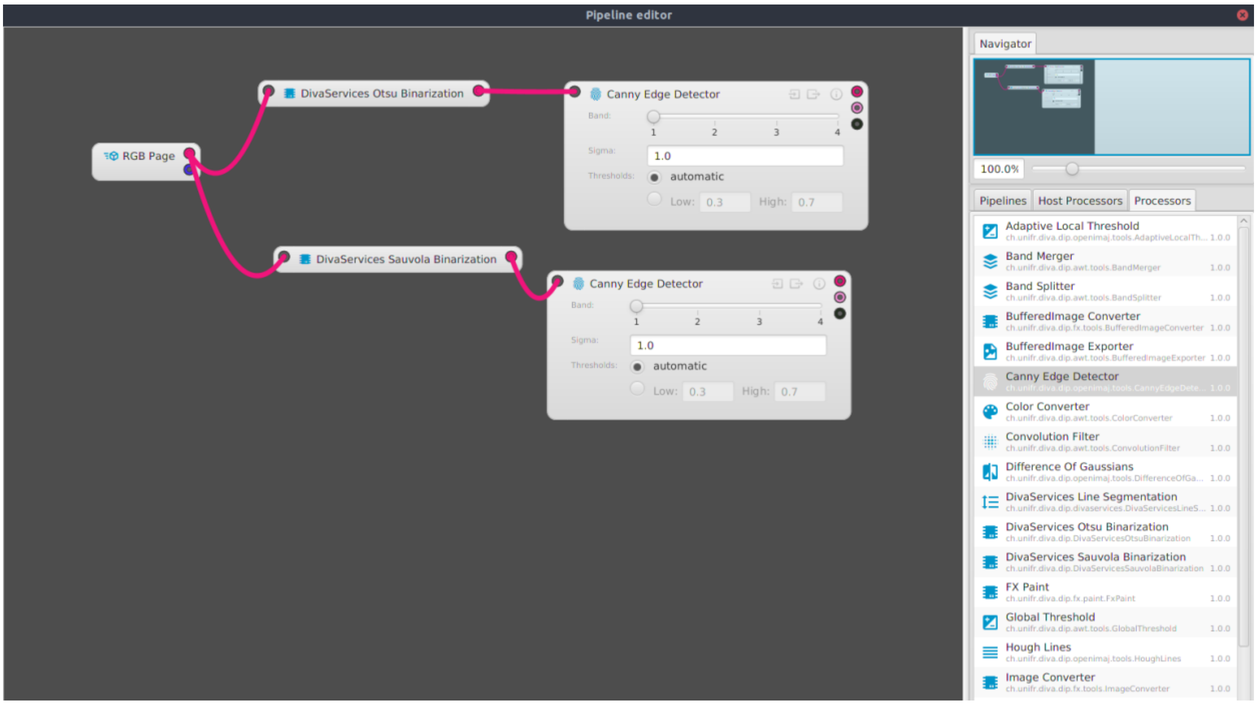

File Edit Select Process View Help
Navigator
36.2%
Pages Pipeline
Stage 2
DivaServices Otsu Binarization
Process Reset
DivaServices Sauvola Binarization
Process Reset
Stage 1
RGB Page

Figure 2 Visualization of a computed result. The user can switch between the various results using the radio buttons on the right side.

The proposed WMS follows the pipeline metaphor for connecting individual methods. In Figure 3 we show the workflow designer. In there, the user can see all the different methods (called *processors*)

[1] DIP stands for **Document Image Processor.**

and can drag them onto the work screen. If one is dragged on to the workspace, each input and output type is color-coded to provide a simple overview what can be connected together. If a method takes additional arguments, they can be provided on the method but also changed later when executing the workflow.



Adaptive Local Threshold	ch.unifr.diva.dip.openimaj.tools.AdaptiveLocalTh...	1.0.0
Band Merger	ch.unifr.diva.dip.awt.tools.BandMerger	1.0.0
Band Splitter	ch.unifr.diva.dip.awt.tools.BandSplitter	1.0.0
BufferedImage Converter	ch.unifr.diva.dip.fx.tools.BufferedImageConverter	1.0.0
BufferedImage Exporter	ch.unifr.diva.dip.awt.tools.BufferedImageExporter	1.0.0
Canny Edge Detector	ch.unifr.diva.dip.openimaj.tools.CannyEdgeDete...	1.0.0
Color Converter	ch.unifr.diva.dip.awt.tools.ColorConverter	1.0.0
Convolution Filter	ch.unifr.diva.dip.awt.tools.ConvolutionFilter	1.0.0
Difference Of Gaussians	ch.unifr.diva.dip.openimaj.tools.DifferenceOfGa...	1.0.0
DivaServices Line Segmentation	ch.unifr.diva.dip.divaservices.DivaServicesLineS...	1.0.0
DivaServices Otsu Binarization	ch.unifr.diva.dip.DivaServicesOtsuBinarization	1.0.0
DivaServices Sauvola Binarization	ch.unifr.diva.dip.DivaServicesSauvolaBinarization	1.0.0
FX Paint	ch.unifr.diva.dip.fx.paint.FxPaint	1.0.0
Global Threshold	ch.unifr.diva.dip.awt.tools.GlobalThreshold	1.0.0
Hough Lines	ch.unifr.diva.dip.openimaj.tools.HoughLines	1.0.0
Image Converter	ch.unifr.diva.dip.fx.tools.ImageConverter	1.0.0

Figure 3 Workflow designer of DIVA-DIP. The various inputs and outputs are color coded for simplicity. Users have the ability to change available parameters on each method.

DIVA-DIP is Open Source, released under LGPLv2.1, making it possible for everyone to add more processors into this application, increasing its usefulness. We started to include methods provided via DIVAServices such that users can easily create workflows based on a growing repository of DIA tools. We hope that over time, DIVA-DIP is able to evolve into a convenient front-end for DIVAServices, exposing the DIA tools not only to computer scientists but also to experts from the humanities, who can profit from the methods for creating dedicated workflows.

References

[1] M. Würsch, R. Ingold, and M. Liwicki, “SDK Reinvented: Document Image Analysis Methods as RESTful Web Services,” in *12th IAPR Workshop on Document Analysis Systems*, 2016, pp. 90–95.

[2] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira Da Silva, M. Livny, and K. Wenger, “Pegasus, a workflow management system for science automation,” *Futur. Gener. Comput. Syst.*, vol. 46, pp. 17–35, 2015.

[3] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M. P. Balcazar Vargas, S. Sufi, and C. Goble, “The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud,” *Nucleic Acids Res.*, vol. 41, no. Web Server issue, 2013.

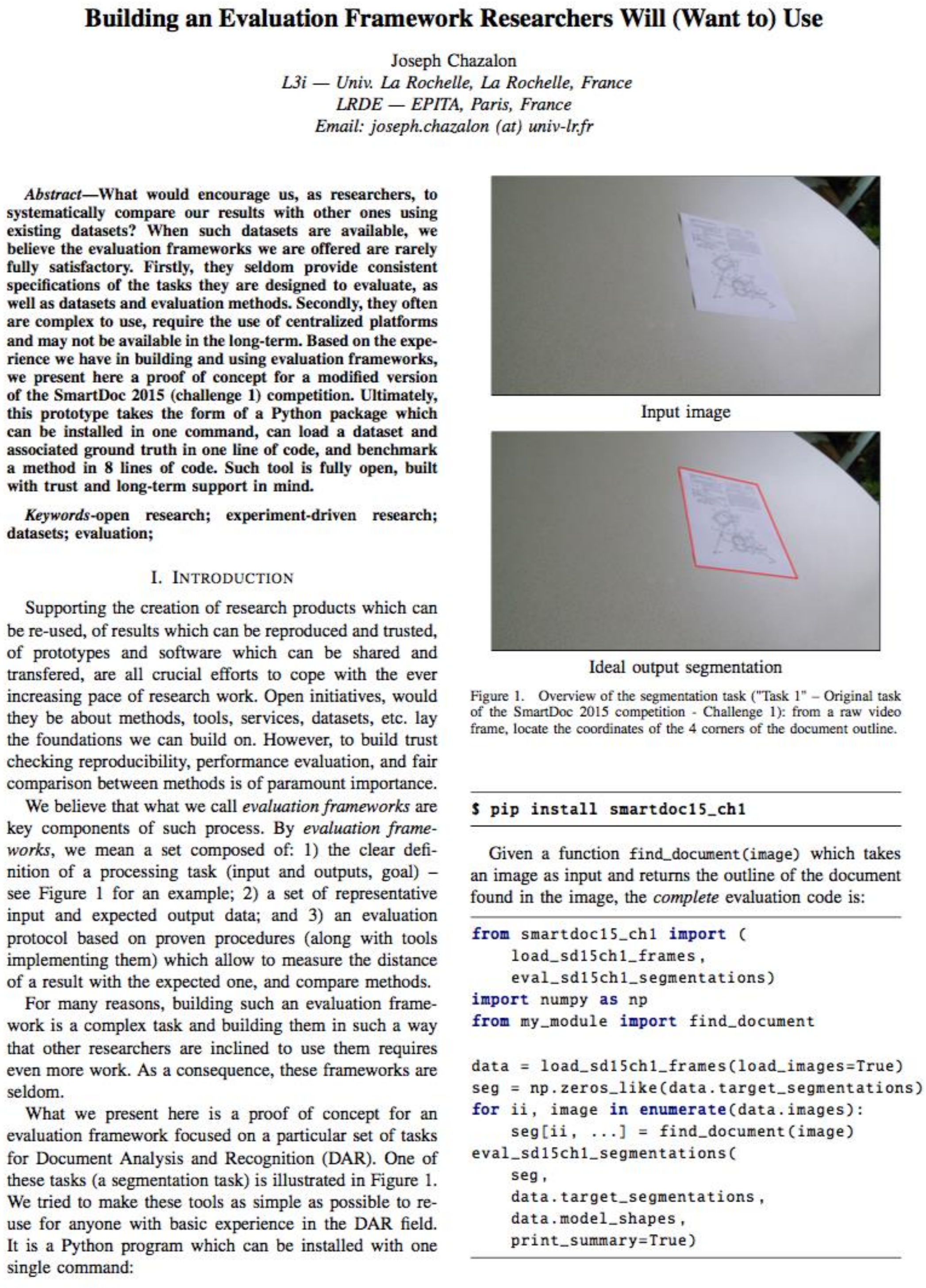
[4] V. Romero, A. H. Toselli, L. Rodríguez, and E. Vidal, “Computer Assisted Transcription for Ancient Text Images,” in International Conference on Image Analysis and Recognition, 2007, pp. 1182–1193.

[5] C. Clausner, S. Pletschacher, and A. Antonacopoulos, “Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments,” in International Conference on Document Analysis and Recognition, 2011, pp. 48–52.

[6] H. Z. Nafchi, S. M. Ayatollahi, R. F. Moghaddam, and M. Cheriet, “An Efficient Ground Truthing Tool for Binarization of Historical Manuscripts,” in International Conference on Document Analysis and Recognition, 2013, pp. 807–811.

[7] A. Garz, M. Seuret, A. Fischer, and R. Ingold, “GraphManuscribble: Interact intuitively with digital facsimiles,” in International Conference on Natural Sciences and Technology in Manuscript Analysis, 2016, pp. 61–63.

Joseph Chazalon (Laboratoire Informatique, Université de La Rochelle, La Rochelle, France)



Building an Evaluation Framework Researchers Will (Want to) Use

Joseph Chazalon
L3i — Univ. La Rochelle, La Rochelle, France
LRDE — EPITA, Paris, France
Email: joseph.chazalon (at) univ-lr.fr

***Abstract*—What would encourage us, as researchers, to systematically compare our results with other ones using existing datasets? When such datasets are available, we believe the evaluation frameworks we are offered are rarely fully satisfactory. Firstly, they seldom provide consistent specifications of the tasks they are designed to evaluate, as well as datasets and evaluation methods. Secondly, they often are complex to use, require the use of centralized platforms and may not be available in the long-term. Based on the experience we have in building and using evaluation frameworks, we present here a proof of concept for a modified version of the SmartDoc 2015 (challenge 1) competition. Ultimately, this prototype takes the form of a Python package which can be installed in one command, can load a dataset and associated ground truth in one line of code, and benchmark a method in 8 lines of code. Such tool is fully open, built with trust and long-term support in mind.**

***Keywords*-open research; experiment-driven research; datasets; evaluation;**

I. Introduction

Supporting the creation of research products which can be re-used, of results which can be reproduced and trusted, of prototypes and software which can be shared and transfered, are all crucial efforts to cope with the ever increasing pace of research work. Open initiatives, would they be about methods, tools, services, datasets, etc. lay the foundations we can build on. However, to build trust checking reproducibility, performance evaluation, and fair comparison between methods is of paramount importance.

We believe that what we call *evaluation frameworks* are key components of such process. By *evaluation frameworks*, we mean a set composed of: 1) the clear definition of a processing task (input and outputs, goal) – see Figure 1 for an example; 2) a set of representative input and expected output data; and 3) an evaluation protocol based on proven procedures (along with tools implementing them) which allow to measure the distance of a result with the expected one, and compare methods.

For many reasons, building such an evaluation framework is a complex task and building them in such a way that other researchers are inclined to use them requires even more work. As a consequence, these frameworks are seldom.

What we present here is a proof of concept for an evaluation framework focused on a particular set of tasks for Document Analysis and Recognition (DAR). One of these tasks (a segmentation task) is illustrated in Figure 1. We tried to make these tools as simple as possible to re-use for anyone with basic experience in the DAR field. It is a Python program which can be installed with one single command:

Input image

Ideal output segmentation

Figure 1. Overview of the segmentation task ("Task 1" – Original task of the SmartDoc 2015 competition - Challenge 1): from a raw video frame, locate the coordinates of the 4 corners of the document outline.

```
$ pip install smartdoc15_ch1
```

Given a function find_document(image) which takes an image as input and returns the outline of the document found in the image, the *complete* evaluation code is:

```
from smartdoc15_ch1 import (
    load_sd15ch1_frames,
    eval_sd15ch1_segmentations)
import numpy as np
from my_module import find_document

data = load_sd15ch1_frames(load_images=True)
seg = np.zeros_like(data.target_segmentations)
for ii, image in enumerate(data.images):
    seg[ii, ...] = find_document(image)
eval_sd15ch1_segmentations(
    seg,
    data.target_segmentations,
    data.model_shapes,
    print_summary=True)
```

The source code for all the material presented here is available at the following URLs:

***
https://github.com/jchazalon/smartdoc15-ch1-dataset
https://github.com/jchazalon/smartdoc15-ch1-pywrapper
***

After a brief review of a selection of notable initiatives which support open or reproducible research, with a focus on DAR (Sec. II), we present our contribution (Sec. III): a proof-of-concept evaluation framework based on a task-oriented definition of the SmartDoc 2015 dataset (Sec. III-A), a new distribution scheme for the raw dataset (Sec. III-B), and an easy to use Python wrapper for data loading and method evaluation (Sec. III-C).

II. Related Work

As mentioned in introduction, we take into consideration 3 aspects of evaluation which constitute a consistent evaluation framework: task definitions, datasets and evaluation methods.

Within the Document Analysis and Recognition (DAR) community, the Technical Committees 10[1] and 11[2] maintain a list of publicly available datasets for research use. While the datasets listed by the curators are generally free to obtain and have research-friendly licenses, it is not uncommon that the download links get broken, causing long-term availability issues because of unreliable hosting solutions. Furthermore, regarding evaluation tools, there are even fewer options and it is very rare that free and open tools are released. A notable exception is the excellent UNLV-ISRI set of OCR evaluation tools [1] which remains used and maintained, even 20 years after its release.

To our knowledge, the first initiative which really embraced the three aspects of an evaluation framework is the Robust Reading Competition (RRC) series [2]. It provides a clear definition of the different problems (or tasks) researchers can evaluate their methods against. Datasets are hosted on the platform; upon submission of method results an evaluation and a ranking of the methods are automatically performed. This approach now faces the issue of raising hosting and maintenance costs. In response, the platform is now progressively opened to avoid interrupting the service. This solution could also solve the issue of having closed-source evaluation procedures and secret ground truth for some datasets.

The Document Analysis and Evaluation (DAE) platform [3] is another important initiative. It was more general in the sense that it aimed at offering an orchestration platform which could be used to compose data processing units (exposed as web services) and data sources. The complexity of the packaging of processing units made the dataset distribution aspect much more successful that the processing one [4]. In order to avoid availability issues for datasets, a new version of the platform (DAE-NG) [4] now focuses on building a federation of synchronized dataset repositories. However this still requires the maintenance of some dedicated platforms.

A last initiative from the DAR community we would like to mention is the DIVA Services platform [5]. It enables to easily package document processing tools using Docker containers and allows to either test or combine them using a REST API. Hosting evaluation methods on this platform is promising, but it requires hosting datasets on a separate platform and can lower the consistency between data and evaluation. Finally, while the authors build the platform with openness in mind, deploying and maintaining a complex platform is necessary to run experiments, penalizing long-term availability.

An approach which conciliates long-term availability, ease of use and a strong consistency between dataset content and evaluation procedures is the Scikit-learn Dataset API [6]. This free and open Python library can be installed with a single command and Scientific Python is a new standard for computer vision research. It often makes evaluation very easy to implement thanks to the many standard functions provided. The Dataset API enables to load many datasets in a computable format with a single line of code. We believe such approach is an interesting alternative to centralized evaluation platforms.

III. Building an Open Evaluation Framework

Based on the original version of the challenge 1 of the SmartDoc 2015 competition, we revised its task definition to enable a wider use of the dataset and the evaluation tools we created (Sec. III-A). In order to facilitate the dissemination and the evolution of this new evaluation framework, we separated the implementation of the raw dataset distribution (Sec. III-B) from the Python library ("wrapper") which enables the manipulation of its content as *computable* objects as well as the straightforward evaluation of any method which complies with the previous task definition (Sec. III-C).

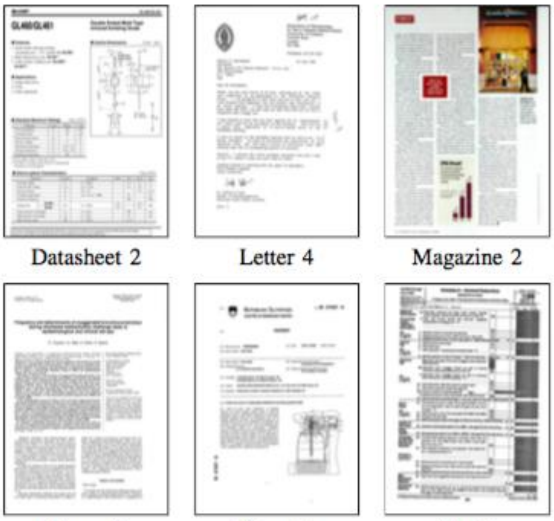
As previously mentioned, all the sources and products of this proof of concept are freely available online. Our main goals were to build an evaluation framework which would be: 1) as easy to use as possible; 2) available in the long term; 3) reliable and trustworthy.

A. New Task Definitions

As previously mentioned, the original dataset we used as the basis of this proof of concept is the SmartDoc 2015 database for document capture (challenge 1) [7]. This dataset was initially created to evaluate the performance of smartphone applications for document image acquisition, focusing on one of the first stages of the pipeline: the segmentation of the document outline in video frames or pictures, in order to allow the correction of perspective distortion.

The dataset was built by capturing 30 document models (5 for each of the 6 different types as shown in Figure 2) under 5 different background scenarios (as visible in Figure 3). Some small noise and margins from the original document images were removed and finally the images

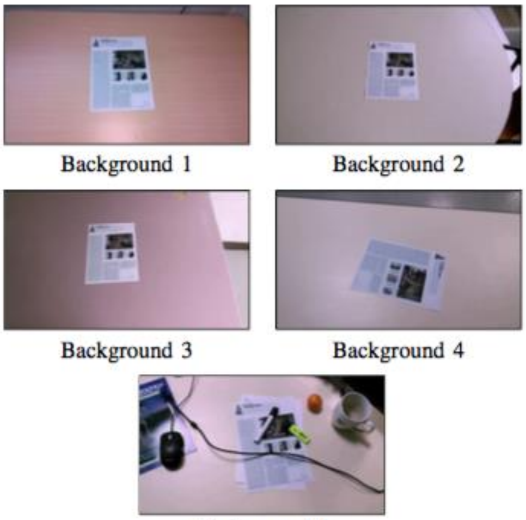
[1] http://iapr-tc10.univ-lr.fr
[2] http://tc11.cvc.uab.es/



Datasheet 2 Letter 4 Magazine 2

Paper 5 Patent 4 Tax 3

Figure 2. Sample documents used in our dataset.



Background 1 Background 2

Background 3 Background 4

Background 5

Figure 3. Sample backgrounds used in our dataset.

were rescaled to all have the same size and fit an A4 paper format, resulting in several variants of the 30 model images. In addition of the video clips, a picture of each of the documents was captured to be used as another set of models variants.

Each of these documents were printed using a color laser-jet and captured using a Google Nexus 7 tablet. The dataset consists of 150 video clips, comprising near 25 000 frames, captured by hand holding and moving the tablet. The video frames present realistic distortions such as focus and motion blur, perspective, change of illumination and even partial occlusions of the document pages. The ground truth of segmentation data was created by semi-automatically annotating the quadrilateral coordinates of the document position for each frame in the collection.

The new version of the dataset makes use of the model images we created and captured. Each of the new tasks we are about to introduce can have two variants: 1) a *model-agnostic* variant with no knowledge of the original document models; 2) a *model-aware* variant based on the knowledge of the complete set of document model images. This second type of scenario allows to test application like augmented reality or form digitization.

Researchers can test their methods against three tasks using this new dataset. For each of them, the *model-aware* variant is obtained by adding one extra input: the set of model image (or the result of the indexation of the latter).

Task 1 – Segmentation: (original task) Inputs are video frames and expected output is composed of the coordinated of the four corners of the document image in each frame (top left, bottom left, bottom right and top right). The evaluation is performed by computing the intersection over union ("IoU" or also "Jaccard index") of the expected document region and the detected region. The frame coordinates are projected onto the document referential in order to allow comparisons between different frames and different document models.

Task 2 – Model classification: (new task) Inputs are video frames and expected output is the identifier of the document model represented in each frame. There are 30 models named "datasheet001" to "tax005". The evaluation is performed as any multi-class classification task would be.

Task 3 – Model type *classification:* (new task) Inputs are video frames and expected output is the identifier of the document model *type* represented in each frame. There are 6 models types, each having 5 members, named "datasheet", "letter", "magazine", "paper", "patent" and "tax". The evaluation is performed as any multi-class classification task would be.

B. Raw Dataset Distribution

We changed the dataset format to adapt it to the new task specifications. We did so with the objective of supporting long-term availability, which also has consequences on the hosting strategy we followed. The resulting product has a dedicated Github project available at: https://github.com/jchazalon/smartdoc15-ch1-dataset. Github hosting has the immediate benefit of making the project easy to find and self-documented thanks to the embedded documentation viewer.

In order to design a dataset still usable in ten years, we reviewed and changed the data format used for the original version of the dataset. The original distribution was based on a set of video files, along with XML files in a custom format for the ground truth. The files were available at some secure file server using a procedure sent by email to users after they registered and accepted the dataset license on the main website. This process is sustainable thanks to email automation but it does not support automated download from a Python script. Furthermore, the explicit license agreement is an unnecessary burden for a standard Creative Common license.

We started by producing a format which makes reading data as simple as possible. We first extracted all the frames from the video files and saved them as JPEG images to minimize decoding issues. We then created a simple CSV file for storing all metadata about each video frame: each line represents a frame observation, the columns store either information about file location or the ground truth for each task. The format is documented to remove any ambiguity about content types. The resulting files (images, metadata, documentation, license, etc.) are packaged into a gzipped TAR archive for maximal compatibility. The model images were packaged using the same process.

Regarding data hosting and distribution, we considered several options with the constraint of being widely accessible and durably available. We chose to use Github[3] *releases* as they feature very simple HTTPS upload and download while supporting version numbers. We also considered (and still plan to use) the EU-funded Zenodo[4] platform which specifically targets the archiving and distribution of research datasets.

The resulting solution makes data archives directly downloaded (and verifiable using SHA256 checksums) with implicit license agreement. Users will be able to download specific versions of the dataset based on the version number they carry. We hope versioning will encourage collaboration and improvement of dataset, while allowing to keep track of which sets of version numbers produce comparable evaluation results.

C. Python Wrapper for Data Loading and Evaluation

Using this reliable dataset distribution, we built a Python wrapper with three goals in mind: 1) making the dataset *"computable"* thanks to the automated loading of the raw archive as ready-to-compute objects; 2) making evaluation so easy researchers will want to play with; and 3) leveraging openness to maximize trust in the correctness of the method, as well as enabling long-term availability.

The resulting code has a very simple API organized around two kind of functions: loading functions (one for the frames, one for the models), which have options to load images, pre-process them and load the associated ground truth for each tasks; evaluation functions (one for each of the three tasks). We tried to comply as much as possible with Python and SciPy philosophies in order to provide researchers with a plug and play library, as illustrated in the listings of the first page. In particular, every data series is a Numpy array which supports all the useful operations researchers are familiar with. For instance, is it possible to directly pass the result of the dataset loading to the automated function of Scikit-learn for separating training and test sets to perform cross validation. The evaluation therefore complies with the usual evaluation pattern for estimators, comparing the target results with actual ones.

In order to maximize the usability of this proposed evaluation framework, we packaged this Python wrapper as a PIP package listed on the central Python Package Index (PyPI) and installable with a single `pip install` command. Finally, the source of these tools can be inspected, forked, maintained, improved and contributed to at: https://github.com/jchazalon/smartdoc15-ch1-pywrapper. Thanks to the online documentation viewer, this address also features a simple and nice entry point for any researcher willing to experiment with this prototype.

ACKNOWLEDGMENT

The author would like to thank the organizers of the Workshop "OpenX for Manuscript Research" for their kind invitation. This work has been conducted within the MOBIDEM project, part of the *Systematic Paris-Region* and *Images & Network* Clusters, funded by the French Government and its economic development agencies.

REFERENCES

[1] S. V. Rice and T. A. Nartker, "The ISRI analytic tools for OCR evaluation," UNLV/Information Science Research Institute, Technical Report TR-96-02, 1996.

[2] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, "ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email)," in *Int. Conference on Document Analysis and Recognition*, 2011, pp. 1485–1490.

[3] B. Lamiroy and D. Lopresti, "The Non-Geek's Guide to the DAE Platform," in *10th IAPR International Workshop on Document Analysis Systems*, 2012.

[4] B. Lamiroy, "DAE-NG: A Shareable and Open Document Image Annotation Data Framework," in *Proceedings of the International Conference of Document Analysis and Recognition (ICDAR)*, Nov. 2017, pp. 31–34.

[5] M. Würsch, R. Ingold, and M. Liwicki, "SDK Reinvented: Document image analysis methods as RESTful web services," in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on.* IEEE, 2016, pp. 90–95.

[6] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[7] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. Luqman, M. Mehri, N. Nayef, J. Ogier, S. Prum, and M. Rusiñol, "ICDAR2015 competition on smartphone document capture and OCR (smartdoc)," in *Int. Conference on Document Analysis and Recognition*, 2015, pp. 1161–1165.

[3] https://github.com

[4] https://zenodo.org

Vinodhrajan Sampath (Department of Informatics, Universität Hamburg, Germany)

Interactive Exploration of Digitized Manuscripts: Introducing the iXMan_Lab

1. Introduction
1.1 Computing in the Context of Humanities

With the recent advances in theories, methods, and applications of various computational techniques (pertaining to image processing and archiving - among others) in the Humanities and the consequent emergence of Digital Humanities (or eHumanities) as a scientific discipline, there has been an explosion of tools enabling thorough manuscript understanding and/or digital paleography. Even though a wide variety of these techniques and tools are aimed at supporting scholarly work, only a few of the tools (particularly in the case of digital paleography) have found wide-spread and consistent acceptance.

The reasons for this can be briefly outlined as follows. Most of the tools have been developed solely from the point of view of Informatics and often do not take into consideration the specific requirements of end-users and, consequently, the methods (and their interfaces) are not particularly tailored to their needs. Hence, this approach typically produces tools, which are probably scientifically well-set and challenging but rank rather low in terms of usability and usefulness in their day-to-day workflow. Unsurprisingly, this results in low tool adoption by the intended user community of scholars and undermines the primary research outcome in terms of actual impact in the intended application domain.

1.2 Approaching Computing in Humanities

Given the complexity of Humanities scholars' workflows and the hitherto co-existence – or even parallelism – of quite different scientific cultures, theoretically grounded computational methods that are intended to be applied to Humanities cannot, or rather, should not be approached through the viewpoint of Informatics alone. It is high time for a paradigm change that involves taking users' perspectives, understanding their workflow, analyzing their requirements and framing relevant scientific challenges, drawing upon well-understood computational theories and methods, and devising a put-the-user-in-control (*a.k.a* keep-the-user-in-the-loop) methodology. We should aim to develop such user-centered systems and attempt to experimentally evaluate such systems through real-world applications, instead of benchmarking them using randomly chosen digitized manuscripts (as in the case of manuscript studies). Such a paradigm change might be novel in the context of Digital Humanities but is intrinsic to modern Informatics. The approach behind has been variously named as *Design Thinking*[2] and *Software Technology for Evolutionary Participative System* or simply *STEPS* (Floyd et al., 1989). As a necessity, the integration of potential users is required from the outset for the sake of arriving at acceptable and truly useful tools with a level of complexity users are able to master.

2. Need for Interactive Exploration

Applied research usually focusses on finding solutions to specific well-defined problems with very particular assumptions and, as a result, any application of the research output is constrained by the target domain. However, real-world applications are rarely well-defined with clear hypotheses and, often, data is imperfect. To complicate things further, very frequently, assumptions made during the development of methods do not actually hold true, particularly with respect to manuscript analysis. Scholars regularly deal with digitized images of very poor quality that usually lack metadata. For problems like word spotting, writer identification or dating, the application (or even the development) of scientific techniques requires the existence of a proper *ground truth* to test the hypothesis/efficiency of the methods or to apply any sort of machine learning. More often than not, such annotated data do not exist, with neither the scholar nor the Informatics researcher having any means to verify, evaluate or even apply a particular method. Also, the scholars have often very little understanding of how to set the method's parameters for the required results, or even to interpret those results. This seriously limits their ability to optimally apply a method to their problems.

It is seldom possible to provide specific computational solutions that are generalizable and applicable to scenarios that scholars face on a day-to-day basis. Therefore, we propose that along with such specific solutions, we must also focus strongly on providing them with required toolsets that would let them explore various methods, deal with digitized images and create custom-built solutions themselves

[2] https://dschool.stanford.edu

in an interactive way. They should be able to choose various techniques or even chain them as necessary. The key is to tailor the interactive exploration to the needs of scholars, while enabling them to master (over time) the unavoidable technical complexity. As mentioned earlier, with several methods needing various input parameters and producing results that are often numbers (without any context), visualization also goes hand-in-hand with enabling the exploration process. It is necessary to adopt proper visualization techniques that will allow users to interpret input/output relations depending on the parameter regimes and perhaps even help them understand how a method works. Visualization can further be employed to create workflows and further assist in performing productive exploration. An appropriate interaction paradigm is thus indispensable in such a context.

An integrated and interdisciplinary approach takes all these aspects into consideration and provides an efficient tool principally aimed at the end-users. It will encapsulate all the technicalities of various methods in a very intuitive and user-friendly way and provide a tool that users can use to create their own solutions for their problems. Apart from a demonstrated competence in both Informatics and Humanities, a prerequisite of such an approach is undoubtedly an eye-to-eye-level communication between the disciplines with the purpose of deeply understanding not only the demands and needs but also the limits of theory and the feasibility of realization. Simply, this is to avoid too high expectations and the resulting sobering frustration. Such a kind of requirement engineering is both painstaking and time-consuming and often not scientifically rewarding in the classical sense due to its trial-and-error nature with a high probability of pitfalls. But this is a necessary evil that needs to be surmounted in order to deliver tools that are *production-ready* in the parlance of Software Engineering.

3. iXMan_Lab

In this context, we introduce the *iXMan_Lab* (**i**nteractive e**X**ploration of **Man**uscripts **Lab**oratory) at the Department of Informatics, University of Hamburg. The underlying motto for the laboratory is to develop concepts, paradigms and prototypes that contribute to the realization of usable and useful software tools for manuscript scholars, which they can use in their day-to-day activities as discussed earlier. The lab consists of an interdisciplinary team utilizing a multi-touch table environment with high-performance computing equipment as a medium for a two-fold aim: First, experimentally designing a manageable processing chain based on computational vision methods for analyzing digitized manuscripts and, second, freezing-in a validated (or even evaluated or benchmarked) processing chain by consensus in order to deliver a useful tool for a broad range of users. In terms of hardware capabilities, the laboratory currently has a custom built 65-inch multi-touch table (MTT) supported by a multi-core gaming engine. Furthermore, the MTT is additionally capable of being adjustable to a wide range of height and angle settings. The laboratory is completely equipped in terms of running GPU-accelerated image processing algorithms, and if necessary, running deep learning methods as well.

Even though primarily situated within the Department of Informatics, the lab is uniquely placed within the Centre for Manuscript Studies as well through its ability to web-interact with various scholars from the sub-projects of *SFB 950 – Manuscript Cultures in Asia and Africa*. This allows the laboratory to perform meticulous requirement engineering due to close interaction between scholars from Manuscript Studies and Informatics.

3.1 Advanced Manuscript Analysis Portal (AMAP)

Currently, the main focus of the iXMan_Lab is concerning the further development of the *Advanced Manuscript Analysis Portal (AMAP)* (Rajan & Stiehl, 2018) equipped with an intuitive interaction paradigm in the context of a multi-touch table. This will allow users to intuitively deal with various advanced image processing techniques and other manuscript-related methods and create their own customized processing chains or perform one-time analyses. The goal is to design and develop AMAP in such a way that even advanced methods could be applied in an easy and intuitive manner by scholars without any technical background. We are designing AMAP to be able to encourage the exploration of various techniques, methods and workflows and, at the same time, to be easy without any steep learning curve. We particularly chose to implement AMAP in a MTT, as we believe touch-based technology is gaining huge traction and has the potential of a primary mode of interaction in the near future. Even now, touch interfaces are becoming more and more popular compared to the traditional Windows-Icons-Mouse-Pointers (WIMP)-based interfaces. Also, having a large-scale interaction/interface area available is necessary to interact with multiple high-resolution images, which is usually the case with analyzing digitized manuscripts. It can further be augmented to allow multiple input modalities that could be harnessed to make the system even more natural by its ability to model and mirror physical real-world interaction, e.g. by speech and deixis, with manuscripts as much as possible. A MTT is also an ideal medium to encourage real-time collaboration of a team of scholars through the provision of a sharable large-scale monitor-based interaction device.

For enhancing AMAP, we are currently implementing an innovative hybrid visual programming language that integrates both a flow-based approach and a block-based approach. The UI paradigm works on the principle of visualizing the digitized documents and computation methods as virtual objects that can be manipulated spatially in relation to each other to perform various chained operations and/or create workflows. The UI is particularly designed to reflect real-world metaphors as much as possible in terms of interaction.

Within the SFB 950, the lab's affiliated scientific service project Z03[3] is currently working on writer identification (Mohammed et al., 2017) and keyword spotting (in continuation of Thomas et al. (2016)). We are also working on integrating both of the tools into AMAP. Our system also offers the ability to integrate other backend systems that provide image processing and analysis techniques as web-based services. This has been actually realized by implementing the methods available at DIVAServices (Würsch et al., 2016) to be a part of AMAP. Such integrations demonstrate the flexibility of our approach as well as the ability to assimilate wide-ranging manuscript-related methods into our platform.

During the workshop, we will provide a live demo of the system to show AMAP in action. We are looking for feedback on the concepts that we are currently drawing upon and perhaps, even suggestions for future ideas that could be developed into viable prototypes are very much welcome.

4. Conclusion

We reported on the current state of computing within Humanities and the way it should be ideally approached. We further discussed the need for interactive exploration of digitized manuscripts that will enable scholars to access various available methods easily. And finally, the current status of AMAP being developed at our lab was outlined. We intend to make the iXMan_Lab a fertile environment that will actively develop, encourage and incubate such ideas, effectively resulting in significant contributions to the field of Digital Humanities in general and the fields of manuscript studies and digital paleography in particular.

References

[1] Mohammed, H., Märgner, V., Konidaris, T., & Stiehl, H. S. (2017). Normalised Local Naïve Bayes Nearest-Neighbour Classifier for Offline Writer Identification. In *Document Analysis and Recognition (ICDAR), 14th IAPR International Conference on (Vol. 1, pp. 1013-1018)*. IEEE.

[2] Konidaris, T., Kesidis, A. L., & Gatos, B. (2016). A segmentation-free word spotting method for historical printed documents. *Pattern Analysis and Applications, 19(4)*, 963-976.

[3] Würsch, M., Ingold, R., & Liwicki, M. (2016). DIVAServices—A RESTful web service for Document Image Analysis methods. *Digital Scholarship in the Humanities*, *32(1)*, i150-i156.

[4] Rajan, V. & Stiehl, H. S. (2018). Bringing Paleography to the Table: Developing an Interactive Manuscript Exploration System for Large Multi-Touch Devices. In *Document Analysis Systems (DAS), 13th IAPR International Workshop*. [Forthcoming]

[5] Floyd, C., Reisin, F. M., & Schmidt, G. (1989). STEPS to software development with users. In European Software Engineering Conference (pp. 48-64). Springer, Berlin, Heidelberg.

[3] funded by DFG, (SFB 950/Z03) 2015 – 2019
https://www.manuscript-cultures.uni-hamburg.de/Poster/Z03_A4_P2.pdf