Article

HisDoc 2.0: Toward Computer-assisted Paleography

Angelika Garz, Nicole Eichenberger, Marcus Liwicki, and Rolf Ingold | Fribourg

Abstract

HisDoc 2.0[1] is a research project on textual heritage analysis and is funded by the Swiss National Science Foundation (SNSF). It builds on the groundwork of the HisDoc[2] project, which concentrated on automated methods for codicological and philological studies. The objective of HisDoc 2.0 is computational paleographical analysis, or more specifically, the analysis of scripts, writing styles, and scribes. While the first project aimed at analyzing simple layouts and the textual content of historical documents, HisDoc 2.0 will be dedicated to complex layouts, including fine-grained text-line localization and script analysis. Furthermore, semantic domain knowledge extracted from catalogs available on databases such as *e-codices*[3] or *manuscripta mediaevalia*[4] is incorporated into document image analysis. In HisDoc 2.0, we perform fundamental research to facilitate the development of tools that build on existing expert knowledge and will support scholars from the humanities who are concerned with examining and annotating manuscripts in the future.

1. Introduction

Document image analysis (DIA) refers to the process of automatically extracting high-level information from digitized images of documents. HisDoc 2.0 will address documents with complex layouts and on which more than one scribe worked (see fig. 1), in other words, documents that have so far been circumvented by the DIA research community. Existing approaches focus more on subtasks such as layout analysis, text-line segmentation, writer identification, or text recognition. These are naturally interrelated tasks which are usually treated independently. Furthermore, the existing approaches presume certain laboratory conditions, i.e. assumptions about the nature of the input are common practice. These assumptions include high-quality separation of the background and foreground – a problem that is only partially solved[5] – (manually) pre-segmented text-line images, or pre-segmented text written by one scribe only. Given a complex document with one or more main text bodies, annotations, embellishments, miniatures, and so on, traditional methods fail, since there are several different challenges to be met simultaneously. Reliable script analysis and text localization, for example, are mutually dependent: scribe identification relies on exact segmentation of homogeneous text regions on a page, which in the presence of various kinds of scripts or writing styles in turn depends on the ability to discriminate scripts.

Based on this argumentation and the fact that the DIA community has produced a vast number of papers on subtasks of DIA,[6] we intend to move forward with HisDoc 2.0 to work on problems which are composed of several tasks. We will start by integrating text localization, script discrimination, and scribe identification into a holistic approach in order to obtain a flexible, robust, and generic approach for historical documents with complex layouts. 'Flexibility' in this context means that the system can be adapted without much effort so as to handle different styles of documents from different sources. 'Robustness' refers to correct results, while 'generic' means that the method is not restricted to a specific type of document. The second focus of the project is to incorporate existing expert knowledge into DIA approaches by extracting data from semantic descriptions created by experts. A long-term goal is to automatically translate results generated by DIA methods into human-readable interpretations, which can then be used to enhance existing semantic descriptions and assist human experts.

[1] http://diuf.unifr.ch/hisdoc2.

[2] Fischer et al. 2012.

[3] http://www.e-codices.unifr.ch/.

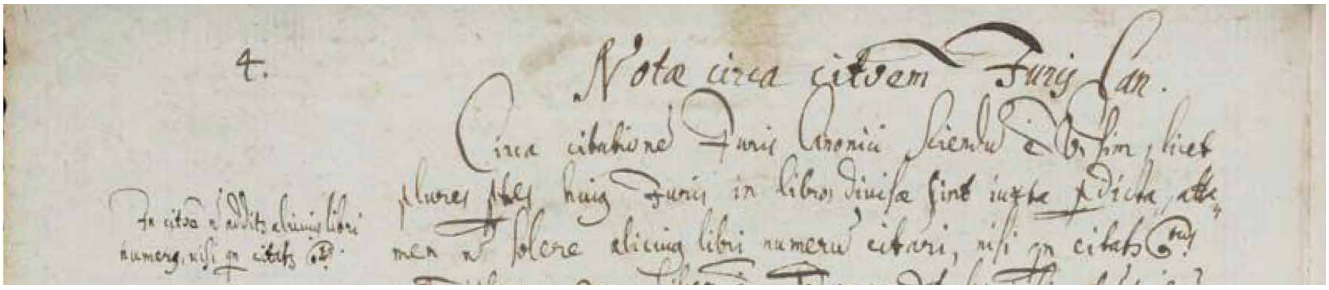
[4] http://www.manuscripta-mediaevalia.de.

[5] Gatos, Ntirogiannis, and Pratikakis 2009; Pratikakis, Gatos, and Ntirogiannis 2010, 2011, 2012, 2013.

[6] See chap. 2, *State of the art*, for a short summary of contributions relevant to the goals of HisDoc 2.0.

a) St. Gallen, Stiftsbibliothek, Cod. Sang. 863, p. 4 (11th century)

b) Sarnen, Benediktinerkollegium, Cod. membr. 8, fol. 9v (1427)



c) Zürich, Zentralbibliothek, Ms. B 124, p. 4 (1655)

Fig. 1: Sample pages of three manuscripts from the e-codices database illustrating several of the challenges to be handled. (a) shows various annotations by different scribes have been added next to the text and between text lines. Separating the annotations from the main text is extremely difficult and is a problem which cannot be solved by existing methods; (b) shows annotations at the top of the page, headings above the images, rubrics, and main text. Methods based on texture analysis are unable to find these text entities, since different scripts have different textures, and there are only a few text lines available for each script. (c) Annotations have been added on this page, and there is a heading placed above the main text body; the textual parts are distinguished by different levels of calligraphical elaboration.

We will concentrate on medieval manuscripts in HisDoc 2.0, but also intend to develop methods that can be adapted to historical documents of different ages and origins without great effort. Three sample pages from medieval and early modern documents depicting several of the challenges to be tackled within the proposed project are shown in fig. 1. These examples demonstrate a variety of scripts, annotation strategies, embellishments, materials, and layouts in different types of manuscripts.

The remainder of the paper is organized as follows: The next section summarizes some of the relevant state-of-the-art approaches to the tasks at hand. Further details are then provided on the projects and the angle we intend to take, followed by a short conclusion.

2. State of the art

Numerous contributions have been published on the issue of solving DIA tasks and subtasks. In this section, we will provide a critical summary of relevant methods with regard to text localization, script analysis, and semantic data.

2.1 Text localization

For the purpose of HisDoc 2.0, the desired outcome of text localization is a set of text lines. While a detailed survey on text-line segmentation with respect to historical documents has been conducted by Likforman et al.,[7] the following presents general research directions along with the most recent approaches. The underlying methods can be categorized as follows:[8] techniques based on projection profiles (PP), smearing techniques, Hough transform techniques, stochastic methods, methods based on thinning operations, and seam-carving methods.

An established method for text-line segmentation in documents with constrained layout is PP, which horizontally accumulates foreground pixels (ink) and results in a histogram encoding a profile of text lines.[9] This method is restricted to constrained documents without any skew or curvature in baselines, however, and fails when applied to documents with complex layouts. Inconsistent inter-line distances as well as touching ascenders and descenders are further problems connected with this method.[10] Various authors[11] have modified global PP in order to correctly segment skewed text blocks and curvilinear text lines by splitting a page into non-overlapping vertical stripes and employing PP only piece-wise. Lines are detected by connecting local minima of the PP of two consecutive stripes.[12]

Smearing methods[13] are based on mathematical morphology; they smudge consecutive foreground pixels along the writing direction. In other words, background pixels between characters are filled with foreground, resulting in an area enclosing the text lines. The fuzzy run length smoothing algorithm (RLSA)[14] is a smearing method which calculates a horizontal fuzzy run length for each pixel resulting in a grayscale image. Text lines are then found by binarizing the run length image. The accuracy of the algorithm depends on the run length chosen and the skew of the text lines. In cases where there are white spaces between words and highly skewed text lines, approaches based on horizontal smearing fail due to similar issues as with PP.

Hough transform has been widely used for purposes of slant, skew, and line detection as well as text-line segmentation.[15] Lines are detected in images based on the peaks in the Hough transform, where gravity centers of connected components (CC) are used as units for Hough transform. This method was modified to a block-based form by Louloudis et al.[16] in order to account for changes in baseline skew.

Recent approaches[17] apply methods of seam carving[18] known from image retargeting. Seams are paths of connected pixels with least entropy, i.e. paths crossing homogeneous areas, not characters, which are used to segment text lines. Each pixel of the image is valued by an energy function, and the seam is then generated by propagating a path of minimum cost through the image. Segmentation of grayscale images is possible, i.e. binarization can be omitted. Drawbacks of these methods include the need for prior knowledge about orientation and number of text lines, however.

An approach for grayscale images which does not depend on prior knowledge about line orientation, number, and curvature has been proposed by Garz et al.[19] This approach exploits density distributions of so-called interest points in order to localize text. Interest points are predominantly found on text and subsequently clustered into lines. Touching components are separated by seam carving.

[7] Likforman-Sulem, Zahour, and Taconet 2007.

[8] Likforman-Sulem, Zahour, and Taconet 2007; Louloudis et al. 2008.

[9] Hashemi, Fatemi, and Safavi 1995; Manmatha and Rothfeder 2005.

[10] Alaei, Pal, and Nagabhushan 2011.

[11] Arivazhagan, Srinivasan, and Srihari 2007; Pal and Datta 2003; Papavassiliou et al. 2010.

[12] Zahour et al. 2001.

[13] Nikolaou et al. 2010; Roy, Pal, and Lladós 2008; Shi and Govindaraju 2004.

[14] Shi and Govindaraju 2004.

[15] Likforman-Sulem, Hanimyan, and Faure 1995; Louloudis et al. 2008.

[16] Louloudis et al. 2008.

[17] Asi, Saabni, and El-Sana 2011; Nicolaou and Gatos 2009; Saabni and El-Sana 2011.

[18] Avidan and Shamir 2007.

[19] Garz et al. 2012, 2013.

2.2 Script analysis

We subsume the terms 'script discrimination' and 'scribe identification' under the term 'script analysis.' These tasks are related, as both examine the properties of script with the target of classification or discrimination. As such, computational features and methods developed for script discrimination can be transferred to the domain of scribe identification and vice versa.

Whereas script discrimination is predominantly based on image statistics and as such is independent of the written content, scribe identification methods can be split into two categories: text-dependent and text-independent methods.[20] The former rely on the comparison of individual character or word images with known textual content and require exact localization and segmentation of the respective entities. The latter extract statistical features from a segmented text block. In order to achieve independence from the textual content, a minimal amount of text is needed.[21] Text-independent methods have the advantage that identification can be performed without the need for handwriting recognition (i.e., extraction of the textual content of an image, which is a non-trivial task for handwriting) or the interaction of a user transcribing and annotating character images. Several comprehensive surveys[22] provide a broad overview of the efforts at text-dependent scribe identification. Text-independent scribe identification approaches prior to 2007 have been reviewed by Schomaker;[23] they can be grouped into texture, structural, and allographic methods.[24]

Approaches based on texture analysis consider a document simply as an image. Features are extracted globally from an image patch extracted from writing areas: Gabor features,[25] angular histograms[26] capturing stroke directions, or combinations which cover slant and curvature, for example.[27] Changes in writing styles, such as differences in word and line spacing, and strokes of varying thickness alter the texture and thus pose certain problems with regard to texture-based methods. A change in scribe between text blocks is easy to deal with. However, a change in scribe between consecutive text lines or even within one line is hard to localize, since an image patch of a certain minimum size is required – usually covering several text lines.

Structural features attempt to capture structural properties of handwriting such as the height of writing zones (x-height, ascenders, or descenders), character width, or slope. They are predominantly extracted from PP and connected components (CC), which requires prior binarization and is problematic if components touch in consecutive lines. Marti and Bunke[28] report a method based on twelve features extracted from binarized segmented text lines: heights of three writing zones extracted from a vertical PP, character width calculated from white runs, slant angle from the character contour, and two features representing the legibility of characters based on fractal geometry. Schlapbach and Bunke[29] propose a stochastic approach using a series of hidden Markov model-based handwriting recognizers and Gaussian mixture models where exactly one model is trained for each scribe, based on nine features which are extracted at text-line level. The output of each recognizer is a transcription along with a log-likelihood score used to rank authors.

The last group of methods is based on the idea that each scribe produces a particular set of personalized and characteristic shape variants of characters – so-called allographs.[30] In computer science literature, the terms *allographic feature* and *writer's invariants* have been used in methods based on writer-specific character shapes. Depending on the actual algorithm[31] that segments (cursive) handwriting into characters, however, a division into allographs cannot be guaranteed. We thus introduce the term *script primitives* to describe meaningful parts of a character which can have a shape ranging from a single stroke to a full character or even a composite of adjacent characters or character parts.

[20] Bulacu and Schomaker 2007; Said, Tan, and Baker 2000.

[21] Brink, Bulacu, and Schomaker 2008.

[22] Impedovo, Pirlo, and Plamondon 2012; Impedovo and Pirlo 2008; Leclerc and Plamondon 1994; Plamondon and Srihari 2000; Rejean Plamondon and Lorette 1989..

[23] Schomaker 2007.

[24] Ibid.

[25] Said, Tan, and Baker 2000.

[26] Bulacu, Schomaker, and Vuurpijl 2003.

[27] Bulacu and Schomaker 2007.

[28] Marti and Bunke 2002.

[29] Schlapbach and Bunke 2007.

[30] Schomaker 2007.

[31] Bensefia, Paquet, and Heutte 2005; Bulacu, Schomaker, and Vuurpijl 2003; Bulacu and Schomaker 2007; Niels, Grootjen, and Vuurpijl 2008; Niels, Vuurpijl, and Schomaker 2007; Schomaker, Bulacu, and Franke 2004; Wolf, Littman, et al. 2010.

A person's handwriting tends to '*entail homogeneous style elements,*'[32] i.e. primitives repeated in different allographs, such as corresponding shapes of descenders or ascenders which can be used to identify a scribe.

Primitives-based methods are applied at character or subcharacter level and are therefore not in principle dependent on the shape of text blocks, baseline curvature, or annotations written between lines. Words are automatically segmented into parts (primitives), a codebook of primitives is computed, and scribe models are built as histograms in the codebook.[33] Several primitives-based methods have been proposed with different classification and retrieval methods. These methods have proven successful for the task of scribe identification on datasets of modern handwriting, and the performance can be boosted when combined with features which capture properties observed at a higher level.[34] An additional future advantage of these approaches over others is the conceivable translation of results into a report which can be easily understood by users.[35]

The character-independent primitives-based approach is fundamentally different from state-of-the-art approaches in human-performed paleography (for Latin and German manuscripts, refer to Bischoff[36] and Schneider[37]). For human writers and readers, the character is the most important reference point: scripts and scribes are discriminated and identified by specific shapes of single characters. A character-independent primitives-based approach therefore introduces a novel perspective to the problem of script analysis, which is different to that of the human experts and could thus be a valuable complement to traditional analysis methods. The crucial problem with regard to the automated approach is the transfer of automatically generated output to a human-understandable and interpretable format so that it can be evaluated and profitably integrated into the work of human experts.

2.3 Semantic data

The most prominent approach to making semantic information accessible for computers is the formalization of ontologies, i.e., the 'formal, explicit specification of a shared conceptualization'[38] in a way that is both human-understandable and machine-readable. The use of such representations facilitates the development of tools to aid humans in identifying, creating, and distributing knowledge in a semi-automatic manner.

The Dublin Core Metadata Initiative[39] constituted a simple standard for metadata descriptions of text, defining information such as title, creator, subject, or publisher. Since this data is best suited for modern texts, several international projects have focused on developing standards for historical documents. The European MASTER[40] project made an attempt to find a unified metadata standard for medieval manuscripts; they defined an XML interface format for machine-readable semantic data. The results of this project have been incorporated into the Text Encoding Initiative (TEI),[41] which defines an XML structure for describing texts in order to make the descriptions machine-readable. Several follow-up projects have focused on defining databases for more specific uses. Kalliope[42] is a database which describes and catalogs literary estates of artists, mainly from the last two centuries. However, the projects mentioned mainly focus on textual contents and relations between documents. They are only partially useful for describing medieval manuscripts, where paleographic information and visual features also play an important role.

Attempts have been made to automatically generate new meta-data using DIA methods,[43] mainly for layout properties. Existing semantic data has not yet been used to improve DIA methods, however, nor have any efforts been made to enhance and verify existing data.

The Genizah project is a project with similar goals to those of HisDoc 2.0.[44] While a sequential approach toward document image analysis has been adopted in the Genizah

[32] Schomaker 2007.

[33] Schomaker 2007.

[34] Bulacu and Schomaker 2007.

[35] Schomaker 2007.

[36] Bischoff 2009.

[37] Schneider 1987, 2009, 2014.

[38] Gruber 1993.

[39] Weibel et al. 1998.

[40] http://xml.coverpages.org/master.html (last accessed: April 14, 2014).

[41] http://www.tei-c.org/ (last accessed: April 14, 2014).

[42] Von Hagel 2004; Shweka et al. 2013.

[43] Le Bourgeois and Kaileh 2004.

[44] Wolf, Dershowitz, et al. 2010.

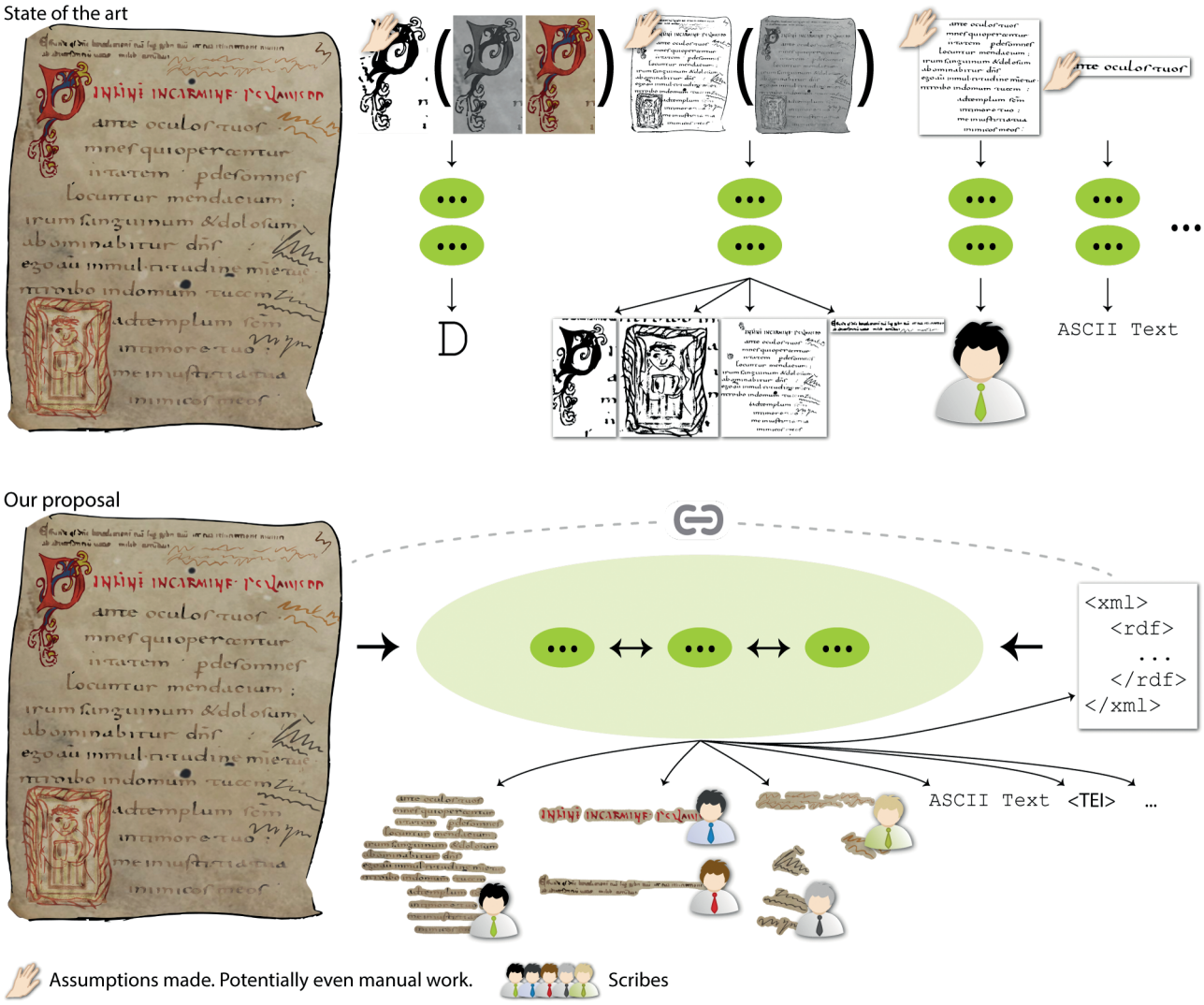


Fig. 2: Comparison between state-of-the-art approaches (top) and the novel proposition in the HisDoc 2.0 project. It is common practice to treat tasks independently, such as processing initials and recognizing the letter they represent, layout analysis, scribe identification, and handwriting recognition. Various assumptions are made about the nature of the input, i.e. high-quality binarization for layout analysis and scribe identification, presegmented texts by one scribe in order to identify the person, or presegmented text line images for handwriting recognition. Contrary to common practice, the tasks are combined into a subproblem in HisDoc 2.0 and are processed in a joint procedure; existing semantic data that is linked to the page image is also incorporated.

project, including manual corrections and processing steps,[45] our aim in HisDoc 2.0 is to exploit interdependencies between several subtasks. Semantic data has been extracted for layout description only, whereas we intend to go a step further in using catalog descriptions and additionally provide data based on the TEI standard.

3. HisDoc 2.0 in detail

The objective of HisDoc 2.0 is to move forward from solving DIA tasks and subtasks toward solving subproblems using an integrated approach combining the tasks of text localization, script analysis, and semantic information for historical documents. Fig. 2 summarizes and describes the difference between the common practice in state-of-the-art methods and our own proposition. The following provides details about the modules of HisDoc 2.0.

3.1 Text localization

Text localization detects the positions of text regions within a page, where the desired outcome is a set of segmented text lines. We will analyze methods for text localization and also propose an algorithm which is not based on assumptions about the layout (such as script, locations, orientations,

[45] Shweka et al. 2013; Wolf, Dershowitz, et al. 2010.

a) Text line localization b) Script discrimination c) Scribe identification

Fig. 3: Transferring knowledge between modules: iterative combination of text segmentation, script discrimination, and scribe identification. The results from each module can be refined by applying knowledge from the other modules.

numbers, types of text entities, and relationships between them) and is capable of handling complex layouts. A paper about a prototype of this method was published as proof of concept at DAS 2012 and received very positive reviews and feedback from the community.[46] An improvement was presented at ICDAR 2013.[47]

3.2 Script discrimination

Script discrimination refers to detecting script changes that occur within a document. We address this issue by unsupervised clustering of uniform textual regions according to their visual similarity, i.e. discriminating scripts of an unknown type and number. The aim is simply to group scripts or writing styles with similar properties, and not to assign a specific scribe or script family. While features that enhance any slight variations between different handwriting are sought for scribe identification, more general features capable of capturing larger variations are needed for script discrimination, since a rather coarse decision is required as to whether or not a region is from the same script. We might therefore rely on a more general subset of features for scribe identification in this module, with the intention of adding further features at a later stage.

3.3 Scribe identification

We regard scribe identification as a more specific task of script analysis. In other words, rather than matching scripts against a database of known writers, we will identify the number of scribes, the point in a text where the scribe changed, or different annotations by the same reader in one manuscript or in a specific part of a manuscript. In addition to analyzing existing features, we intend to study the applicability of interest points to segment script primitives – for both script analysis tasks. They are capable of describing parts of different sizes and can be applied at different granularities, i.e. they can capture a range of details as regards handwriting, from small parts to whole characters and character composites.

3.4 Combining the modules

We aim to combine the three modules of text segmentation, script discrimination, and scribe identification into one holistic approach. There are three conceivable fusion methods for integrating text localization and script discrimination: sequential processing, where script discrimination is performed and the results are included in regions defining text localization within which text lines can be concatenated; joint processing, which includes script discrimination in text-line segmentation; and an iterative approach. Fig. 3 illustrates the process of knowledge transfer between modules. If text locations are known, script analysis can be performed either on text lines or text blocks. Information generated in the script discrimination process helps distinguish different text blocks and facilitates the analysis for scribe identification. Furthermore, we can generate statistical information about handwriting in the text segmentation module, which can be incorporated into the script analysis process.

[46] Garz et al. 2012.

[47] Garz et al. 2013.

3.5 Semantic data

The second major topic of HisDoc 2.0 is semantics. Databases of document collections published online are predominantly annotated with textual descriptions in natural language. This poses the challenge of transforming them into a machine-readable format. While there is a certain amount of structure using XML, the relevant textual descriptions (for example, 'Textura von zwei Händen') are not normalized in terminology and content. Furthermore, the quality and level of detail can vary from one database to another and even within a single database, since different cataloging projects use different guidelines[48] and have a different focus. The first step in this task is to define an ontology for the semantic description of historical documents. We intend to build upon existing database designs. Together with scholars in the humanities who are interested in the scope of the HisDoc and HisDoc 2.0 projects, we will enhance these descriptions by adding axioms for the inference of new knowledge.

The crucial step of deriving computer-readable information from existing textual descriptions will be tackled as follows: existing structured data will be used directly; making use of unstructured information is not as straightforward, however. We plan to use state-of-the-art natural language processing tools to extract information from the textual descriptions, i.e. we will identify entities which are defined in the ontology and automatically derive relations between the instances. The resulting semantic information will enable further automatic processing of the catalog entries in the future.

4. Conclusion and outlook

So far, existing DIA approaches have focused on laboratory conditions for subtasks. While layout analysis and script discrimination methods have been evaluated on simple historical documents only, scribe identification has been performed predominantly on modern handwriting. HisDoc 2.0 will be the first attempt in the DIA community to process text localization and script analysis using a holistic approach that makes use of existing expert knowledge. Our approach is intended to handle complex historical handwritten documents with complicated layouts, additional artifacts, heterogeneous backgrounds, and several scripts within one page, for example. The second major novelty of HisDoc 2.0 is the incorporation of existing semantic information into the DIA process.

While the HisDoc 2.0 project is fundamentally research-based, powerful support tools for scholars from the humanities can be developed based on its results in future. The potential of integrating computational methods into traditional paleographical and codicological analysis performed by human experts is considerable, especially when comparing a number of manuscripts beyond the processing capacity of a single person. In order to benefit from this potential, interdisciplinary collaboration is needed between computer scientists and scholars in the humanities, with the aim of translating computational output into a human-readable format and allowing for its integration into the scholar's work.

ACKNOWLEDGEMENTS

This work is funded by the Swiss National Science Foundation project 205120-150173.

REFERENCES

Alaei, Alireza, Pal, Umapada, and Nagabhushan, P. (2011), 'A New Scheme for Unconstrained Handwritten Text-Line Segmentation.' *Pattern Recognition,* 44.4: 917–28.

Arivazhagan, Manivannan, Srinivasan, Harish, and Srihari, Sargur (2007), 'A Statistical Approach to Line Segmentation in Handwritten Documents', in *Document Recognition and Retrieval XIV*.

Asi, Abedelkadir, Saabni, Raid, and El-Sana, Jihad (2011), 'Text Line Segmentation for Gray Scale Historical Document Images', in *Workshop on Historical Document Imaging and Processing*, 120–25.

[48] Deutsche Forschungsgemeinschaft 1992.

Avidan, Shai, and Shamir, Ariel (2007), 'Seam Carving for Content-Aware Image Resizing', *ACM Transactions on Graphics,* 26.3: 10.

Bensefia, Ameur, Paquet, Thierry, and Heutte, Laurent (2005), 'A Writer Identification and Verification System', *Pattern Recognition Letters,* 26.13: 2080–92.

Bischoff, Bernhard (2009), *Paläographie des römischen Altertums und des abendländischen Mittelalters*, 4th ed. (Berlin: Erich Schmidt Verlag).

Brink, A, Bulacu, Marius, and Schomaker, Lambert (2008), 'How Much Handwritten Text Is Needed for Text-Independent Writer Verification and Identification', in *International Conference on Pattern Recognition*, 1–4.

Bulacu, Marius, and Schomaker, Lambert (2007), 'Text-Independent Writer Identification and Verification Using Textural and Allographic Features', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 29.4: 701–17.

——, Schomaker, Lambert, and Vuurpijl, Louis (2003), 'Writer Identification Using Edge-Based Directional Features.' in *International Conference on Document Analysis and Recognition*, 937–41.

Deutsche Forschungsgemeinschaft, Unterausschuß für Handschriftenkatalogisierung, 1992, *Richtlinien Handschriftenkatalogisierung*, 5th ed. (Bonn-Bad Godesberg: Deutsche Forschungsgemeinschaft).

Fischer, Andreas et al. (2012), 'HisDoc: Historical Document Analysis, Recognition, and Retrieval', in *Digital Humanities, Book of Abstracts*, 94–97.

Garz, Angelika, Fischer, Andreas, Bunke, Horst, and Ingold, Rolf (2013), 'A Binarization-Free Clustering Approach to Segment Curved Text Lines in Historical Manuscripts', in *International Conference on Document Analysis and Recognition*, 1290–94.

——, Fischer, Andreas, Sablatnig, Robert, and Bunke, Horst (2012), 'Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering', in *International Workshop on Document Analysis Systems*, 95–99.

Gatos, Basilis, Ntirogiannis, Konstantinos, and Pratikakis, Ioannis (2009), 'ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)', in *International Conference on Document Analysis and Recognition*, 1375–82.

Gruber, Thomas R. (1993), 'A Translation Approach to Portable Ontology Specifications', *Knowledge Acquisition,* 5.2: 199–220.

Von Hagel, Frank (2004), 'Kalliope-Portal: Fachportal für Autographen und Nachlasse', *Bibliotheksdienst. Organ der Bundesvereinigung deutscher Bibliotheksverbände,* 3.38: 340–47.

Hashemi, M. R., Fatemi, O., and Safavi, R. (1995), 'Persian Cursive Script Recognition', in *International Conference on Document Analysis and Recognition*, vol. 2, 869-873.

Impedovo, Donato, and Pirlo, Giuseppe (2008), 'Automatic Signature Verification: The State of the Art', *IEEE Transactions on Systems, Man, and Cybernetics, Part C,* 38.5: 609–35.

——, Pirlo, Giuseppe, and Plamondon, Rejean (2012), 'Handwritten Signature Verification: New Advancements and Open Issues', in *International Conference on Frontiers in Handwriting Recognition*, 367–72.

Le Bourgeois, Frank, and Hala Kaileh (2004), 'Automatic Metadata Retrieval from Ancient Manuscripts', in *International Workshop on Document Analysis Systems*, 75–89.

Leclerc, Franck, and Plamondon, Rejean (1994), 'Automatic Signature Verification: The State of the Art – 1989–1993', *International Journal of Pattern Recognition and Artificial Intelligence,* 08.03: 643–60.

Likforman-Sulem, Laurence, Hanimyan, A., and Faure, C. (1995), 'A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents', in *International Conference on Document Analysis and Recognition*, 774–77.

——, Zahour, Abderrazak, and Taconet, Bruno (2007), 'Text Line Segmentation of Historical Documents: A Survey', *International Journal on Document Analysis and Recognition,* 9.2: 123–38.

Louloudis, G., Gatos, B., Pratikakis, I., and Halatsis, C. (2008), 'Text Line Detection in Handwritten Documents', *Pattern Recognition* 41.12: 3758–72.

Manmatha, R., and Rothfeder, J. L. (2005), 'A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27.8: 1212–25.

Marti, U. V. and Bunke, Horst (2002), 'The IAM-Database: An English Sentence Database for Offline Handwriting Recognition', *International Journal on Document Analysis and Recognition* 5.1: 39–46.

Nicolaou, Anguelos, and Gatos, Basilios (2009), 'Handwritten Text Line Segmentation by Shredding Text into Its Lines', in *International Conference on Document Analysis and Recognition*, 626–30.

Niels, R. M. J., Grootjen, F. A., and Vuurpijl, L. G. (2008), 'Writer Identification through Information Retrieval: The Allograph Weight Vector', in *International Conference on the Frontiers of Handwriting Recognition*, 481–86.

Niels, Ralph, Vuurpijl, Louis, and Schomaker, Lambert (2007), 'Automatic Allograph Matching in Forensic Writer Identificaiton', *International Journal of Pattern Recognition and Artificial Intelligence,* 21.01: 61–81.

Nikolaou, Nikos et al. (2010), 'Segmentation of Historical Machine-Printed Documents Using Adaptive Run Length Smoothing and Skeleton Segmentation Paths', *Image and Vision Computing,* 28.4: 590–604.

Pal, U. and Datta, S. (2003), 'Segmentation of Bangla Unconstrained Handwritten Text', in *International Conference on Document Analysis and Recognition*, 1128–32.

Papavassiliou, Vassilis, Stafylakis, Themos, Katsouros, Vassilis, and Carayannis, George (2010) 'Handwritten Document Image Segmentation Into Text Lines and Words', *Pattern Recognition* 43.1: 369–77.

Plamondon, R., and Srihari, S. N. (2000), 'Online and Off-Line Handwriting Recognition: A Comprehensive Survey' *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22.1: 63–84.

——, and Lorette, Guy (1989), 'Automatic Signature Verification and Writer Identification – The State of the Art', *Pattern Recognition,* 22.2: 107–31.

Pratikakis, Ioannis, Gatos, Basilis, and Ntirogiannis, Konstantinos (2010), 'H-DIBCO 2010 – Handwritten Document Image Binarization Competition', in *International Conference on Frontiers in Handwriting Recognition*, 727–32.

——, (2011), 'ICDAR 2011 Document Image Binarization Contest (DIBCO 2011)', in *International Conference on Document Analysis and Recognition*, 1506–10.

——, (2012), 'ICFHR 2012 Competition on Handwritten Document Image Binarization (H-DIBCO 2012)', in *International Conference on Frontiers in Handwriting Recognition*, IEEE, 817–22.

——, (2013) 'ICDAR 2013 Document Image Binarization Contest (DIBCO 2013)', in *International Conference on Document Analysis and Recognition*, IEEE, 1471–76.

Roy, Partha Pratim, Pal, Umapada, and Lladós, Josep (2008), 'Morphology Based Handwritten Line Segmentation Using Foreground and Background Information', in *International Conference on Frontiers in Handwriting Recognition*, 241–46.

Saabni, Raid, and El-Sana, Jihad (2011), 'Language-Independent Text Lines Extraction Using Seam Carving', in *International Conference on Document Analysis and Recognition*, 263–68.

Said, H. E. S., Tan, T. N., and Baker, K. D., (2000), 'Personal Identification Based on Handwriting', *Pattern Recognition,* 33.1: 149–60.

Schlapbach, Andreas, and Bunke, Horst (2007), 'A Writer Identification and Verification System Using HMM Based Recognizers', *Pattern Analysis Applications* 10.1: 33–43.

Schneider, Karin (1987), *Gotische Schriften in deutscher Sprache: I. Vom Späten 12. Jahrhundert bis um 1300* (Wiesbaden: Ludwig Reichert).

——, (2009), *Gotische Schriften in deutscher Sprache: II. Die Oberdeutschen Schriften von 1300 bis 1350* (Wiesbaden: Ludwig Reichert).

——, (2014), *Paläographie und Handschriftenkunde für Germanisten. Eine Einführung,* 3rd ed. (Berlin – Boston: De Gruyter).

Schomaker, Lambert (2007), 'Advances in Writer Identification and Verification', in *International Conference on Document Analysis and Recognition*, 1268–73.

——, Bulacu, Marius, and Franke, K., (2004), 'Automatic Writer Identification Using Fragmented Connected-Component Contours', in *International Workshop on Frontiers in Handwriting Recognition*, 185–90.

Shi, Z., and Govindaraju, Venu (2004), 'Line Separation for Complex Document Images Using Fuzzy Runlength', in *International Workshop on Document Image Analysis for Libraries*, 306–12.

Shweka, Roni, Choueka, Yaacov, Wolf, Lior, and Dershowitz, Nachum (2013), 'Automatic Extraction of Catalog Data from Digital Images of Historical Manuscripts', *Literary and Linguistic Computing,* 28.2: 315–30.

Weibel, Stuart, Kunze, John, Lagoze, Carl, and Wolf, Misha (1998), 'Dublin Core Metadata for Resource Discovery', *Internet Engineering Task Force RFC,* 2413: 222.

Wolf, Lior, Nachum Dershowitz, et al. (2010), 'Automatic Palaeographic Exploration of Genizah Manuscripts', in *Kodikologie und Paläographie Im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*, 157–79.

——, Littman, Rotem, et al. (2010), 'Identifying Join Candidates in the Cairo Genizah', *International Journal of Computer Vision,* 94.1: 118–35.

Zahour, Abderrazak, Taconet, Bruno, Mercy, Pascal, and Ramdane, Said (2001), 'Arabic Hand-Written Text-Line Extraction', in *International Conference on Document Analysis and Recognition*, 281–85.