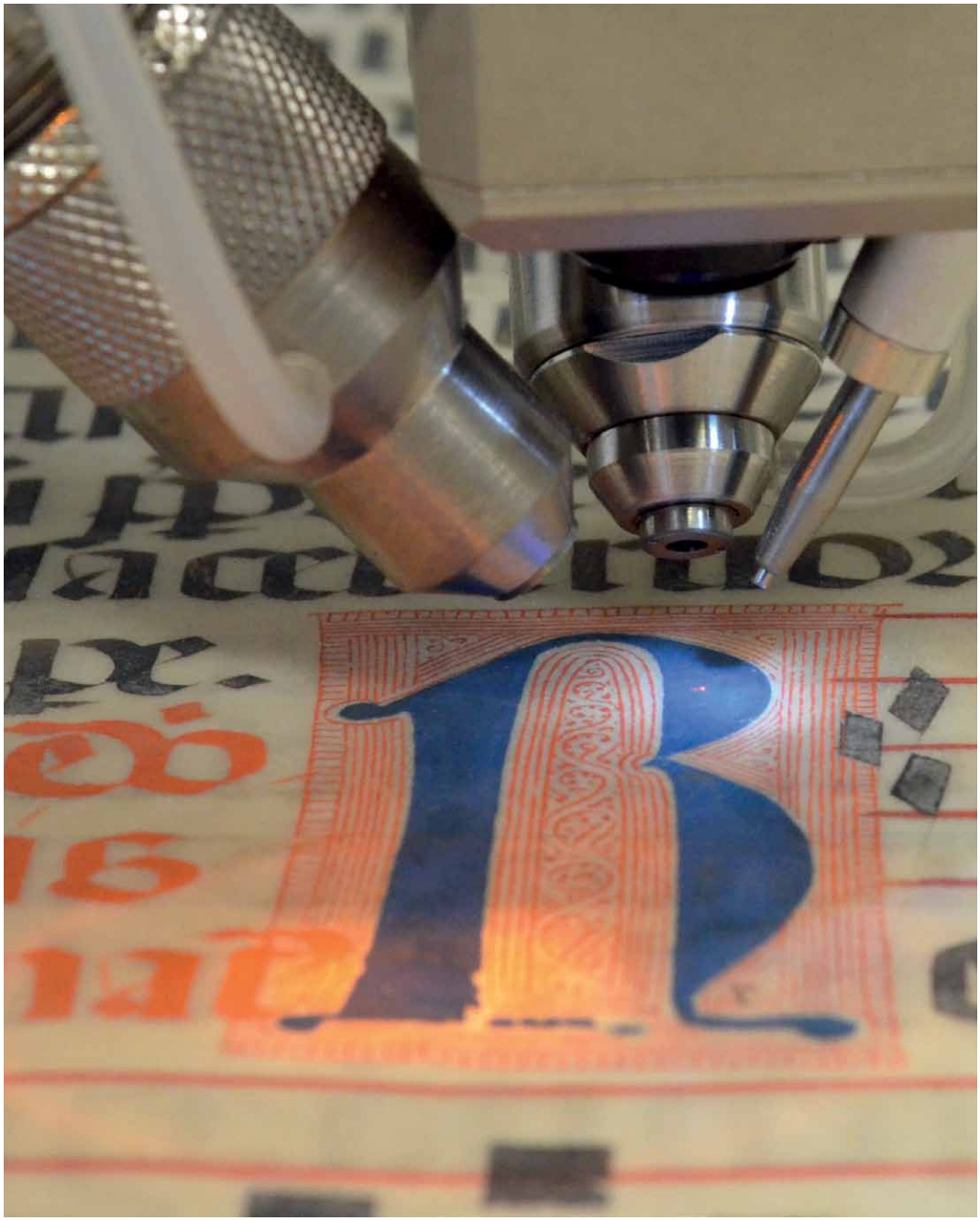
mc Nº 7 2014

manuscript cultures

Hamburg | Centre for the Study of Manuscript Cultures ISSN 1867–9617



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CENTRE FOR THE STUDY OF MANUSCRIPT CULTURES

DFG

Publishing Information

Natural Sciences and Technology in Manuscript Studies
Edited by Christian Brockmann, Michael Friedrich, Oliver Hahn, Bernd Neumann, and Ira Rabin

Proceedings of the 'Conference on Natural Sciences and Technology in Manuscript Analysis' at the University of Hamburg, SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa' – Centre for the Study of Manuscript Cultures, 4–6 December 2013.

Editors
Prof Dr Michael Friedrich
Universität Hamburg
Asien-Afrika-Institut
Edmund-Siemers-Allee 1/ Flügel Ost
D-20146 Hamburg

Tel. No.: +49 (0)40 42838 7127
Fax No.: +49 (0)40 42838 4899
michael.friedrich@uni-hamburg.de

Prof Dr Jörg Quenzer
Universität Hamburg
Asien-Afrika-Institut
Edmund-Siemers-Allee 1/ Flügel Ost
D-20146 Hamburg

Tel. No.: +49 40 42838 - 7203
Fax No.: +49 40 42838 - 6200
joerg.quenzer@uni-hamburg.de

Editorial Office
Dr Irina Wandrey
Universität Hamburg
Sonderforschungsbereich 950
"Manuskriptkulturen in Asien, Afrika und Europa"
Warburgstraße 26
D-20354 Hamburg

Tel. No.: +49 (0)40 42838 9420
Fax No.: +49 (0)40 42838 4899
irina.wandrey@uni-hamburg.de

Layout
Astrid K. Nylander

Cover
Photo: Mobile µ-XRF Spectrometer "Artax" scanning a medieval Latin music manuscript with Fleuronné initials in blue and red; private collection. © CSMC

Translations and Copy-editing
Carl Carter, Amper Translation Service

Print
AZ Druck und Datentechnik GmbH, Kempten
Printed in Germany

ISSN 1867–9617

www.manuscript-cultures.uni-hamburg.de

CONTENTS

Editorial

Natural Sciences and Technology in Manuscript Studies

Dear Reader

In recent years, the emerging field of manuscript studies has come to provide a platform for dialogue between the humanities and the natural sciences, helping to define issues to be tackled by non-destructive technologies developed in the natural and applied sciences. The analyses of visual, physical and chemical properties of manuscripts provide important data for answering questions that cannot be solved by historical and philological methods alone. Multispectral imaging, for example, is already being widely used to recover erased text in palimpsests. Non-destructive material analysis contributes to the classification of writing materials and provenance studies and can potentially be employed to determine the age of manuscripts as well. Finally, image-processing techniques are also gaining recognition in the field of palaeography and codicology.

Growing recognition of the potential of physical and chemical diagnostics for a wide range of applications from conservation and restoration of artefacts to scholarly disciplines from archaeology to philology has led to an increasing number of meetings and publications by researchers. To the best of our knowledge, however, there has not yet been any attempt to assemble experts using and developing methods from the natural and applied sciences that focus exclusively on manuscripts.

The first International Conference on Natural Sciences and Technology in Manuscript Analysis was held at the premises of the Centre for the Study of Manuscript Cultures in Hamburg on 4–6 December 2013. It brought together scientists and scholars engaged in this field of research and provided a forum for discussion and for presenting new methods and results.

This special issue of *manuscript cultures* contains a selection of the papers presented in Hamburg. The articles were solicited for original research work illuminating the role of the natural sciences and technology in manuscript analysis and covered areas such as:

- the recovery of lost writing.
- image analysis of visual manuscript features.
- material analysis of writing materials and writing supports.
- cutting-edge techniques.

All in all, this special issue represents the state of the art, illustrating how different techniques and varying methodologies can be successfully applied to analytical investigations in the field of manuscript analysis. We hope that it will help to integrate the natural and applied sciences into the field of manuscript studies.

We would like to express our gratitude to the German Research Foundation (DFG) and the University of Hamburg for their financial support, to all the authors for submitting persuasive, up-to-date papers, to all the anonymous reviewers for their valuable and constructive comments and finally to the editorial office for their own fruitful contribution to this issue.

Christian Brockmann, Michael Friedrich, Oliver Hahn, Bernd Neumann, and Ira Rabin

Article

Abbreviations in Medieval Latin Handwriting

Björn Gottfried, Marius Wegner, Marianna Spano, and Mathias Lawo | Bremen – Berlin

Abstract

Abbreviations were used extensively in medieval Latin manuscripts. One reason for this was to allow economical use of parchment or other kinds of writing materials, which were relatively expensive during the Middle Ages. This article elaborates on the employment of abbreviations in medieval Latin handwriting and how they can be extracted from digitized documents with the support of an image-processing software system. The main objective of extraction is to characterize different writing hands. According to our research hypothesis, abbreviations reveal a great deal about specific scribes. As a first step, the stability of this criterion is analyzed. As it turns out, a similar amount of abbreviations and a similar distribution regarding their positions within individual words can be found for the same copyist.

1. Introduction

The objectives of the present research are to analyze handwritings and establish the correspondence between the document image and transcription. This allows texts, and particularly abbreviations, to be found and compared in the original documents. Above all, abbreviations require a level of analysis which goes further than the word level. There are several systems being developed to support the transcription process,[1] but current efforts are either text line-based[2] or word-based.[3] A correspondence between the document image and transcription at the level of glyph images and characters would be desirable, however. This would enable one-to-one mapping between all glyph images and their corresponding characters in the transcription. As soon as a word includes abbreviation characters, one-to-many mapping is required. This entails a number of problems regarding the extraction of abbreviations and how they are represented. The current issue, however, concerns the question as to what abbreviations tell us about the copyist (or copyists) of a text. Are there any tendencies regarding the use of abbreviations by specific writing hands?

The next section discusses the general background to the employment of abbreviations in medieval Latin handwritings. Problems arising from automatic detection of glyphs and abbreviations in documents are then presented as well as possible methods for dealing with them. The resulting transcription can be exported for further processing. There are specific conditions which apply to abbreviations in this context, and it is necessary to explain how abbreviations can be found in original document images after they have been transcribed. Finally, a number of documents are analyzed in an evaluation aimed at showing whether abbreviations alone provide a useful criterion for characterizing individual scribes.

2. Abbreviations

There are two general types of abbreviations to be distinguished in Latin texts: abbreviations consisting of letters or combinations of them, and so-called conventional signs, which cannot be connected to letters but may have their origins in ancient tachygraphy (such as $_{9}$ for *com-/con-*, ^{9} for *–us*, $_{7}$ and *&* for *et*, ÷ for *est*, and so on). Literal abbreviations can be divided into 'suspensions' (*Suspensionskürzungen* in German), which are words interrupted after a certain number of letters (≥ 1) regardless of the syntactical connection of the word, and 'contractions' (*Kontraktionskürzungen* in German) consisting of at least the initial letter of a word and its ending.

One-letter suspensions are the oldest type.[4] They appear in a number of public documents dating back to the Roman republic (509–27 BCE), and the majority of them are familiar to us from the uniquely transmitted work of the grammarian Marcus Valerius Probus (fl. 1st century CE), a well-known example being SPQR for *Senatus Populusque Romanus*.

[1] Serrano 2013, Wüthrich et al. 2009, and Romero 2007.

[2] Romero 2007.

[3] Serrano 2013.

[4] Bischoff 2009.

Depending on the context, however, it might also stand for another case such as the accusative *Senatum Populumque Romanum*. This ambiguity was already recognized as a problem in antiquity, and the use of this type of abbreviation in legal texts was prohibited in 438 CE.

The use of contractions arose from the *nomina sacra* ('holy names') in Christian religious texts describing, for instance, the Holy Trinity as *dn̄s p̄r* (*deus pater* = God the father), Jesus Christus as *Ih̄s Xp̄s*, and the Holy Spirit as *sp̄s sc̄s* (*spiritus sanctus*), where all abbreviations are marked by a simple stroke above the central letter or whole combination of letters. It is possible to trace the Greek origin of the middle example, since *h*, *X*, and *p* most certainly stand for Greek *eta*, *chi*, and *rho*. We can therefore assume that many of the issues summarized in this paper for text written in Latin characters can also be applied to other letter systems. Contractions also pose certain problems for readers and editors of ancient manuscripts: given that the most common handbook of abbreviations by Adriano Cappelli[5] comprises not more than around 15,000 abbreviations, it is quite clear that the majority of inflected forms – especially inflected verb forms – are missing. For Cappelli and his human readers, this was no trouble at all, since they had the philological understanding to combine (correctly as well as freely) the abbreviated roots and endings of a word into a known form. Completely different conditions apply to computers and machine-readable resolutions of abbreviations, of course. Thus, the *Abbreviationes* database by Olaf Pluta[6] currently (as of September 2014) 'comprises over 70,000 entries containing a total of 80,098 references to manuscripts', but even that does not contain every single form. The imperfect tense of the Latin verb *habere* (= to have), for example, comprises 24 different forms: six in the active voice (three singular, and three plural) and six in the passive voice, each of them in the indicative and subjunctive moods. Only three of the twenty-four forms can be found in Cappelli:[7] the third-person singular (indicative *h^{s}ebat*, *hēb^{a}t* = *habebat*, subjunctive *hab^{s}&*, *he^{s}t*, *hēr^{s}t*, *h²; = haberet*) and the first-person subjunctive *hr̄em* = *haberem*. As can be seen, there are different ways of abbreviating the same form, and there is no reason why the omitted forms should not be abbreviated in a similar way. It is therefore hardly surprising that the more extensive *Abbrevationes* database includes eight different forms, namely *habebant*, *habebat*, *habebatur*, *haberem* (three types of abbreviation), *haberent*, *haberes*, *haberet* (abbreviated in nine ways), and *haberetur*. On the other hand, the same abbreviation can be used for completely different words. Thus, *mr̄m* might stand for *matrimonium*, *martyrum*, *monstrum* or – not mentioned by Cappelli[8] – also for *magistrum*, *matrum*, *meliorem* if you compare it with *mr̄i*, for example.

As can be seen from the examples cited for abbreviating *habebat* and *haberet*, it is not only the sequence of the letters that matters (as in Cappelli or the search function in *Abbreviationes*), but also the shape of the abbreviation and its position (an issue considered by the otherwise paleographically normalized entries in *Abbreviationes*). This is evident when we look at letters – there is undoubtedly a difference between reading $\underline{g}^{i}$ (= *igitur*) and $\underline{g}^{o}$ (= *ergo*) – or combinations with conventional signs written above the line: $\underline{p}$ usually stands for *per*, $\underline{p}^{i}$ for *pri*, p^{o} for *post*, $\bar{p}$ for *pr(a)e*, and p for *pro* (cf. fig. 2), no matter whether it is used as an isolated word (if possible), as a prefix, or as a substitute for the corresponding combination of letters within the word. Where, however, the most common and clearly ambiguous abbreviation signs in manuscripts are concerned, i.e. straight or curved lines, the question may be posed as to whether not only the *position* of the signs above certain letters might be specific to specific writers, but also the length, shape, and direction of the signs. This is one of the fields where computer science produces results which are far more prolific, easily comparable, and reliable than the results obtained by human analysis. Another area of investigation might be the ratio of abbreviations in a larger or smaller context: not only their quality (which abbreviations are used, are there different types of abbreviation for the same word, and so on), but also their quantity (are there more abbreviations at the beginning of the text than at the end or is there a constant proportion, does the ratio of abbreviated and long forms of the same word vary within the text, and so on).

3. Extracting glyphs

In order to analyze the use of abbreviations, it is necessary to obtain a mapping between glyphs within the original document image and in the transcription. A single glyph maps

[5] Cappelli 1929.

[6] *Abbreviationes* 2014.

[7] Cappelli 1929, pp. 157ab, 159a, 160a, 164b, 165a.

[8] Cappelli 1929, p. 226a.

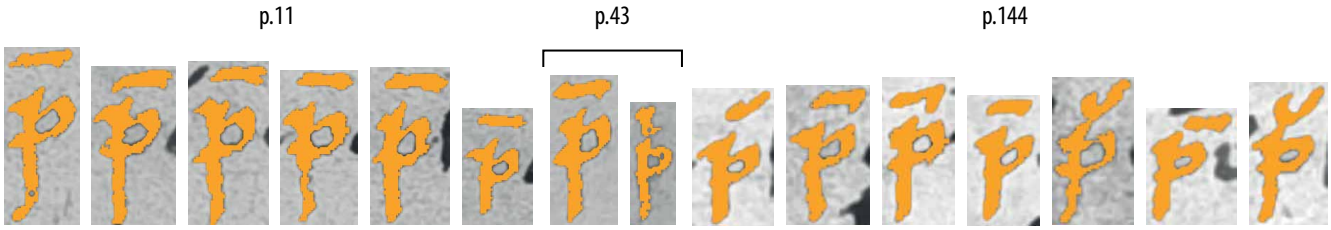


Fig. 1: The abbreviation glyphs of the string 'pre' are shown for three document pages (fol. 11r, fol. 43r, and fol. 144r) distributed over a manuscript of around four hundred pages. Each page is indicated by one of the three boxes. The order of the glyphs corresponds to the order of their appearance in the book. It turns out that there are clear differences in the shape of the glyphs depending on how late in the book they appear. The abbreviation bars in particular differ from case to case. In the last box, the bar is sometimes straight and sometimes curved, whereas the first occurrences look more uniform (box one).

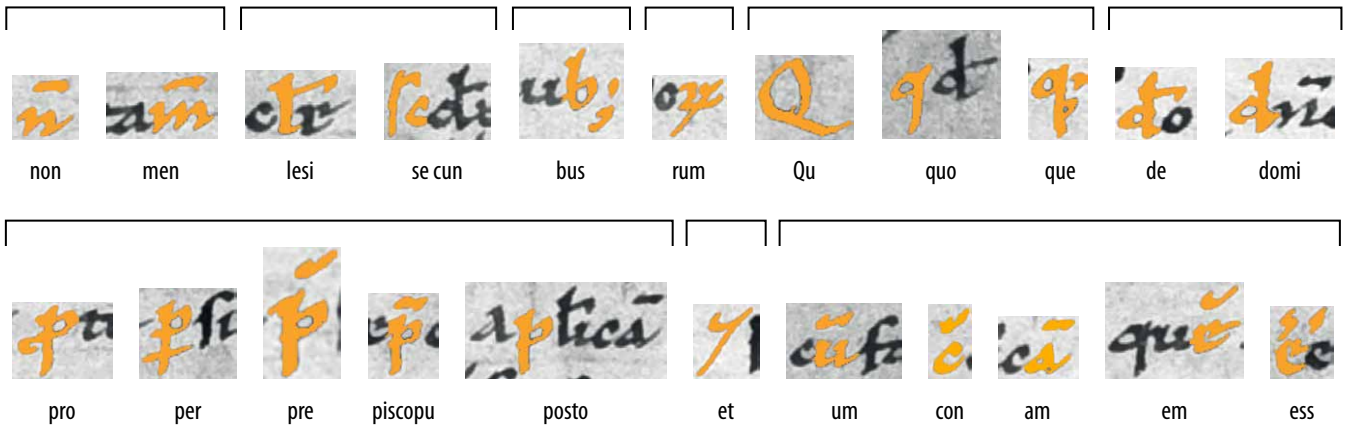


Fig. 2: The most frequent abbreviations employed by Hugh of Flavigny, grouped with respect to visually similar characteristics of the glyph images. These abbreviations are taken from fol. 144r of Philipps 1870.

to a whole string in the case of abbreviations. For this purpose, the software system *Diptychon* employs *$* signs which are used in order to mark the beginning and the end of a string in the transcription. In this way, each chain of characters that is enclosed by *$* signs represents the completed string of an abbreviation. An example is provided in fig. 1, which shows eleven different instances of the character *p* with a bar written above it. This abbreviation expands to *pre*.

The *Diptychon* software system tries to separate adjacent glyphs automatically. A number of difficult cases might result in glyphs being separated incorrectly inasmuch as *Diptychon* does not make any assumptions about the underlying handwriting. Interactive methods are available in cases such as these in order to correct the proposed separations.[9] Abbreviations present an additional difficulty, since they frequently consist of two or even more disconnected components, such as the *p* with a bar above it, as shown in fig. 1. In these cases, the user can let the system know which disconnected regions pertain to the same logical unit.

Each logical unit is mapped to a string in the transcription. This mapping is also done automatically as follows: the rows of the document image are detected and, for each text row, an input field for the transcription is generated on the user interface. The line by line correspondence between document image and transcription breaks down the correspondence problem to single rows, and the correspondence is established for each single row based on the linear order of the glyphs in the image. Apart from some specific difficulties such as supplements between the lines which the user has to correct manually, this approach enables seamless mapping between glyphs and transcription. By employing *$* signs, it especially includes the handling of abbreviation glyphs and the corresponding strings in the transcription.

Some of the most frequent abbreviations used by 11th-century author Hugh of Flavigny in his chronicle are shown in fig. 2. They are grouped with respect to the visual characteristics of the glyph images, showing that similar or even identical glyphs might represent different abbreviations, for example *pre* and *piscopu*. In these cases, the context of the abbreviations is required in order to translate them

[9] Gottfried, Wegner, and Lawo 2013; Gottfried, Wegner, and Lawo 2014.

correctly. A horizontal bar is frequently written above a glyph, indicating an abbreviation. This is not always the case, however; a high degree of expertise is required if paleographers are to recognize abbreviations correctly.

4. Exporting abbreviations

While the user interface of the *Diptychon* system requires the user to employ *$* signs in order to mark abbreviations, these signs are not desired when exporting the transcription to a file. For the critical edition of a manuscript text, however, it can be helpful to indicate where abbreviations are found in the original document. Moreover, the edited text should make it clear which characters mark the beginning of an abbreviation and are hence shown as glyphs, in contrast to the abbreviated letters that do not appear in the original document. All the abbreviated letters are therefore enclosed in parentheses in the export file, and the characters shown as glyphs are placed before the parentheses. This enables the editor to subsequently analyze the abbreviations purely on the basis of the exported transcription. For instance

 is transcribed in *Diptychon* as *concilium* and is exported as *c(on)ciliu(m)*. The isolated abbreviation

 becomes *et* in the transcription and is exported as *e(t)*, and

 is transcribed as *Francorum* and exported as *Francor(um)*.

There are also examples of glyphs which are not the characters that start the abbreviation, such as *qđ* in fig. 2, which stands for *(quo)d*. There are certain abbreviation characters which seem to have been invented solely for the purpose of abbreviating strings. Most other abbreviations are only distinguished by an extra sign extending a common character, such as a vertical bar above the abbreviation glyph, or a point or semicolon to the right of the glyph.

5. Searching for abbreviations

Once the transcription of a document is available, it is possible to search for all occurrences of a string and in particular for occurrences of abbreviations. While the underlying search methods make use of the transcription which is stored in a symbolic form, each character or abbreviation string is linked to the corresponding glyph within the original document

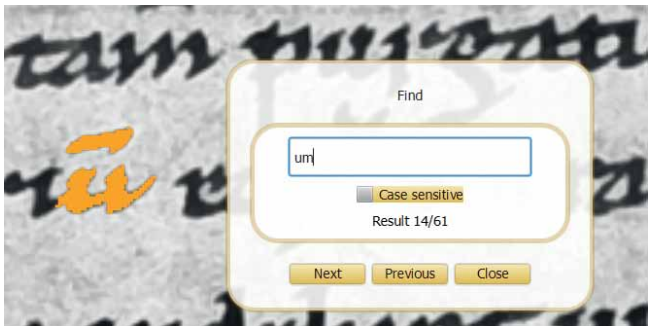


Fig. 3: The search dialog provided by *Diptychon*.

image. The whole region of the latter is accessible, meaning that it can be emphasized by color highlighting. The user gets the impression that the glyphs are found in the original document image. This enables the paleographer to gain an insight into the distribution of abbreviations, their visual appearance, and the contexts in which they are used. Fig. 3 shows the search dialog.

It is possible to search for an arbitrary string which might specifically contain space characters. This allows a search for certain types of abbreviations such as those that start a word. For this purpose, the search string should start with a space character followed by the requested abbreviation. Another option is to search for abbreviations that are found at the end of a word or form an entire word in themselves. In the latter case, the abbreviation has to be enclosed by space characters. Likewise, a broader context of adjacent words can be taken into account. Abbreviations are not used in every case, however, and sometimes all the characters of a word are given. Nevertheless, the search result contains every occurrence of the relevant string, either as part of an abbreviation or given explicitly, in order to show how abbreviations are used in the relevant document.

6. A case study on abbreviation criteria

The long-term strategy is to employ different criteria as a means of characterizing writers. One such criterion concerns the way in which a writer uses abbreviations. Before this criterion can be applied, however, it has to be analyzed to determine its robustness. If a writer uses abbreviations to a similar degree in different documents or in pages of documents written at different times, the employment of abbreviations can be deemed stable and can therefore be regarded as a robust criterion.

As an example, the robustness of abbreviations is analyzed for Hugh of Flavigny, who was a late 11th-century chronicler. A manuscript (probably an autograph) by this writer still

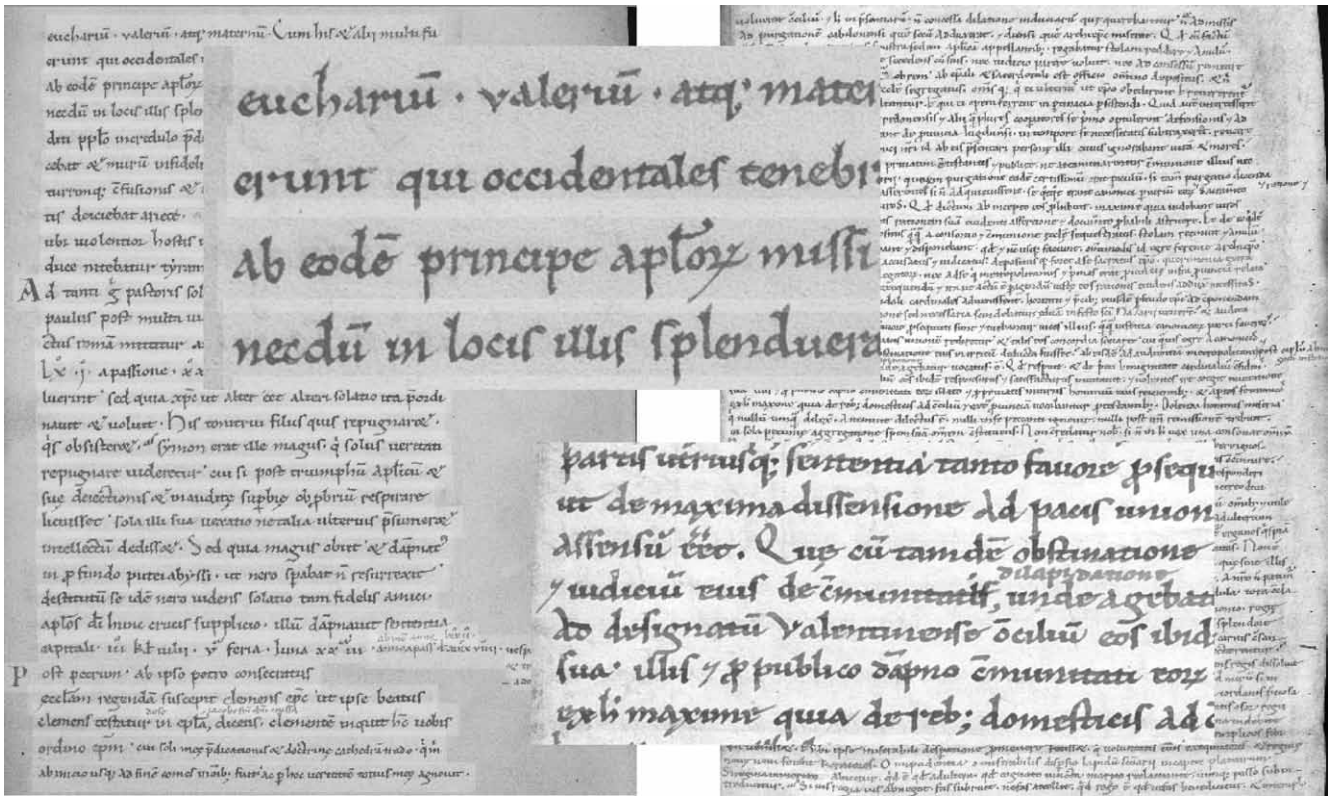


Fig. 4: Hugh of Flavigny, fol. 11r (left) and fol. 144r (right).

exists in the Berlin State Library. It comprises approximately 400 pages in two volumes. This is extensive enough to allow a comparison between the use of abbreviations at an early stage in the codex and how they are used in another passage which comes later in the same book. Fig. 2 shows two pages – the one on the left is from the initial part of the first volume, and the page on the right is from the end of the work. The first page looks relatively unobtrusive, whereas the latter page is much more complex, since the writer is trying to exploit the full range of the parchment. To be precise, folio 11r comprises 1,497 glyphs in 38 text lines, and folio 144r consists of 4,872 glyphs in 60 lines.

While 46 different abbreviations are used on fol. 11r, there are 62 different abbreviations on fol. 144r. This results from the fact that the text on fol. 144r is 3.25 times longer in terms of the number of glyphs than the text on fol. 11r. In other words, 7.5% of the glyphs on fol. 11r are abbreviations, while there is a similar amount of abbreviations on fol. 144r (7%).

Four types of abbreviations can be distinguished:

a) at the beginning of a word

b) at the end of a word

c) at an intermediate position in the word

d) discrete from other glyphs (abbreviations that form a whole word in their own right).

The following holds for fol. 11r: (a) 26%, (b) 38%, (c) 16%, and (d) 21%. This shows that the frequencies are quite similar to fol. 144r: (a) 28%, (b) 37%, (c) 13%, and (d) 22%. Basically, a writer might employ abbreviations in each of the four different categories to an arbitrary degree. It can therefore be concluded that both pages together, which are placed more than two hundred and fifty pages apart, provide a first indication that Hugh of Flavigny used abbreviations consistently.

Both pages have 23 abbreviations in common, which are shown in fig. 5 together with their frequency distribution. This means that 50% of the abbreviations on fol. 11r are common to both pages, and 37% of those on page 144. Of the common abbreviations, five out of the first third have a similar ranking concerning their frequencies, namely *um*, *et*, *em*, *am*, and *per*.

In order to confirm these results, a third page was analyzed which was positioned right in the middle of the other two pages. The distribution of abbreviations differs slightly in comparison to the other two examples, but the tendency is the same (the values for fol. 11r and fol. 144r are given in

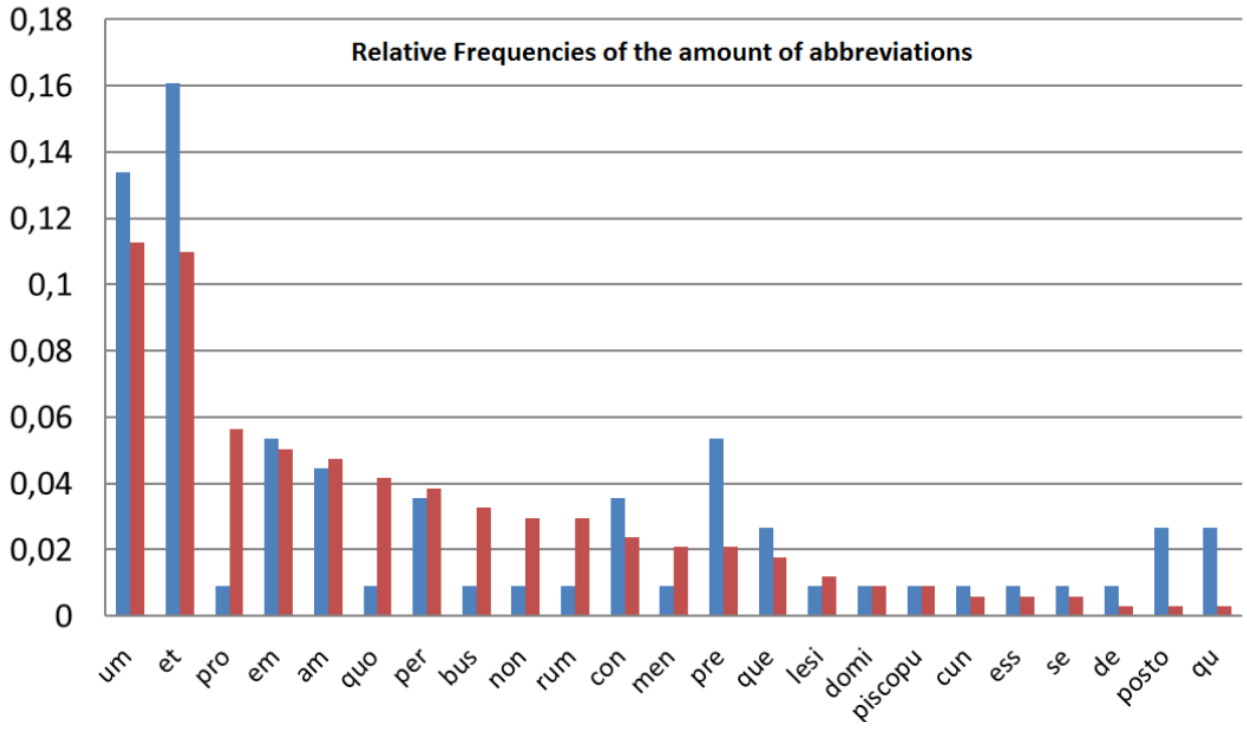


Fig. 5: There are 23 common abbreviations used on fol. 11r (blue) and fol. 144r (red). Their relative frequencies are similar as regards some of the most commonly used abbreviations and somewhat different when it comes to other abbreviations.

parentheses): The following holds for fol. 43r: (a) 13% (26%, 28%), (b) 52% (38%, 37%), (c) 17% (16%, 13%), and (d) 17% (21%, 22%). In other words, most abbreviations are found at the end of words in the third document sample – the same as on the other two pages. A similar amount of abbreviations appear in intermediate positions, but there are slightly fewer at the beginning of words as well as isolated abbreviations on page 43. There are a total 9.8% of abbreviations on this page with 2,232 glyphs (7.5% on fol. 11r, and 7% on fol. 144r). Having analyzed the robustness of the abbreviation criterion in this way, future work will have to reveal how stable the use of abbreviations is for other writers as well.

We could also ask how many abbreviations are found within documents written by other writers. In order to shed light on this question, we examined another document: a charter by Emperor Charles IV issued in the year 1361.[10] This contains 1,128 glyphs with an abbreviations ratio of just less than 5%, thus showing a less frequent use of abbreviations. 29% of the abbreviations are found at the beginning of words, while 38% are found at the end. This appears to be similar to the use of abbreviations by Hugh of Flavigny. There is, however, a clear difference concerning isolated abbreviations, namely that there are only around 5% in the imperial document, but an average of 20% in the chronicle by Hugh of Flavigny. Another difference concerns abbreviations that are situated in the middle of a word: there are around 29% in the imperial document, and approximately half as many in the other case. Although this demonstrates a clear difference in the employment of abbreviations, it is only regarded as a first indicator for our assumption that abbreviations may have been used differently by different writers.

7. Summary

There is a striking use of abbreviations in medieval Latin manuscripts. It is therefore of interest to examine exactly how abbreviations were used by different writing hands, and whether they can be isolated as a distinguishing feature. The *Diptychon* software system we use has been developed for the precise extraction of glyph images and the investigation of abbreviations in medieval documents. Initial indications that our research criterion is stable for single scribes, but that there might be clear differences with regard to different writers have been found in a dataset containing 9,729 glyph images. These images were extracted from the documents of just two scribes. Evidently, results will have to be compared for many more different writers in future work in order to substantiate our hypothesis.

[10] Bayerisches Staatsarchiv Nürnberg, RU Nürnberg 1086.

ACKNOWLEDGEMENTS

We gratefully acknowledge the funding support provided by the German Reseach Foundation (DFG) under Grant Numbers GO 2023/4-1/LA 3066/4-2 and LA 3007/1-1 under the project name *Diptychon*. We are thankful to Jan-Hendrik Worch for supporting the software development.

REFERENCES

Abbreviationes™, ed. by Olaf Pluta, (http://www.ruhr-uni-bochum.de/philosophy/projects/abbreviationes/index.html).

Bayerisches Staatsarchiv Nürnberg, *RU Nürnberg 1086*.

Bischoff, B. (2009), *Paläographie des römischen Altertums und des abendländischen Mittelalters,* 4. Aufl. (Berlin: Erich Schmidt; Grundlagen der Germanistik 24), 202–223.

Cappelli, A. (1929), *Lexicon Abbreviaturarum. Dizionario di Abbreviature Latine ed Italiane,* 4th edition (Milan: Hoepli; several reprints).

Gottfried, B., Wegner, M., and Lawo, M. (2013), 'Diptychon: A transcription assistant system for the separation of glyphs in medieval manuscript texts', in E. Angelopoulou et al., *20th SAOT Workshop on Automatic Pattern Recognition and Historical Document Analysis, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. June 14–15, 2013* (The Digital Library of the Göttingen Academy of Sciences and Humanities).

——, ——, and —— (2014), 'Towards the interactive transcription of handwritings: anytime anywhere document analysis', *International Journal on Document Analysis and Recognition* (IJDAR), (DOI 10.1007/s10032-014-0234-7), 1-15

Lawo, M. (2010), *Studien zu Hugo von Flavigny* (Hannover: Verlag Hahnsche Buchhandlung; Monumenta Germaniae Historica. Schriften, 61).

Romero, V., Toselli, A. H., Rodriguez, L., Vidal, E. (2007), 'Computer-assisted transcription for ancient text images', in M. S. Kamel and A. C. Campilho (eds.), *Image analysis and recognition (ICIAR). 4th international conference*, *(Montreal, Canada), Lecture Notes in Computer Science*, vol. 4633: 1182–1193.

Serrano, N., Gimnez, A., Civera, J., Sanchis, A., Juan, A. (2013), 'Interactive handwriting recognition with limited user effort', *International Journal on Document Analysis and Recognition*, 1–13.

Staatsbibliothek zu Berlin, Preußischer Kulturbesitz, Ms. Phill. 1870, fol. 11r and fol. 144r.

Wüthrich, M., Liwicki, M., Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., Stolz, M. (2009), 'Language model integration for the recognition of handwritten medieval documents', in *10th International Conference on Document Analysis and Recognition (ICDAR)* (Piscataway, NJ: IEEE Computer Society), 211–215.

Article

Synthetic-based Validation of Segmentation of Handwritten Arabic Words

Laslo Dinges, Ayoub Al-Hamadi, Moftah Elzobi, and Sherif El-Etriby | Magdeburg – Shebeen El-Kom

Abstract

Suitable and comprehensive databases are crucial for training and validation of various document analysis methods. However, those databases are limited, and their ground truth is often not sufficient for methods like word segmentation. In this article, we propose an active-shape model-based approach for Arabic handwriting synthesis that enables parameterization of several real-world Arabic handwriting distortions, e.g., skew, slant, variations in letter size inside a word, and length of their connections. Furthermore, we develop a technique to render such handwriting so that it reflects the features of real writing instruments. Finally, we test and validate a segmentation method on a large dataset of synthetic samples as well as real-world samples. The initial results are promising, suggesting the synthetic approach will be useful in various handwriting recognition areas.

1. Introduction

Document analysis methods, including automatic document classification, layout analysis, segmentation or text recognition, can be very helpful when dealing with huge amounts of data such as the Ottoman Archives. To test and compare these methods, comprehensive databases are essential. However, the problem of lacking satisfactory handwriting databases is very obvious for Arabic handwriting. Two main word databases that are free are available: the IFN/ENIT (Pechwitz et al. 2002), which contains the names of Tunisian towns, and our IESK-ArDB (Elzobi, et al. 2012), which contains international town names and common terms. Since the most expensive part of Ground Truthing databases is due to the information required for validating segmentation methods, the number of available samples that include such information is still limited. To bypass this problem, it would be very helpful if databases that are adapted to general or specific needs of a concrete document analysis task could be produced automatically. This includes precise manipulation of the degree of challenge, which is hardly possible with traditional databases.

Approaches to text synthesis can be classified into two main categories. The first category includes all research that generates synthetic text by perturbing real text samples, which can be complete units such as text lines or words, but also glyphs, e.g., groups of letters (Guyon 1996), Varga and Bunke 2003; Thomas, Rusu, and Govindaraju 2009; Miyao and Maruyama 2006). The second category encompasses all approaches that are based on deformable models in synthesizing text. Words are composed by randomly permuting and linking glyphs to obtain an unlimited number of synthetic samples (Cheung et al. 1998; Al-Zubi 2004; Shi, Gunn, and Damper 2003). Only a few approaches are known for Arabic. The first approach to Arabic handwriting synthesis uses character images from two writers in the IFN/ENIT database and composes them into words (Elarian, Al-Muhsateb, and Ghouti 2011). To achieve smooth word shapes, only letters are connected that have a similar angle and width (at their juncture), but this limits the number of different samples that can be generated for each word. Shortly thereafter, a second approach was proposed, which composes letter trajectories that are acquired by tablets (Saabni and El-Sana 2012). This approach allows multiple kinds of handwriting of a given Piece of Arabic Word (PAW) to be generated, but without any diacritical marks like dots, which are, however, distinctive features of Arabic letters.

In order to enable intuitive generation of synthetic Arabic text databases, we have developed a user interface (UI) that allows specific features to be manipulated, such as the slant, how well-written the characters are and the size of the *kashida*, which is the connection between two letters. Unlike connections in Latin-based scripts, a *kashida* is often wider than a letter itself, but sometimes vanishes entirely.

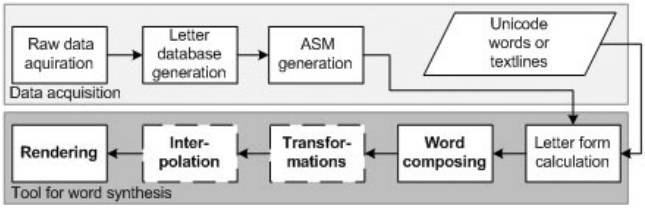


Fig. 1: Overview of the proposed synthesis approach (stippled boxes indicate optional processes).

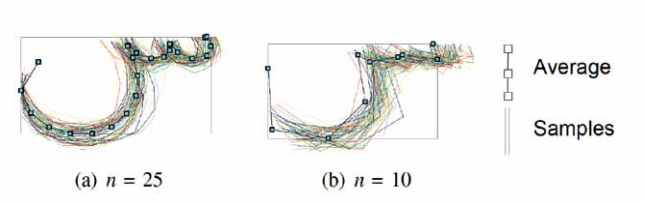


Fig. 2: Normalized samples and average shape of the letter *sin* in an isolated form using a) 25 landmarks (sufficient for all classes) and b) 10 landmarks (insufficient for most classes).

In the preview section of the UI, one can enter Arabic Unicode or type the Latin names of letters. Thereafter, a few synthesis examples can be generated in order to examine the effect of the current settings. Moreover, interpolation settings as well as predefined pigmentation textures (that are used to render the synthetic images) can be employed to achieve the desired outcome.

The UI enables quick generation of databases that focus on specific features. In that way, it is possible to investigate how robust methods are against different writings styles, carefully or roughly written text, and primitive features such as the slant and skew of the writing. We use these options in section 3 to validate a segmentation technique that we proposed in Elzobi et al. 2012.

2. Methodology

In this section, we describe our approach to synthesizing Arabic handwriting from Unicode. Note that the synthesized samples have either an image (png, bmp) or vector graphic format (pdf, eps or svg), but vector graphics will only be used for illustration here.

Fig. 1 shows the main modules of our synthesis approach. The first module includes the concept of data acquisition, mainly generating and processing raw data. The second module uses the acquired data as well as requests from the user to synthesize Arabic handwriting. This includes word composition, directed manipulation of features (like word slant) and the rendering technique.

The basic idea of handwritten Arabic word synthesis from Unicode sequences is to select polygonal samples with correct context-sensitive shapes and subsequently connect them to obtain Pieces of Arabic Words (PAWs), words and complete sentences. These are then modified by affine transformations, smoothed by B-spline interpolation and composed to text, as shown in fig. 3. Finally, the writing is rendered and saved, including ground truth (xml files). In this way, our system produces offline pseudo-handwritten samples with variations in shape and texture. At the end of this section, we briefly discuss word segmentation as an example of image-processing methods that can be validated using synthetic handwriting.

2.1 Data acquisition

Even though the synthesis module is built to create new letter and word samples automatically, suitable and sufficient samples of handwritten Arabic letters are needed in the first place. These samples are crucial to achieve a natural and comprehensive outcome. A total of 28,046 online samples from multiple writers were created to compute Active Shape Models (ASMs) for over one hundred letter classes, since experiments with offline samples led to poor results for most letters that have complex shapes (Dinges, Al-Hamadi, and Elzobi 2013). ASMs are statistically deformable models. To build an ASM, we defined a fixed number of *n* landmarks for all samples in the same class (see fig. 2).

The ASMs contain information about the position of these landmarks and how these positions statistically vary in relation to themselves and all other landmarks. ASMs reflect the main characteristics of the input samples used, which can be specifically manipulated by their eigenvectors. ASMs are used to generate unique letter representations for each synthesis. Since we intended to synthesize offline handwriting, we chose to use online pens instead of tablets, since the former can be used as ordinary biros and do not distort the handwriting style.

2.2 Letter-form calculation

In this and the following section, we describe how to build the trajectory of an Arabic handwriting sample from Unicode. Unicode strings represent every letter as a plain number. Although there are special Unicode signatures for the isolated, end, middle and beginning form (final, medial and initial), only the general form is used for normal Arabic text. Letter forms have a strong influence on letter shape, however, so they have to be determined first. Six letters (ر ا

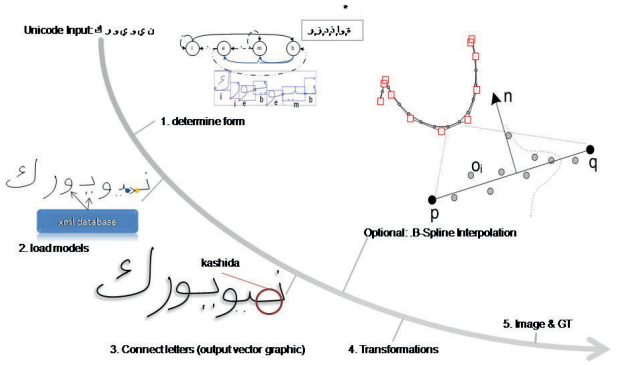


Fig. 3: Scheme shows how Unicode input is turned into a synthetic handwriting image.

د ز ذ و) can only assume an isolated or end form, while all the others also have a middle or initial form — take the letter *ayn*, for example (علعلع ع). Letters that have only two forms split an Arabic word into PAWs which consist of one or more letters. If the first letter comes from the set (ا ر د ز ذ و), it has an isolated form or an initial form. The forms of all the other letters are computed as in fig. 3. If any letter is followed by a space token, it initiates a new word and must assume an isolated or end form.

2.3 Word composition

After all the letter classes have been defined by their names and forms, the corresponding ASMs are loaded. These are used to generate unique shapes for all occurrences of a letter class in order to avoid piecemeal identical syntheses (original samples can be used optionally, though). In order to compose words from these letter shapes, each letter has to be connected to its predecessor in an end or medial form, as illustrated in fig. 3.

Since our samples are currently limited to the 28 regular letters and *tamarbuta* (ة), we substituted special characters, such as *alif* with *hamza* above (أ), with their regular form, *alif* (ا), before starting the synthesis.

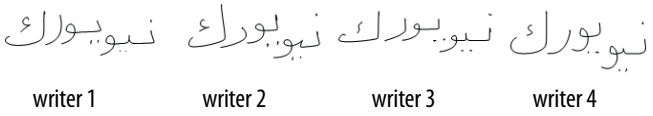


Fig. 4: Syntheses of the Arabic word for 'New York' using ASMs of different writers (vector graphic format).

The height of a PAW in relation to the lower baseline depends on the letter classes the PAW is composed of. Thus, we extracted the average distance (and standard deviation) between the baseline and the center of a letter from a manually created ground truth of our (real) word database IESK-arDB (Elzobi, et al. 2012). We corrected the y-position of all PAWs according to the average heights of all the letters within a PAW.

Finally, the horizontal gap between a PAW and its predecessor has to be defined. Therefore, a user-dependent parameter is applied that may be negative in order to simulate overlapping PAWs. 'Overlapping' means that a letter is above or beneath another one, but its trajectory should not intersect it; if two overlapping PAWs do intersect, the gap is increased by 25% of the average letter width until the intersection is solved. Some examples of such results are shown in fig. 4.

2.4 Transformations

ASMs already contain variations in slant, width and connection size. Nevertheless, these variations are limited by the samples used. In order to increase and control these variations, affine transformations are used that allow optional manipulations of letter and PAW shapes (see fig. 5).

The user interface (UI) of our synthesis tool allows the average μ and standard deviation σ of a Gaussian distribution to be set for all affine transformations. Global variations, in particular, can be achieved this way (e.g., slant or skew). The affine transformations used are scaling, translation, sharing and rotation. The influence these have on the resulting word

stretching	slant	size	skew	position	connection

Fig. 5: Effects of affine transformations and *kashida* cutting (image format).

Fig. 6: Example of syntheses of the PAW في using different *kashida* lengths. The example on the left shows an extremely reduced *kashida* (actually a 'negative' *kashida* that blends two letters). The middle example is created using a moderately reduced *kashida*. The *kashida* almost reaches its maximum size in the right-hand example.

image is shown in fig. 5. Stretching is performed by scaling the components of each letter point. The word slant can be set by sharing the word, the skew of a PAW by rotation. The size of the complete word or single letters can be adjusted by equal scaling, which can be used to control the resolution of the synthesized word images or to increase variation in the letter size.

As we found letter connections of the acquired samples of certain writers to be excessively long, we allow up to 25% of the letter points to be deleted to simulate stretched or missed connections, which often occur in real Arabic handwriting.

In Arabic handwriting, some characters, such as *ya* (ي), are written beneath rather than beside their predecessors. As a result, they resemble a single character, which impedes segmentation and recognition tasks. This effect can be simulated by a strong reduction of the *kashida*, as shown in fig. 6. However, this feature has not yet been completely implemented, since a list of all pairs of letters that typically show this behavior needs to be created first.

2.5 Rendering technique

As described in Dinges, Al-Hamadi, and Elzobi 2013, we compute the coordinates of all the pixels that will be influenced by a virtual pen and then apply an artificial texture to them. To prepare high-resolution syntheses, B-spline interpolation can be used before painting, as shown in fig. 7 (a–b).

To allow a more accurate simulation of writing instruments, such as biros, pencils, fountains pens or quills, we extended our rendering technique. This mainly meant implementing two features: writing speed and the shape of the top of the writing instrument, subsequently called Pen Shape.

Large lines or bows, as with the left part of *sin* (س), are usually written faster than more complex structures. A high writing speed often causes a lack of pigmentation that leads to brighter or dappled lines. However, there is no need to reconstruct writing speed, for the online letter samples we employed already contain this information. We used the normalized writing speed to reduce the pigmentation of our syntheses up to 20% of the average value, as shown in fig. 7 (c).

If Pen Shape is modeled as ellipsoid, the line width depends not only on Pen Shape size but also on the Pen Shape and line angle. We used an angle of 45° for Pen Shape, changing it continuously with a maximum deviation of ±15°. According to the features of Pen Shape, the contact with the paper and

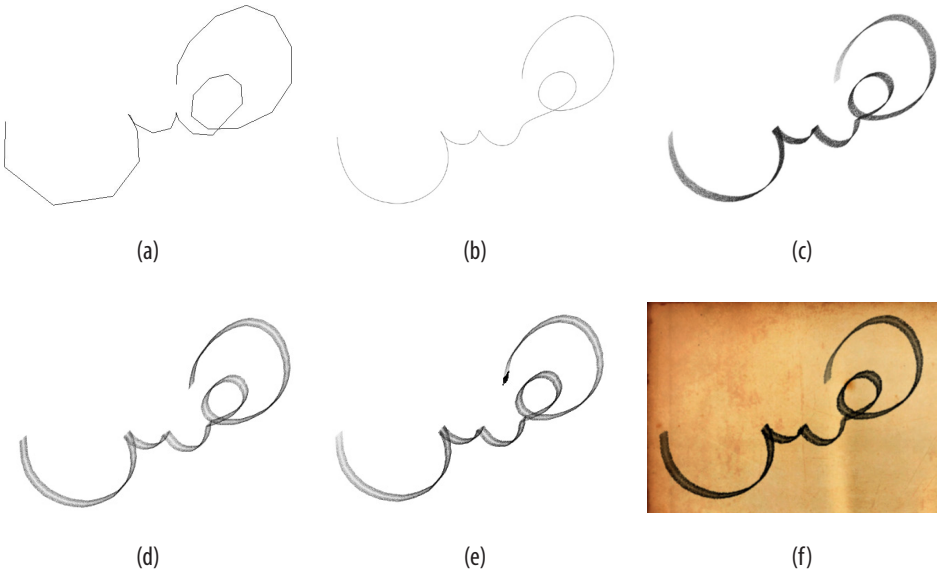
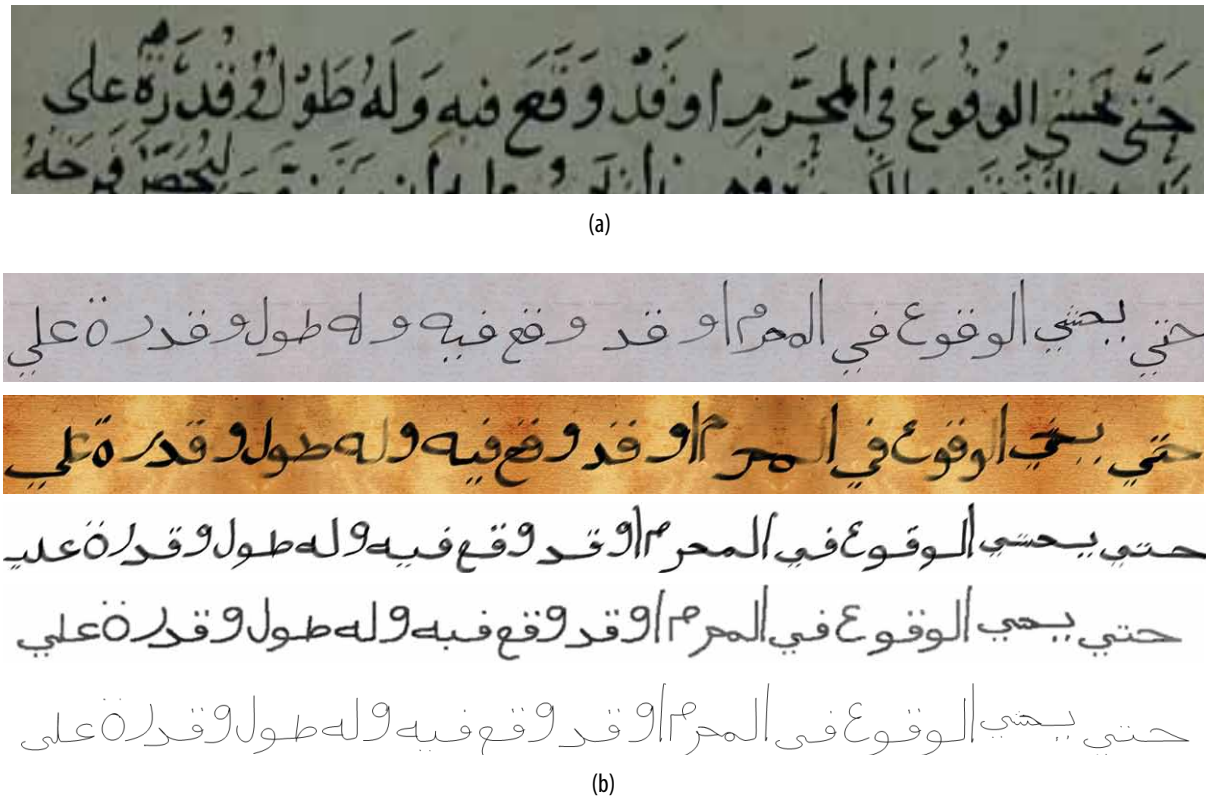


Fig. 7: From polygons to images of synthetic handwriting: (a) simple lines, (b) result after applying B-spline interpolation. Images (c) to (f) show the proposed painting technique using an ellipsoid Pen Shape to simulate a quill-like writing instrument: (c) influence of pen speed, (d) influence of Pen Shape, (e) combination of both, (f) result of (e) combined with an artificial ink texture and transparently drawn over a texture of old paper.



(a)

(b)

Fig. 8: a) First line 'حتي يحشي الوقوع في المحرم او قد وقع فيه وله طول وقدرة علي' of a historical document that is taken from the IESK-arDB manuscript database (file im5_1/gt5_1); (b) five syntheses. The first three images are rendered using an improved rendering technique, where the first two are drawn on a degraded background but the third is on a perfect white background. The fourth image was created using the technique described in Dinges, Al-Hamadi and Elzobi 2013. The last one is a vector graphic (PDF).

the resulting pigmentation can vary. This is simulated by a function that defines the pigmentation potential for each position of Pen Shape. An emphasized example is shown in fig. 7 (d), while fig. 7 (e) is also influenced by writing speed. We created fig. 7 (f) by combining our original rendering technique with the one based on Pen Shape and writing speed as well as using a non-uniform background. We therefore applied a transparent texture to all the pixels in fig. 7 (e) that do not belong to the background texture. Small, random irregularity is simulated this way.

In the examples shown, a black color is used, as most documents are written with dark ink. Pigmentation intensity is implemented as transparency, which allows simulating pigment accumulation at crossing lines or for a textured background. However, opaque or semi-opaque ink can also be simulated. This could be interesting in the context of historical documents, since important passages are often highlighted using red ink. Another important feature of historical documents is degradation of the material on which they have been written. Currently, we simply use images of real paper or parchment as background textures, which are scaled or tiled if they are smaller than the synthesized document.

2.6 Sample method to test applicability: segmentation of Arabic words

Since the data synthesis module is built to ease the development and validation of methods that are related to document analysis, it is not only of interest whether syntheses look realistic or not. In fact, it is crucial how image-processing methods behave when fed with synthetic instead of real data. This is investigated in the following section, where a method that segments handwritten Arabic words into letters is validated using such data. We chose word segmentation as the example due to its sensitivity to character shapes as well as global features, such as overlapping PAWs and varying *kashida* length.

One of the earliest segmentation-based approaches suggested for the recognition of handwritten Arabic text is the one proposed by Almuallim and Yamaguchi 1987, but no segmentation results were reported. Xiu et al. proposed

a probabilistic segmentation model in which tentative, contour-based over-segmentation was first performed on the text image. As a result, a set of what they called graphemes was produced (Xiu et al. 2006). The approach differentiated between three types of graphemes. The confidence of each character was calculated according to the probabilistic model while respecting other factors, such as recognition output, geometric confidence and logical constraint. The authors experimented with the proposed methodology on five different test sets, achieving a success rate of 59.2%.

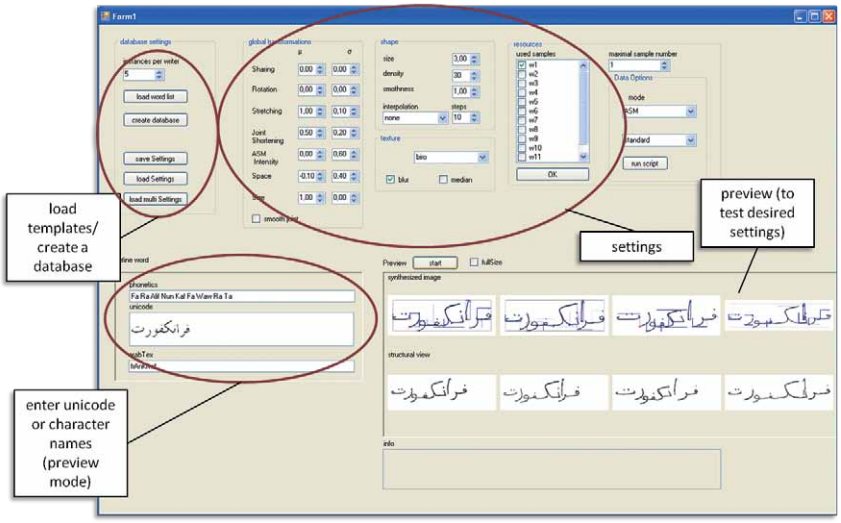


Fig. 9: a) Screen shot of the user interface, which enables handwriting to be synthesized; b) ground truth that is used to validate segmentation.

The segmentation method used for the following experiments is described in Elzobi et al. 2012. It is based on topological features and a set of rules that reduce all candidates to a final set of points that divide two neighboring letters. Unlike other approaches, candidates are not minima that indicate the middle of a *kashida*, but typically the following branch point.

3. Experiments using synthetic databases

The synthetic samples, which are created by the proposed approach, are meant to be used as training or testing data for different document analysis methods. To investigate whether these syntheses can be used instead of real samples, we created synthetic samples (png images + ground truth) of all words in our IESK-arDB database to build IESK-arDB-Syn. Then, we applied a segmentation technique on both the original and the synthesized database. We found that the detected error rates are comparable, as shown in Table 1. This proves at least that the proposed synthesis method is capable of reflecting those features of real Arabic words that are critical for segmentation. Furthermore, we investigated robustness of the segmentation method against the influence of particular features such as slant or skew. Modified versions of IESK-arDB-Syn were created for each experiment, where only the investigated feature differed for samples by the same writer (except the last experiment, which was done using the original IESK-arDB-Syn). We used cross-validation for all the following experiments.

Main experiment: Since ASMs for letters are currently only available from four writers (due to limited resources), multiple ways of writing a word were created for each writer to get sufficient samples. In order to achieve variations close to those of the 12 writers of the IESK-arDB, some optional features, such as slant or *kashida* length modification, have

Table 1: Experimental results.

	IESK arDB		IESKarDB-Syn	
Measure	μ	σ	μ	σ
Error per word	1.67	0.13	1.74	0.024
Error per letter	0.35	0.026	0.34	0.0045
Over-segmentation (per word)	0.79	0.097	0.86	0.0066
Under-segmentation (per word)	0.87	0.071	0.88	0.019
Perfect segmentation (per word)	0.17	0.0019	0.13	0.0067

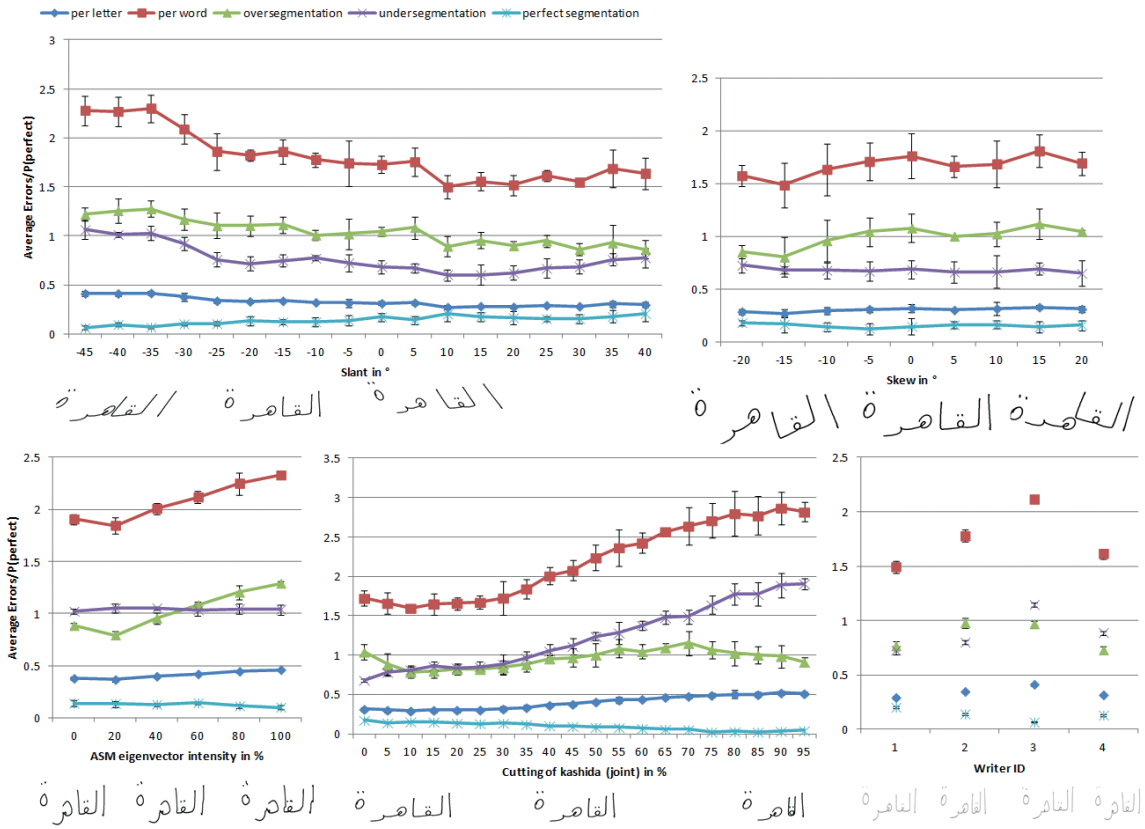


Fig. 10: Influence of different handwriting features on accuracy of the segmentation method.

been parameterized by the user interface. This is shown in fig. 9.

As the main experiment, we segment all samples of the IESK-arDB database (2540) and the IESK-arDB-Syn database (9000) into letters and compare the results. However, a manual validation of the segmentation results of thousands of words is barely feasible. Hence, the ground truths of both databases are used to perform automatic validation. Each point where two letters are connected is defined as a dividing point. We experimentally found a suitable range that defines the area around each dividing point in which exactly one automatic segmentation proposal is expected (shown in fig. 9 b). The validation method measures *under-segmentation* within a word for each area without any segmentation proposal and *over-segmentation* for each additional segmentation proposal and proposals outside the defined areas.

According to the automatic cross-validation, the segmentation method used causes an average of 1.7 errors per word, and the standard deviation σ is 0.13 when applied to the IESK-arDB. Nearly the same accuracy is measured for IESK-arDB-Syn, which shows that validations based on synthetic samples indeed enable one to predict how well a method may perform in real-world situations. This also enables automatic optimization of parameters and comparison of different methods. Nevertheless, the measured perfect segmentation rate is less for IESK-arDB-Syn and the *over-segmentation* is 7% higher. This could be caused by both well and badly written samples occurring in the IEKS-arDB, a property that is not yet mimicked sufficiently by our synthesis approach. Furthermore, the simulation of letter joints has to be improved for some cases to avoid accidental structures that are excessively hard to segment.

The results of the following experiments are shown in fig. 10.

Slant: Since the segmentation method only uses segments at rows with exactly one foreground pixel, extreme slants can cause segmentation errors due to overlapping ascenders. The experiment shows that even a slant of about 20° improves the segmentation results, especially if the *kashidas* between the letters are relatively long. Hence slant correction is a useful preprocessing step for our segmentation method.

Skew: Although skew correction is a common preprocessing step, this experiment shows that this segmentation method is robust against a skew of +/- 20°.

Kashida length: In Arabic handwriting, the length of *kashidas* is highly variable. The experiment shows that a valid segmentation is especially difficult for very small *kashidas*, since most structures that indicate a potential dividing point are hidden or have vanished in such cases. In contrast to slant and skew variation, this problem cannot be solved using a simple preprocessing technique.

ASM eigenvector intensity: The intensity of the used eigenvectors defines the similarity of a computed letter shape to the average shape, where maximal intensity often causes deformed shapes that are hard to classify. The experiment confirms that eigenvector intensity is also proportional to segmentation error. However, even strongly deformed letter shapes are not as critical as the reduction of *kashida* length.

Writer: As one can see, the segmentation method is more sensitive to the writer-dependent style than to the precision of the writer. The best performance was achieved for writer #1, whose letters are well defined and have long *kashidas*.

4. Conclusion

We have presented an efficient approach for generating pseudo-handwritten Arabic words or text lines from Unicode (including diacritical marks). Online Sample and Active Shape model-based glyphs from multiple writers as well as affine transformations allow the researcher to generate various images for a given Unicode string to cover the variability of human handwriting. Features such as the slant of a letter can be controlled manually if desired. To increase variability and realism, a rendering technique that simulates the physical attributes of various writing tools has been developed.

In order to validate how accurately the syntheses reflect features of real handwriting, we compared the results of a segmentation method applied to real and synthetic samples. In addition, the use of specifically manipulable features was investigated. Therefore, we experimentally detected the influence of slant, skew and writing style on segmentation accuracy. Due to the comprehensive ground truth, special preprocessing methods, such as detection and classification of diacritical marks, can be trained, optimized and validated. Such methods can be used to improve different approaches to word recognition, segmentation and spotting.

In our future work, we intend to reduce the amount of synthesized words by means of clustering techniques, such as affinity propagation, to avoid redundant samples. Furthermore, samples by additional writers must be acquired and additional character classes, such as digits or common ligatures, need to be included to improve the quality of syntheses. We will also extend our approach to allow the generation of whole pages of text. This will help synthesize compact but representative databases for training and testing of preprocessing, segmentation and recognition methods for document analysis tasks.

The handwriting synthesis and segmentation approaches proposed here are just modules that are part of a larger project. There are huge archives written in Arabic. We plan to build a system that can segment these documents into characters. Then feature extraction, a priori information as well as a vocabulary of valid Arabic words will be used to recognize the handwriting. The vocabulary should be dynamically defined by the user to check if a document contains keywords from various fields (religion, history, medicine, etc.) or names of specific people or regions. Using explicit segmentation, the recognition process can be separated into a static and a dynamic part that uses contextual information provided by the user. All image-processing methods are assigned to the static part, which only needs to be done once. The proposed synthesis approach will be used to validate this system.

ACKNOWLEDGMENTS

Research project funded by King Abdulaziz City for Science and Technology, (KACST). Project code: 13-INF604-10.

REFERENCES

Almuallim, H., and S. Yamaguchi (1987), 'A method of recognition of Arabic cursive handwriting.' *IEEE Trans. Pattern Anal. Mach. Intell.* (IEEE Computer Society), 9: 715–722.

Al-Zubi, Stephan (2004), 'Active Shape Structural Model', in *Tech. rep*. *Otto-von-Guericke University of Magdeburg*.

Cheung, Kwok-Wai (Student Member), Chin, Roland T., Dit-Yan Yeung, and Chin, T. (1998), 'A Bayesian Framework for Deformable Pattern Recognition With Application to Handwritten Character Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20: 1382–1388.

Dinges, Laslo, Al-Hamadi, Ayoub, and Elzobi, Moftah (2013), 'An Approach for Arabic Handwriting Synthesis based on Active Shape Models', in *12th International Conference on Document Analysis and Recognition (ICDAR),* 1260–1264.

Elarian, Y. S., Al-Muhsateb, H. A., and Ghouti, L. M., 'Arabic Handwriting Synthesis', *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 4.4: 131–143.

Elzobi, Moftah, Ayoub Al-Hamadi, Zaher Al Aghbari, and Dinges, Laslo (2012), 'IESK-ArDB: a database for handwritten Arabic and optimized topological segmentation approach.' In *International Journal on Document Analysis and Recognition*, 16.3: 295–308.

Guyon, Isabelle (1996), 'Handwriting Synthesis From Handwritten Glyphs', *Proceedings of the Fifth International Workshop on Frontiers of Handwriting Recognition*, 309–312.

Miyao, Hidetoshi, and Maruyama, Minoru (2006), 'Virtual Example Synthesis Based on PCA for Off-Line Handwritten Character Recognition', in *Document Analysis Systems*, 96–105.

Pechwitz, Mario, Maddouri, Samia S., Märgner, Volker, Ellouze, Nouredine, and Amiri, Hamid (2002), 'IFN/ENIT – database of handwritten Arabic words', *Proc. of CIFED 2002*, 129–136.

Saabni, R. M., and El-Sana, J. A. (2012), 'Comprehensive synthetic Arabic database for on/off-line script recognition research', *International Journal on Document Analysis and Recognition,* 16:285–294.

Shi, Daming, Gunn, Steve R., and Damper, Robert I. (2003), 'Handwritten Chinese Radical Recognition Using Nonlinear Active Shape Models', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 25: 277–280.

Thomas, Achint Oommen, Rusu, Amalia, and Govindaraju, Venu (2009), 'Synthetic handwritten CAPTCHAs', *Pattern Recognition* (Elsevier Science Inc.), 42: 3365–3373.

Varga, T., and Bunke, H. (2003), 'Generation of synthetic training data for an HMM-based handwriting recognition system', *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings,* vol. 1, 618–622.

Xiu, Pingping, Peng, Liangrui, Ding, Xiaoqing, and Wang, Hua (2006), 'Offline Handwritten Arabic Character Segmentation with Probabilistic Model', in *Document Analysis Systems VII 7th International Workshop*, ed. by H.orst Bunke and A. Lawrence Spitz (Springer), 402–412.

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/.

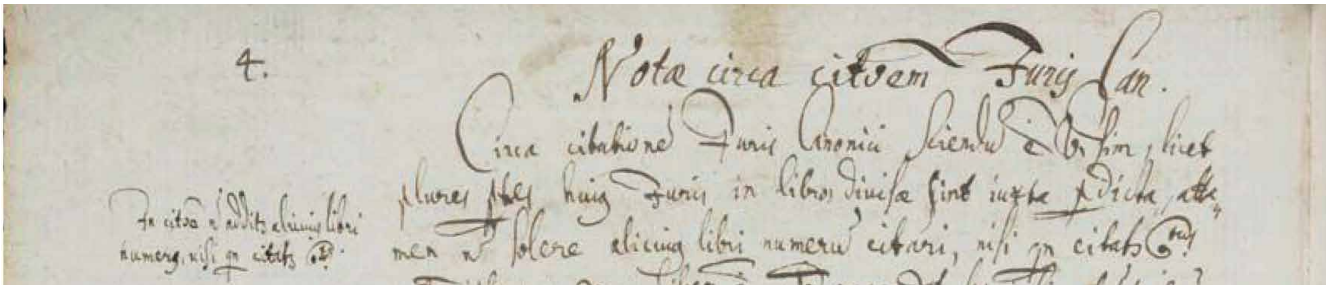
[4] http://www.manuscripta-mediaevalia.de.

[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.

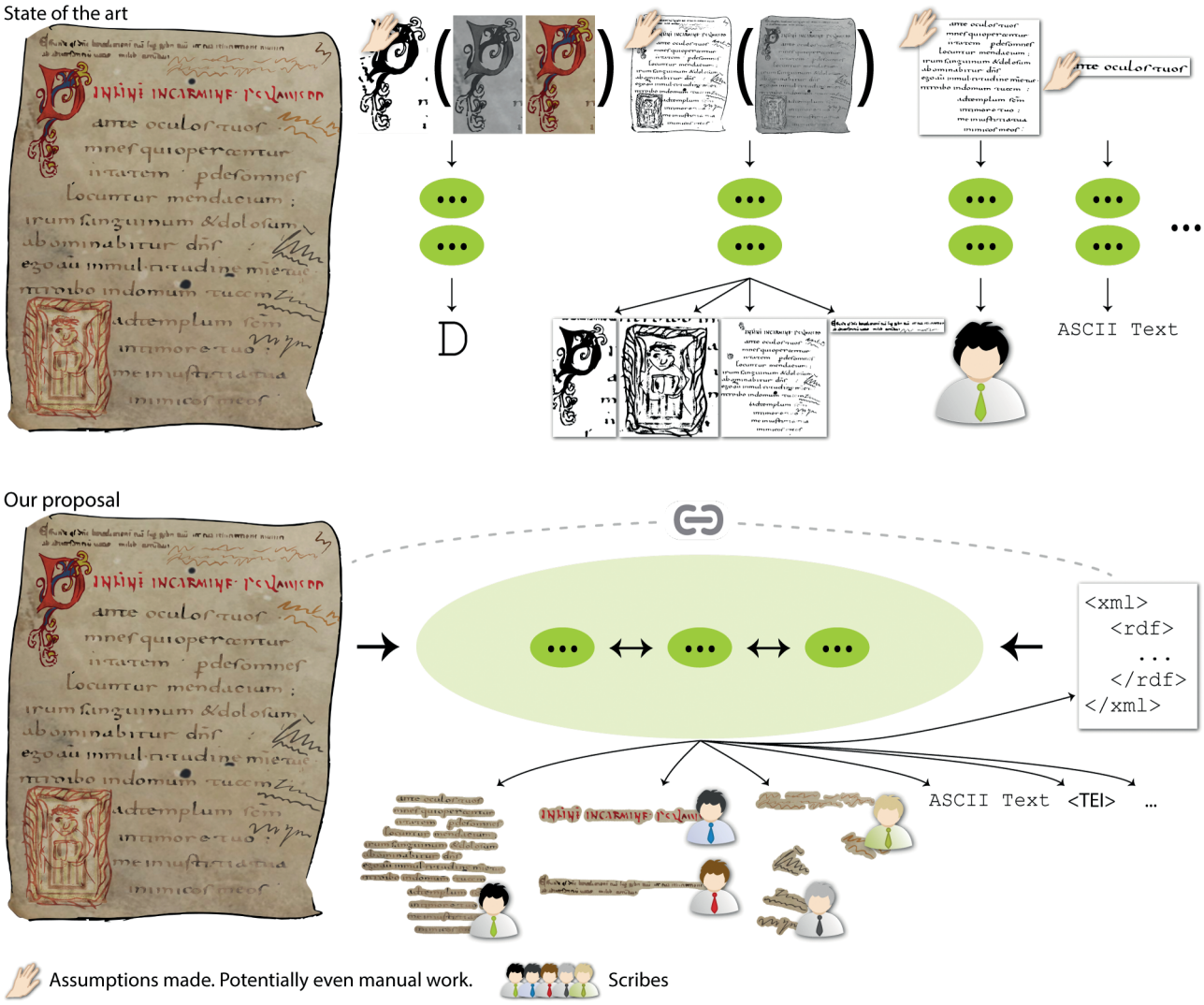


Fig. 2: Comparison between state-of-the-art approaches (top) and the novel proposition in the HisDoc 2.0 project. It is common practice to treat tasks independently, such as processing initials and recognizing the letter they represent, layout analysis, scribe identification, and handwriting recognition. 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.

Article

In the Shadow of Goitein: Text Mining the Cairo Genizah

Christopher Stokoe, Gabriele Ferrario, and Ben Outhwaite | Cambridge

Abstract

The widespread digitization of manuscripts has brought about an era of unprecedented access to a range of important historical collections. However, the lack of substantive metadata associated with these online digital collections represents a significant barrier to those wishing to navigate them in order to identify manuscripts relevant to a particular research question or theme. We propose a novel solution to cataloguing based around text mining published editions, commentaries and other secondary literature in order to automatically generate a rich searchable electronic catalogue. This research explores a range of techniques from the fields of Information Retrieval (term-weighted vocabularies), Natural Language Processing (named entity recognition) and Text Analysis (topic models). Our initial results demonstrate the potential for these approaches to produce significant volumes of descriptive metadata which, when evaluated in the context of retrieval effectiveness, provide suitable evidence on which to perform analysis and make discoveries. A search engine which recommends manuscripts based on the contents of our automatically derived catalogue achieves a Precision @ 10 of 0.54, which significantly beats a baseline strategy of random selection.

1. Introduction

The Taylor-Schechter Genizah Collection at Cambridge University Library is the single most important collection of medieval Jewish manuscripts in the world.[1] As of June 2013, its 193,000 manuscripts have now been completely digitized and are in the process of being made available online as part of the Cambridge University Digital Library.[2]

Whilst mass digitization has significantly improved access to the collection, it is clear that discovery – the act of directing researchers to a particular manuscript that will answer a given information need – remains a key challenge. This is largely due to the sheer size of the collection coupled with the lack of any substantive metadata describing the content of individual manuscripts. The inability to navigate the collection by content presents a substantial roadblock to the diverse group of scholars looking to exploit this unique source of information about the history of the Mediterranean and Near East. The following catalogue entry describes fragment T-S 24.64 (see fig. 1), which aptly demonstrates the full extent of the problem:

> *T-S 24.64 — letter*
> 55 × 14; 74 lines + marginalia (recto); 7 + 3 lines (verso)
> Paper; 1 Leaf; Torn; Judaeo-Arabic
> A lengthy letter from Ḳalaf b. Isaac to Abraham b. Yiju, middle of 12th c.

Manual efforts to improve the quality of our catalogue by the Genizah Research Unit (GRU) are ongoing, but the scale (in terms of number of manuscript fragments) and complexity (manuscript condition, language constraints and required subject expertise) make the cost of full description, transcription and translation prohibitive. In light of this, we have elected to explore the potential for a technology-based solution.

2. Related work

Recent advances in technology have opened up the possibility of automatically deriving catalogue data from digital images of manuscripts. In particular, the Friedberg Genizah Project[3] has experimented with extracting the shape, size and con-

[1] See Reif and Reif 2002.

[2] Cambridge University Digital Library (2014), URL: http://cudl.lib.cam.ac.uk (accessed on March 14, 2014).

[3] Shweka, Choueka, Wolf, and Dershowitz 2013.

Fig. 1: T-S 24.64, a letter from Ḵalaf b. Isaac to Abraham b. Yiju, middle of 12th c.

dition of a manuscript through the use of image analysis and machine learning. Another recent effort[4] attempted to analyze handwriting features and scribal practice as a means of determining authorship, date and point of origin. Whilst these techniques provide one potential source of descriptive metadata about the physical artifact, they do little to unlock the content of the manuscripts. Most notably, optical character recognition (OCR) accuracy on handwritten manuscripts remains problematic for retrieval purposes.[5] Even if OCR accuracy could be improved, machine translation remains out of reach because of a lack of parallel corpora for the vast range of languages present in the Genizah (in particular, Judeo-Arabic).

3. Methodology

Our approach centers on exploiting the 110 years of scholarship that surrounds the Genizah in order to automatically derive a content-based catalogue from the secondary literature. Extensive written material has been published about the manuscripts in the form of published editions, commentaries and academic papers, many of which include full text translation. Through the use of text mining, we propose taking this rich source of knowledge and using it to produce metadata that will facilitate content-based retrieval of data on the manuscripts.

Our methodology consists of the following steps:

1. Bibliometric analysis in order to identify a suitable corpus of secondary literature.
2. Corpus construction involving OCR and automatic segmentation of the text.
3. Term weighting through association of text from the corpus to a specific fragment.
4. Extraction of named entities and temporal expressions from associated texts.
5. Classification of fragments using topic modeling.

The resulting catalogue has been indexed using a search engine and we have evaluated the output in the context of its potential to provide adequate evidence for resource discovery.

3.1 Bibliometric analysis

For the past 30 years the GRU has been compiling an extensive bibliography[6] by tracking those scholarly works that make use of Genizah fragments as their primary source material. As of December 2013, this bibliography contains over 113,786 citations to 64,265 unique manuscript fragments across 3,643 scholarly works. As part of this research, we have undertaken a detailed analysis of this dataset focusing on the number of fragments cited in each work, co-citation of fragments by author and the co-occurrence of fragments across scholarly works. These measures combine to give us a picture of which authors and works to target in order to maximize our coverage of the collection.

Fig. 2 shows a visualization of the citation information for the fragments in the Taylor-Schechter collection based upon clustering co-occurrence across scholarly works. Note

[4] Levy, Wolf, and Stokes 2013.

[5] Naji and Savoy 2011.

[6] Genizah Research Unit Bibliography (2014), URL: http://cudl.lib.cam.ac.uk/bibliographies/genizah (accessed on March 14, 2014).

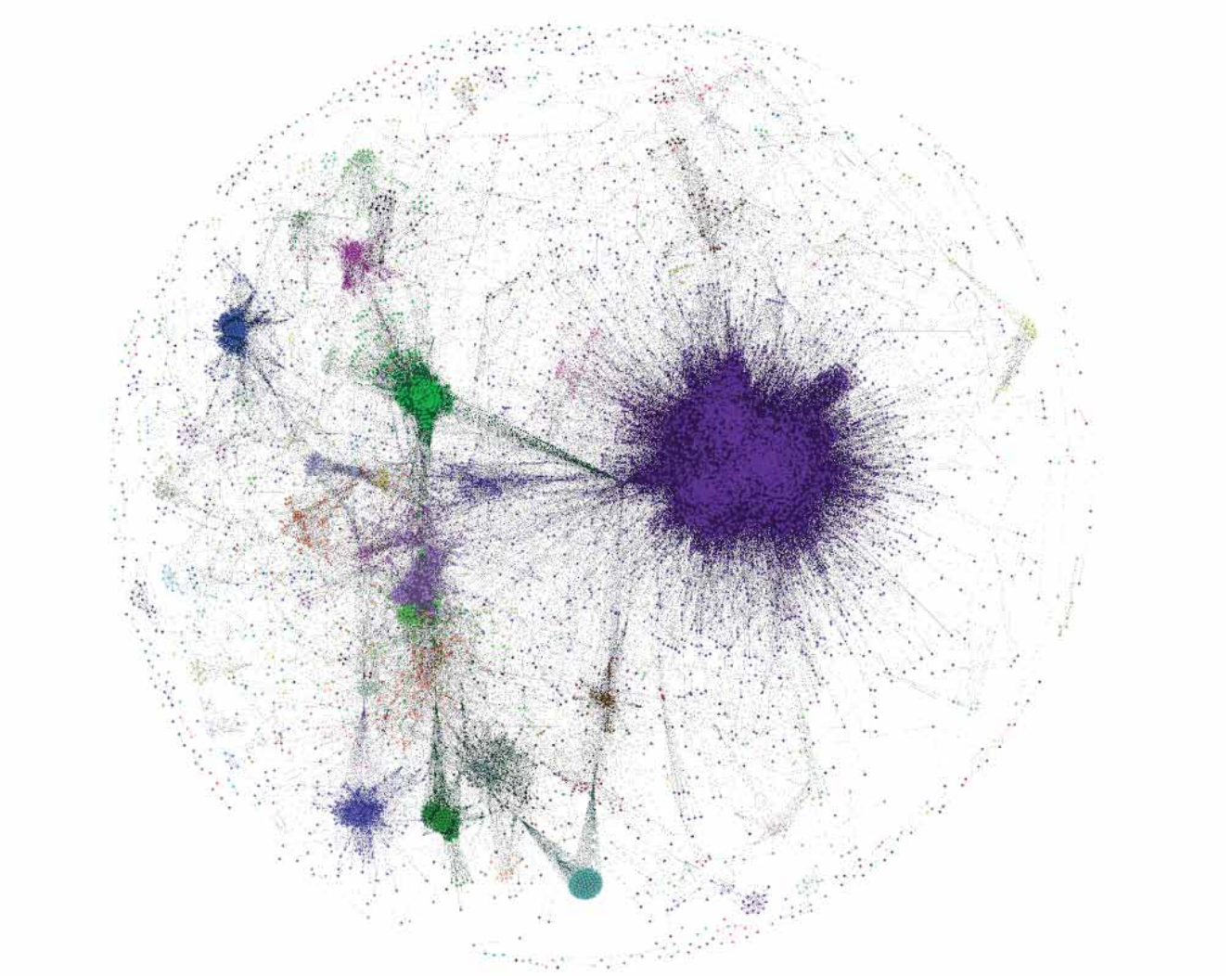


Fig. 2: Visualization showing Genizah fragments clustered based upon co-occurrence in the secondary literature.

that the largest cluster of fragments belongs to a subset of the Genizah that the literature[7] describes as documentary. These fragments consist of the everyday ephemera of life in the classical Genizah period, e.g. letters, accounts, merchants' papers, court depositions and other day-to-day writings. Our analysis clearly shows that the largest single contribution to the literature, measured in terms of fragments discussed, comes from the writings of Shelomo Dov Goitein, principally his six-volume work *A Mediterranean Society: the Jewish communities of the Arab world as portrayed in the documents of the Cairo Genizah*.[8] In addition to the scholarship surrounding the documentary Genizah, there are several other distinct clusters that represent the breadth of Genizah studies, which includes magic, literary works, medicine, liturgy and religious law. In order to maximize our coverage of the collection, we have tried to target a cross section of works that encompasses all of these clusters.

If we consider the example of manuscript T-S 24.64, then our analyses of the bibliography identified 15 scholarly works that are known to have discussed this fragment to varying degrees. The full texts of seven of these were available to us for inclusion within our corpus.

3.2 Corpus construction

The process of building our corpus is ongoing, and scholarly works continue to be added as and when we clear the rights. Accessioning new texts involves format shifting the source material into a machine-readable format (UTF-8) and then automatically segmenting the raw text according to its structure (e.g. page boundary, subsection, chapters). As of December 2013, our corpus contains 38 scholarly works by

[7] Frenkel 2010.

[8] Goitein 1967-1993.

Fig. 3: The top 50 terms associated with T-S 24.64 scaled according to term weight.

25 different authors and represents over 6,500 pages of text about the Genizah. This includes Goitein's *A Mediterranean Society* as well as works by other prominent Genizah scholars including Moshe Gil and Jacob Mann. In total, our corpus contains references to 6,322 fragment classmarks.

Many of the recent works have a native digital edition, but where an electronic source is not available we have had to resort to using a cradle scanner to image a physical copy and then OCR in order to produce machine-readable text. Therefore, approximately half of our corpus consists of text produced using an OCR process that has a reported error rate of approximately 3% when applied to printed material.

3.3 Deriving a term-weighted vocabulary

Using the citation information from the bibliography along with pattern matching for classmark recognition, we automatically associate blocks of text from the secondary literature with a given fragment. With regard to context, our approach tries to identify the boundaries of the paragraph containing the classmark, but if one cannot be detected, then it defaults to including the whole page. Once this mapping is performed, our system generates a vocabulary of the words used to describe the fragment and assigns a set of term weights.

The term-weighting scheme we have used is length-normalized TF-IDF,[9] which in this instance represents a single value measure of the importance of a word to a given fragment relative to the frequency of the word across all fragments. Term-weighting measures are the foundation of the vector space model[10] used in modern information retrieval systems and thus provide a representation which is suitable to enable 'full text' search of the fragments. Fig. 3 shows an example of the resulting vocabulary for manuscript T-S 24.64 expressed as a word cloud with the terms scaled according to their term weight. In total, the vocabulary of terms that we have associated with this manuscript fragment is 3,802, but for the purposes of our experiments we have only used the 50 most commonly occurring terms for inclusion in our catalogue. As we can see from fig. 3, the top 50 terms closely track what little we know about the manuscript, but we have significantly expanded the number of terms that would cause

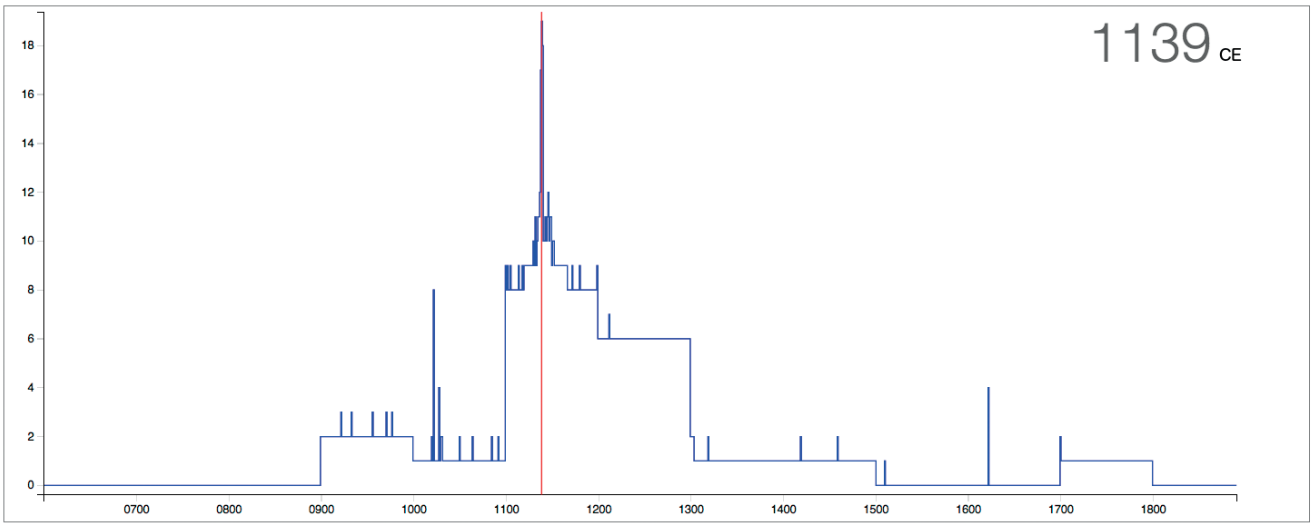


Fig. 4: Temporal graph showing dates associated with T-S 24.64.

[9] Salton and Buckley 1988.

[10] Salton, Wong, and Yang 1975.

this fragment to be surfaced within a search engine or the discovery layer of a library catalogue.

Whilst we recognize that a bag-of-words approach is not necessarily optimal in terms of producing a human-readable catalogue, it is important to highlight the fact that most modern discovery layers simply reduce any free text to this form of representation for the purposes of performing a search. As such, unless a library chooses to surface its raw catalogue information, the user is not necessarily aware that their search results are based upon automatically derived catalogue data.

3.4 Information extraction

Taking our approach a step further, we then attempted information extraction from the blocks of text we had associated with each fragment. In particular, we were interested in extracting any proper names and locations as well as any dating information in the form of temporal expressions.

Our approach to named entity recognition was based on the implementation of a Conditional Random Field (CRF) classifier contained in the Stanford coreNLP toolkit.[11] This is a supervised machine learning approach which required training using a gazetteer of medieval Muslim and Jewish personal names augmented to account for the wide variance in accepted spelling/transliterations commonly found in the literature. We associated names with over 3,500 fragments with varying degrees of success. For fragment T-S 24.64 we identified seven names, which were either derivatives of the known authors or closely related family members.

For dating, we developed a simple rule-based approach (based loosely on the techniques described in Mani and Wilson 2000) in order to extract temporal expressions from the text, focusing on assigning a creation date to each fragment. An extensive list of hand-crafted rules attempts to account for the range of dates present in scholarship which documents a period of over 1200 years using at least three disparate calendar systems. An example of the resulting temporal graph for fragment T-S 24.64 is shown in fig. 4. We can see that we identified a number of potential candidate dates from the literature, but a candidate date of 1139 CE appears over 18 times.

3.5 Topic modeling

By hand-labeling the clusters of fragments that we identified from the bibliography (section 3.1, fig. 2), it was possible to assign an approximate classification to each fragment. This approach only yielded an extremely coarse-grained view of the collection, however.

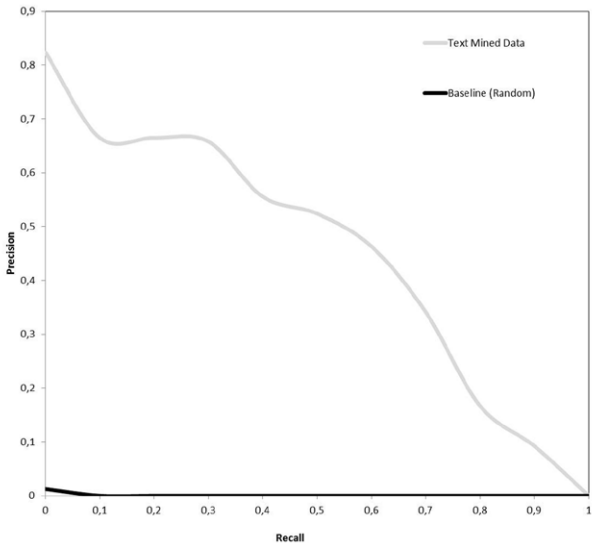


Fig. 5: Interpolated precision / recall graph contrasting retrieval effectiveness of our baseline vs. the text-mined data.

A more fine-grained classification approach was to use Topic Modeling (specifically Latent Dirchlet Allocation[12]) to identify clusters of words that commonly co-occurred within our corpus of secondary literature. Having identified a set of 75 topics, we asked a subject expert to hand-label these based on the top 50 terms most frequently associated with each topic. These labels then became the basis of our classification system. We then applied our topic model to the term-weighted vocabulary that we had generated for each fragment, thus allowing us to fit the fragments to our subject-based classification scheme. The resulting classifications provide a suitable mechanism to browse the collection using content-based themes and highlight the diverse range of topics covered in the Genizah (e.g. travel, religion, society, business, finance, medicine and the occult).

4. Evaluation

The main aim of this work was to generate usable catalogue data to facilitate text retrieval on the fragments. Our approach to evaluation has followed the Cranfield paradigm,[13] which is widely used in information retrieval. Put simply, we have taken a number of real-world queries and made use of a subject expert to identify a set of known relevant fragments. Consider the following query:

[11] Finkel, Grenager, and Manning 2005.

[12] Blei et al. 2003.

[13] Voorhees and Harman 2005.

<title>sufism or mysticism
<narr>Texts referring to Sufi practices, to mystical practices or to interest in 'sodot', qabbala, numerology etc

The title field represents the keywords that a user might enter into a search engine, whilst the narrative contains the guidance provided to the subject expert to help interpret the underlying information need.

Using the metrics of precision and recall, the graph in fig. 5 contrasts the retrieval effectiveness of search results based upon our text-mined data against a baseline strategy of returning 100 random fragments from the 6,322 for which we have catalogue data. As expected, the baseline performs badly with the likelihood of returning a relevant document by random selection being less than 1 in 1000. Contrast this with the results of searching our catalogue and we can see that this approach provides a significant source of evidence as the basis for discovery. In terms of precision @ 10 documents retrieved (which effectively models the first page of a search engine's results), then on average 1 in 2 documents returned were judged relevant by a subject expert, and this holds true all the way to precision @ 30 documents retrieved.

5. Conclusions

In this paper we have outlined a methodology for combining rich citation data with a corpus of secondary literature in order to automatically generate a content-based catalogue for the Taylor-Schechter collection. Our evaluation demonstrates that this approach has produced a weighted vocabulary for 6,322 fragments that, when used as the basis of performing retrieval, significantly outperforms a strategy of randomly selecting fragments. Given the sparseness of existing metadata for this collection, any solution that can recommend relevant fragments is a significant step forward. In this context an average of five relevant results in the first ten retrieved is extremely encouraging. Our exploration of Named Entity Recognition and Topic Modeling has been positive, but our ability to evaluate the work is limited by the lack of a ground truth or gold-standard metadata. Our next step is to engage with several pilot communities in order to produce an appropriate test collection to evaluate these elements more formally.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the Andrew W. Mellon Foundation, which is providing funding for Cambridge University Library's project 'Discovering history in the Cairo Genizah' (2012–14).

REFERENCES

Blei, D., et al. (2003). 'Latent Dirichlet allocation', *The Journal of Machine Learning Research*, 3.4–5: 993–1022.

Cambridge University Digital Library (2014), URL: http://cudl.lib.cam.ac.uk (accessed on March 14, 2014).

Finkel, J., Grenager, T., Manning, C. (2005), 'Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling', in *43rd Annual Meeting of the Association for Computational Linguistics*, 363–370.

Frenkel, M. (2010), 'Genizah Documents as Literary Products' in B. Outhwaite and S. Bhayro (eds.), *From a Sacred Source: Genizah Studies in Honour of Professor Stefan C. Reif* (Cambridge Genizah Studies Series, 1), 139-156.

Genizah Research Unit Bibliography (2014), URL: http://cudl.lib.cam.ac.uk/bibliographies/genizah (accessed on March 14, 2014).

Goitein, S. D. (1967–1993), *A Mediterranean Society: the Jewish communities of the Arab world as portrayed in the documents of the Cairo Genizah*, 5 vols. and index vol. (University of California Press).

Levy, N., Wolf, L., Stokes, P. (2013), 'Document classification based on what is there and what should be there', in *Digital Humanities 2013: Conference Abstracts* (Lincoln, NE: University of Nebraska–Lincoln), 279–82.

Mani, I., and Wilson, G. (2000), 'Processing of News', *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, 69–76.

Naji, N., and Savoy, J. (2011), 'Information Retrieval Strategies for Digitized Handwritten Medieval Documents', in *Proceedings of the Asian Information Retrieval Symposium*, Dubai, LNCS #7097 (Berlin: Springer), 103–114.

Reif, S., and Reif, S. (2002), *The Cambridge Genizah Collections: Their Contents and Significance* (Cambridge University Press).

Salton, G., Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information Processing and Management,* 24.5: 513–523.

——, Wong, A., and Yang, C. S. (1975), 'A Vector Space Model for Automatic Indexing', *Communications of the ACM*, 18.11: 613–620.

Shweka, R., Choueka, Y., Wolf, L., Dershowitz, N. (2013), 'Automatic extraction of catalog data from digital images of historical manuscripts', *Literary and Linguistic Computing,* 28.2: 315–330.

Voorhees, E. and Harman, D. (2005), *TREC: Experiment and Evaluation in Information Retrieval* (The MIT Press).

Article

Statistical Processing of Spectral Imagery to Recover Writings from Erased or Damaged Manuscripts

Roger L. Easton and David Kelbe | Rochester

Abstract

Imaging techniques for recovering historical writings that have been deliberately erased to make palimpsests or otherwise damaged have a long history. These writings often may be recovered from images collected under different illuminations over a range of wavelength bands. These spectral images are combined in subsequent image processing to enhance the visibility of the desired text. This paper considers the application of several processing algorithms based on the spectral statistics of the image data. These methods may be tailored to the specific condition of the manuscript and may enhance erased or damaged writings better than other techniques.

1. Introduction

Many historical manuscripts have been damaged by a variety of mechanisms: the ravages of natural processes over time, inappropriate storage, and deliberate action. Inks fade, manuscripts may have been scorched or charred by fire or damaged by animal excretions, parchments were often deliberately erased and reused to make palimpsests, etc. These faded, erased, or otherwise damaged manuscripts may be impossible to read without artificial aids to human vision. Moreover, simple devices (such as optical magnifiers to enlarge the text) and more complicated tools (e.g. ultraviolet lights to enhance the contrast of erased text relative to the background parchment) only improve the appearance to the human eye. The ultimate limit to the success of these visual aids is the human vision system itself, which is limited by the capability of the eye lens, the photochemical mechanism for sensing light, and the perception of images by the brain. One example of the constraint of the visual system is its narrow range of spectral sensitivity, which is limited to wavelengths between approximately 400 nm (blue) and 700 nm (red). Significant information is often present at wavelengths outside this range.

To obtain even better readings from the manuscripts, the limitations of the human visual system in terms of wavelength and sensitivity must be overcome. In other words, it is necessary to render subjectively imperceptible information so that it becomes visible to the eye. The need for better renderings has long been recognized and led to attempts to enhance the legibility of manuscripts by applying the very earliest imaging technologies. Chiara di Sarzana has presented an excellent history of photography applied to manuscripts.[1] In 1840, only one year after the accepted date of the first photograph, Jean Baptiste Biot reported that William Henry Fox Talbot had made images on sensitized paper of a Hebrew psalm, a Persian newspaper, and a 13th-century charter in Latin.[2] Apart from this early work, the use of photography to document historical writings did not become common until the technology of monochrome emulsion imaging matured in the 1880s, when flexible cellulose film substrates became available, thus eliminating the requirement for fragile glass plates. James Rendel Harris was an early imaging experimenter, having photographed manuscripts at St. Catherine's Monastery in Sinai in 1888.[3] Harris also assisted the Scottish twin sisters Agnes Smith Lewis and Margaret Dunlop Smith Gibson when they imaged the *Codex Sinaiticus Syriacus* on a subsequent visit to St. Catherine's in 1892.[4] It is interesting to note that the thousand or so film sheets exposed during that trip to the monastery were not processed until the travelers returned to England several months later, which prevented them from

[1] Di Sarzana 2006.

[2] Biot 1839-1840, *Chronique of Bibliothèque de l'École des Chartes* I, p. 408.

[3] Harris and Harris 1891.

[4] Gibson 1893.

assessing the value of their images until it was much too late to make any adjustments.

These first efforts at photographing manuscripts were not directed at improving the visibility of the text, but rather at creating copies of the manuscripts that could be studied intensively at a more convenient location. The first attempts to apply scientific technology to enhance the visibility of texts, rather than just to document them, apparently also took place in the mid-1890s, when Ernst Pringsheim and Otto Gradenwitz pioneered the use of multiple photographic images to enhance the erased 'undertext' of a palimpsested manuscript relative to the later text. They created a 'sandwich' of positive and negative photographic transparencies that had been exposed and processed specifically to make the later 'overtext' less visible than the original 'undertext' in the final image.[5] This was probably the first use of 'image processing' to enhance a manuscript that was difficult to read. The single documented result was quite successful.[6]

This early work set the stage for a long history of imaging and image processing for recovery of erased and damaged texts. In the early 1900s, the Benedictine monk Raphael (Gustav) Kögel photographed fluorescence emission from manuscripts after illumination with ultraviolet light generated by a variety of sources then available, including electric arc, mercury vapor, and metal wire lamps. As bandpass filters, he used various liquids in glass cuvettes placed in front of the camera lens. Kögel's goal had been to use fluorescence photography to enhance the visibility of erased text and therefore to replace chemical treatments of palimpsests with tincture of oak galls[7] or Gioberti's tincture[8] to enhance the original texts. Such treatments sometimes rendered the metallic ink more legible for a short time, but left severe damage in their wake. Kögel's goal was to improve the readability without inflicting any lasting damage. In the dedication to his detailed monograph on palimpsest photography in 1920,[9] Kögel acknowledged the pioneering contributions of Pringsheim and Gradenwitz as his inspirations. Though difficult and tedious to implement and requiring long exposures (up to 24 hours), the potential capability of Kögel's method was evident, as he reported improvements in readability over previous results by up to 50%.[10] Kögel also helped initiate the establishment of the Institut für Palimpsestphotographie at the Archabbey of Beuron. In 1913, the Institute published a book of 152 black-and-white plates of the leaves of Codex Sangallensis 193 that were produced using the techniques developed by both Kögel and by Pringsheim and Gradenwitz.[11] This volume was intended to be the first of a series that would document the undertexts of palimpsests, but the effort was interrupted by World War I and apparently not resumed thereafter. Nonetheless, Kögel's work provided the first examples of what has become the basis of modern imaging methods used to recover text from palimpsests. An apparently similar technique was developed independently in Italy by Luigi Pampaloni.[12]

The methods used to image historical manuscripts advanced relatively little over most of the rest of the twentieth century. Color and infrared emulsions were first invented in the 1920s and were used later to assist the reading of manuscripts,[13] but the technology and capability of emulsion photography changed only in minor details from that time forward. For this reason, imaging methods to assist in the reading of manuscripts changed little until late in the 20th century, when there was a revolution in imaging and illumination hardware and processing software, which itself had been driven by the revolution in computing technologies. This advance in technology has resulted in new 'lights', 'eyes', and 'brains' that may now be applied to the task of recovering damaged or erased writings.

Several generations of the new imaging technologies have been applied to assist scholarly readings of manuscripts. The first arguably was the combination of broadband illumination and discrete bandpass filters to collect sets of spectral images, as had been used in the early experiments on the *Archimedes Palimpsest*.[14] The Forth-Photonics 'MUSIS' camera exemplified a later generation; it used a tunable optical filter to obtain images at a number of spectral bands from ultraviolet to near-infrared with a spatial resolution of

[5] Pringsheim and Gradenwitz 1894; Schnauss 1900.

[6] Pringsheim and Gradenwitz 1901.

[7] Edmonds 1998.

[8] Albrecht 2012.

[9] Kögel 1920.

[10] Kögel 1914.

[11] Dold 1913.

[12] Rostagno 1915.

[13] Haselden 1935.

[14] Netz, Noel, Tchernetska, and Wilson 2011.

1280 x 960 pixels (approximately 1.2 megapixels), which is quite coarse by today's standards.[15] Cameras belonging to the newest generation have much better spatial resolution (up to 50 megapixels) and use different illumination wavelengths generated by the new 'lights' to collect spectral image sets. These lights are made from light-emitting diodes (LEDs), which emit radiation generated by changes in electronic states rather than as a byproduct of heat, which makes them much cooler and safer to use for imaging of historical artifacts. Another useful feature of LEDs is their narrow emission bandwidth, which is typically between 10 nm and 50 nm. The illumination may be configured to interact with the manuscript in the usual reflection mode, but also in transmissive mode and in fluorescence, where the ink and substrate absorb incident radiation and emit longer wavelengths that are characteristic of the material. To ensure accurate measurements of the reflectance, transmittance, or fluorescence over the range of available wavelength bands, which is necessary for statistical analysis, standard reflectance and transmittance targets in the field of view are used to calibrate the collected spectral images.

The photoelectronic receptors in the sensors of the new eyes can 'see' light that is invisible to the human visual system, including visible light that is too faint to be perceived and wavelengths outside the human's range of sensitivity. For example, the silicon charge-coupled device (CCD) detector is usefully sensitive over the range of wavelengths from approximately 350 nm (in the near-ultraviolet region) to 1100 nm (near infrared), which is much broader than the range of human vision from approximately 400 nm (blue) to 700 nm (red). The 'invisible' bands of energy that are both shorter and longer than the range of human vision convey useful information about a manuscript. This extended range of vision is useful by itself, but when combined with LED illumination, the new 'eyes' can see a larger number of 'colors' than the three discrete color sensors of the human eye. If additional optical bandpass filters are incorporated in the optical path when using short-wavelength illumination (ultraviolet or blue), spectral images of the fluorescence may be collected, which have proven to be useful for some manuscripts, such as one undertext in the *Archimedes Palimpsest*, a commentary on Aristotle's treatise entitled *Categories*.[16]

The spectral images used in these projects typically include 12 reflective bands, as many as eight fluorescence bands, and often several transmittance bands (particularly if the subject is a palimpsest or a paper watermark that might benefit from enhanced visibility in this mode). The obvious negative aspect of transmissive illumination is that text on both sides of the leaf is often visible in the images, but this may be segmented by subsequent statistical processing.

It is important to recognize that the camera lens in the imaging system must transmit light and maintain focus over the wider range of wavelengths in spectral imaging from the near-ultraviolet to the near-infrared. This means that standard photographic lenses designed for visible light are not satisfactory. Because the glass components of standard lenses severely attenuate any ultraviolet reflection or emission from the object being examined, it is necessary to fabricate elements from crystalline quartz, which is transparent at these wavelengths. Also, if the lens is focused on an object at wavelengths in the middle of the visible range, wavelengths at the extrema of the transmitted ultraviolet and infrared bands will be unfocused. Fortunately, such 'UV-VIS-IR spectral lenses' have already been designed and are available on the market,[17] although they are significantly more costly than otherwise-similar standard lenses designed for visible light.

The camera system that forms the new 'eye' used for collecting these spectral images was constructed by Megavision, Inc., and has been documented elsewhere.[18] The rapid access to the imagery available from such a camera is a far more desirable state of affairs than that faced by the Smith sisters during their trips to Sinai. This imaging system has been used in several important projects, including the study of the *Archimedes Palimpsest* and the current project to image the palimpsests in the 1975 'New Finds' at St. Catherine's Monastery in Sinai.

The images collected at the different wavelength bands and using the different imaging modes are analyzed and processed in the new computer 'brains', with the goal of creating 'new' images that allow subtle differences in reflectance and color of the features to be distinguished. The choice of processing algorithm to enhance the desired feature depends on the specific situation. For example, it is occasionally possible for text that is invisible to the

[15] Rapantzikos and Balas 2005.

[16] Bloechl, Hamlin, and Easton 2010.

[17] E.g., http://www.jenoptik-inc.com/coastalopt-standard-lenses/uv-vis-nir-60mm-slr-lens-mainmenu-155.html.

[18] Easton et al. 2010.

Fig. 1: Visual appearance (on left) compared to pseudocolor rendering (on right) of a section of an *Archimedes Palimpsest* leaf (f. 94r–91v of the *Euchologion* overtext) including gutter. The visibility of the undertext relative to the parchment is enhanced because of the rendering in a contrasting color.

human eye to be read directly from a single image at a single wavelength; this is most often true for manuscripts that were scorched or carbonized, in which case the writing may be visible in an image collected under near-infrared illumination. More often, it is necessary to combine images collected in different bands to enhance subtle variations in the spectral reflectance of the text of interest. The algorithms for combining image bands may be rather loosely classified as 'deterministic', where the same combinations of bands are used for more than one leaf, or as 'statistical', where the band combinations are calculated from the statistics of the gray values in the ensemble of spectral bands for each class of object. In either case, the condition of the leaf (hair or flesh side) and variations in the degradation across the leaf ensure that the optimum choice of bands and processing method generally mean that the processing method also changes.

2. Deterministic processing methods

Although the focus of this paper is on statistical methods, a brief introduction to deterministic renderings may be useful to the reader. Perhaps the simplest technique of this kind is the evaluation of the difference in gray values of two spectral bands, so that regions on the leaf with similar reflectances subtract to small numerical values, while features with different gray values take on an extremum value – either positive or negative – that may be rendered as a more visible feature. Another useful example of deterministic processing is pseudocolor rendering of the image, where monochrome image bands generated under different conditions of illumination are inserted into the red, green, and blue channels of a visual color image. If judiciously chosen, the visibility of the text of interest may be enhanced in a particular combination of bands. This was the primary

means for rendering image data that was used in the study of the *Archimedes Palimpsest*, where images through a blue filter under ultraviolet illumination and through a red filter under tungsten lights were combined to render the overtext in neutral gray or black and the undertext in a reddish tint.[19] The resulting color 'cue' helps the reader distinguish between the two texts.

Because they do not require evaluation of spectral statistics, deterministic methods may usually be implemented quickly, which is a very distinct advantage in large projects with many leaves to be transcribed. For this reason, they are often used productively as a 'first pass' in image processing. If the resulting images are sufficient for scholarly transcription, no further processing is necessary. For those images that are not readable from images processed by deterministic methods, a second pass involving more computationally intensive statistical methods is applied.

3. Statistical processing methods

Statistical image-processing methods analyze the ensemble of gray values at each pixel over the range of spectral bands with the same goal as deterministic processing: to find linear combinations (weighted sums) of spectral bands that enhance the desired text. Numerous techniques of this kind exist; many were originally developed for military purposes (such as camouflage detection) or for environmental applications (such as characterizing ground conditions or assessing the health of crops). The same methods are directly applicable to the goal of enhancing subtle differences in reflectance spectra of the different features on a manuscript.

Consider a set of images collected under N distinct conditions of illumination that may include reflective, transmissive, and fluorescent modes. The integer gray values of a specific pixel measured under the N conditions form a vector with N components. The ensemble of N-dimensional vectors may be plotted, at least in theory, as an N-dimensional histogram, which will exhibit 'clusters' of pixels belonging to the same object class. For example, the gray values of a pixel in images of the same manuscript under green and red light form a two-dimensional vector, and the ensemble of such vectors from all the pixels in the image forms a statistical probability distribution that may be analyzed to look for correlations among object features. If a third illumination is added, then the histogram is formed from the three-dimensional vectors. Of course, the dimensionality of the vector is equal to the number of bands. A rule of thumb is that the number of bands should equal or exceed the number of features to be distinguished. In these projects, it is common to include a dozen or more spectral bands, although it is often useful to use subsets of the images based on observations of the text's visibility.

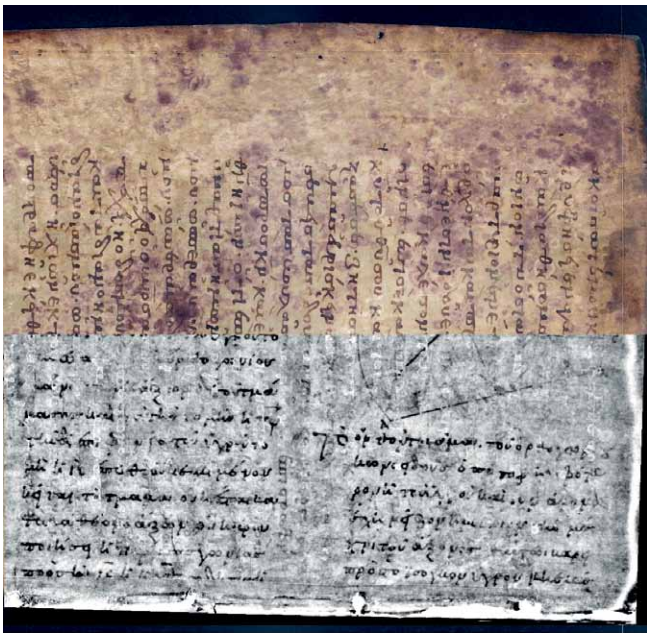


Fig. 2: Visual appearance (top half) compared to result of spectral unmixing (bottom half) of contiguous halves of f. 28v of the *Archimedes Palimpsest*. The unmixing process effectively removes the overtext. In particular, the diagram is much easier to see in the output image.

4. Spectral unmixing

One method that has proven useful in text analysis that requires significant user interaction is 'spectral unmixing', where pixel regions belonging to specific object classes are identified by the user first, e.g. parchment, mold, overtext, erased text, etc. The algorithm then calculates the class membership of each pixel in the image based on the similarity of its spectrum to each of the specified classes. Although it is intensive both in terms of human interaction and subsequent computation time, this method was applied with success during the early experiments on the *Archimedes Palimpsest* and also to spectral image data collected with a MUSIS camera.[20] Spectral unmixing of manuscript imagery deserves additional study, particularly in situations where the feature spectra are known or can be determined a priori.

[19] Netz and Noel 2007.

[20] Knox et al. 2001.

In our applications, we have most frequently employed two similarly named statistical methods for enhancing the visibility of erased or damaged text: principal component analysis (PCA) and independent component analysis (ICA). PCA 'rearranges' the data in a set of N spectral images to create a different and equivalent set of N images that satisfy two properties: (1) the derived images are uncorrelated and (2) they are arranged in descending order of variance. Interested readers are advised to consult the very accessible introduction to PCA by Schlens.[21] ICA also rearranges the set of N spectral images to make a different set of N images, but the output bands are distinguished by statistical 'independence', which will be discussed shortly. Hyvärinen and Oja have written a useful introduction to ICA.[22]

5. Principal Component Analysis

PCA is implemented by evaluating the difference in the gray value of each pixel for each combination of two spectral bands and then evaluating the expectation value for these differences over all pixels in the image. The result is the real-valued and symmetric covariance matrix. The eigenvectors of this matrix are the orthogonal axes of the principal components. The eigenvectors are ordered in sequence based on the magnitudes of the corresponding eigenvalues, with the eigenvector associated with the largest eigenvalues first. The implementation of principal component analysis may be viewed as 'projecting' the image pixels of an N-band image onto each of these N orthogonal axes, followed by rendering each pixel as a gray value based upon its location on the particular axis. In other words, the act of data 'projection' onto each axis evaluates a weighted sum of the original N images. The end result of the projections onto the N orthogonal axes is a set of N 'new' monochrome images that are equivalent to the original N image bands.

The first PC band results from the projection of the data in the N-dimensional histogram onto the axis that spans the widest possible range of variation of the image data, so that the first PC image exhibits the widest possible range of variance of the statistics or the widest range in 'contrast' of image features. In the application to manuscript imaging, the range of gray values of the first PC band is determined by the pixels in the areas of the image that are brightest and darkest overall, such as the light parchment and darkest overtext characters.

The second PC image is the projection of the N bands of data onto the axis of the N-dimensional histogram with the largest possible range of variation in a direction that is orthogonal to the axis used for the first PC image, so that this second band exhibits a smaller range of variation than the first. The second PC band is also a weighted sum of the original N spectral images, but the fact that the two axes are orthogonal means that the first and second PC images are 'uncorrelated.'

The process of determining the orthogonal axis with the next largest variation and projecting the data from the histogram onto that axis is repeated to generate a total of N mutually orthogonal PC bands. The monotonic decrease in the sequence of eigenvalues corresponding to each axis means that the low-order bands (evaluated first) are dominated by large-scale variations in the original N bands of data, while the high-order bands (evaluated last) are dominated by small-scale variations in the original data, which may be random fluctuations ('noise') or image features with very little contrast (such as erased undertext).

It is important to recognize that the projected data values are floating-point numbers that must be mapped to integer gray or color values for display, usually in an 8-bit format with $2^8 = 256$ possible values per color that can be displayed on a computer screen. The process requires selection of the 'lightest' and 'darkest' values to be 'quantized' to the extrema of the available integer values. This process of quantizing the floating-point value to an 8-bit integer means that a range of different numerical values will be mapped to a single integer gray value. This means in turn that different features may be rendered at one gray value or over a small range of gray values, so the features may not be distinguishable for a particular choice of the limiting light and dark pixels. For this reason, it is essential to have the option to change the range of values for the rendering of the image. Similar problems appear in other applications, such as medical X-ray computed tomography, where the radiologist often changes the limits of the grayscale rendering to visualize subtle features of the pathology. The importance of the choice of rendering is easy to overlook.

In the best possible result of PCA processing, each feature class in the scene (e.g., parchment, overtext, undertext, etc.) would appear exclusively in a single specific band in the new set. In this case, each class of feature would dominate the range of gray scale in the specific PC band, while pixels belonging to the other classes of feature would exhibit the same gray value and thus 'disappear' into the background in that band. In fact, this happy occurrence is rare; traces

[21] Shlens 2009.

[22] Hyvärinen and Oja 2000.

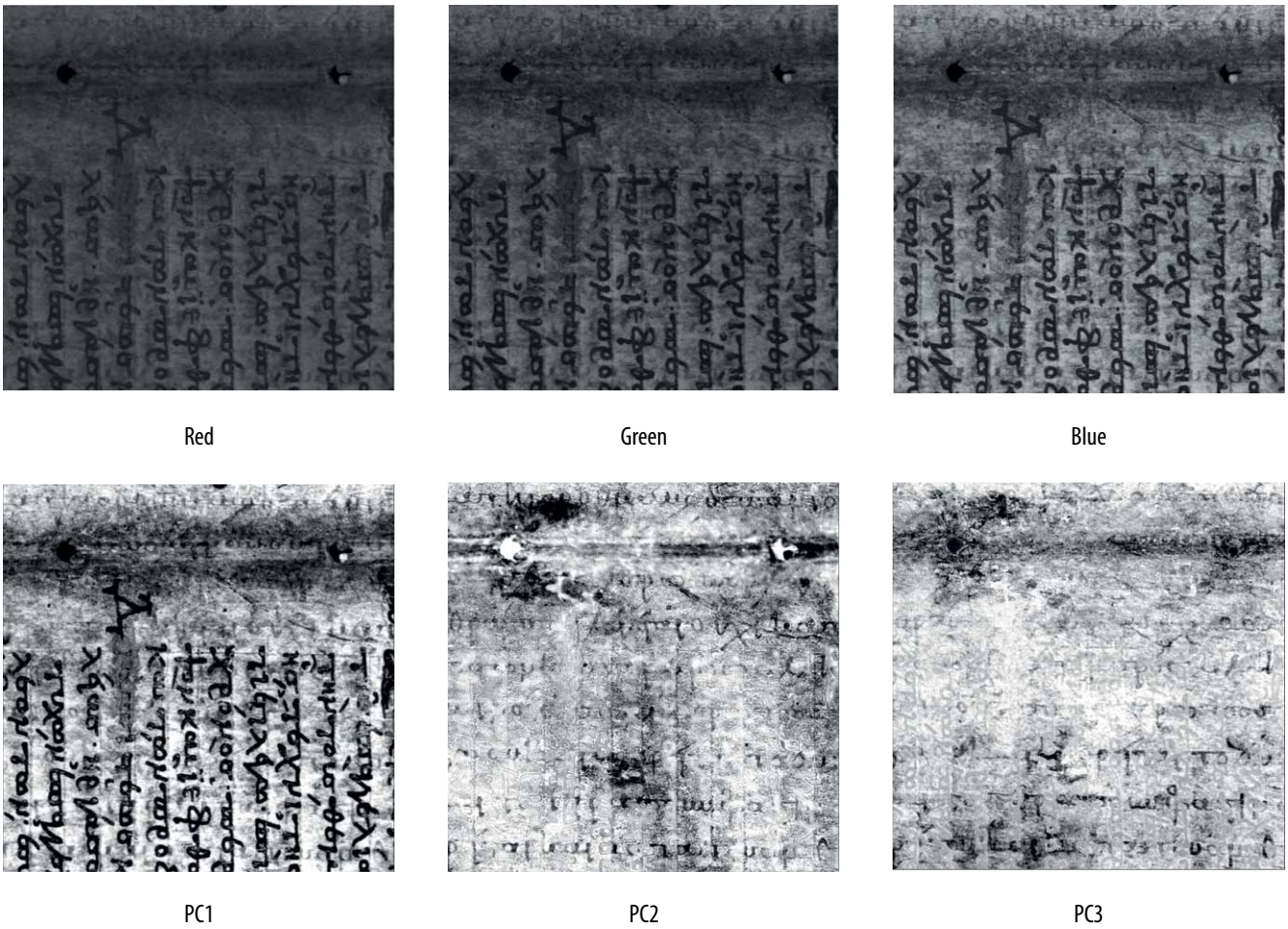


Fig. 3: The top row shows the red, green, and blue color channels of an image of fluorescence generated by ultraviolet light. The images in the bottom row are the corresponding principal components. Note that the first principal component closely resembles the blue channel, which is considerably lighter than the red or green channels. The dark overtext that is so visible in the first principal component is rendered with approximately the same gray value as the parchment in the second and third principal components, so the overtext 'disappears' into the parchment and the erased undertext is far more visible in these bands.

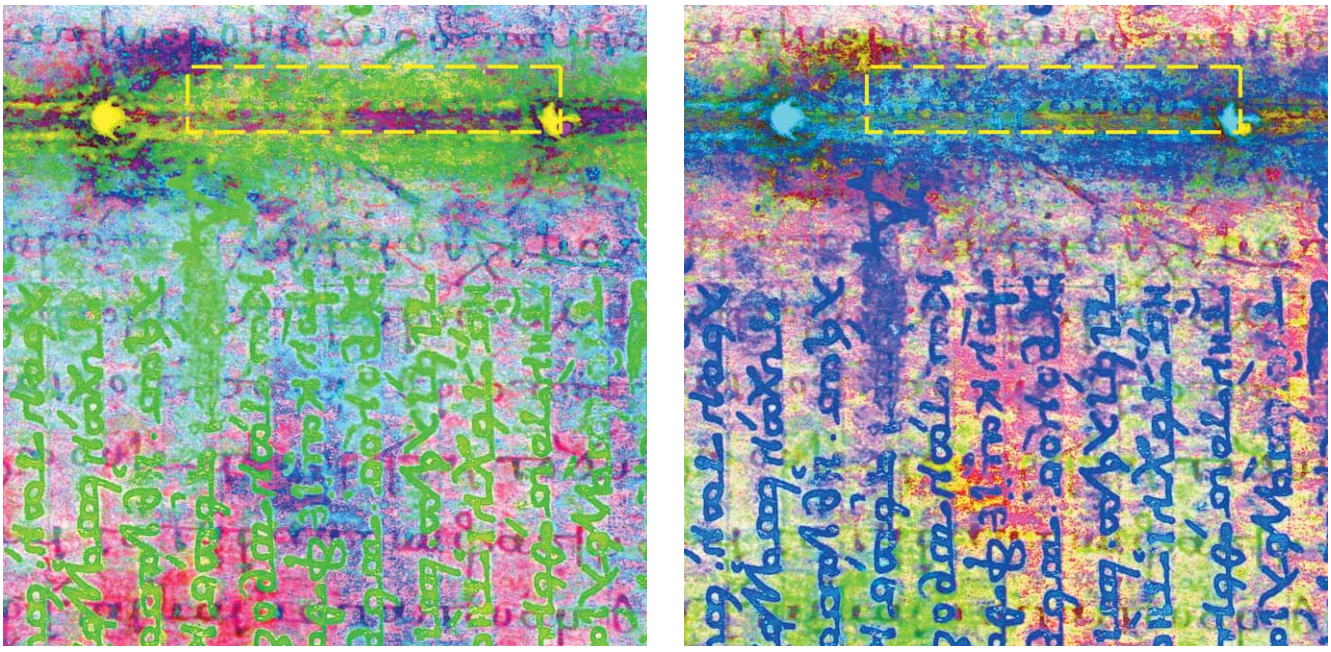
of overtext usually appear in those bands where the erased undertext is most visible, for example, and the variation in statistics across the scene (e.g. due to variations in erasures) ensures that the desired erased text often appears in more than one PC band. This last observation results in the strategy of inserting two or three bands into the channels of a pseudocolor image, and the rendering is altered interactively by the observer, as will be described.

Software to implement principal component analysis is widely available, including versions in open-source imaging software such as ImageJ.[23] The processing implemented here was performed with the ENVI software toolkit that was written for environmental remote sensing by Exelis Visual Information Systems.

Because the 'rearrangement' of the original N input image bands is based only on the statistics evaluated from the N-dimensional histogram, there are no selectable parameters other than the choice of the specific original bands and of the region in the scene where the statistics are evaluated (both of which are important). This independence from input parameters means that PCA is widely applicable for many types of data sets (not just images), but it also means that it may not succeed in separating the features unless the statistics of the different features are truly orthogonal or readily distinguishable.

An example of PCA processing of a leaf of the *Archimedes Palimpsest* is shown in fig. 3. The undertext is a commentary on Aristotle rather than one of the treatises by Archimedes that occupy most of the leaves. This text was particularly difficult to read in the deterministic renderings, but was made quite

[23] ImageJ is available from the National Institutes of Health at http://imagej.nih.gov/ij/.



Hue angle = 0° Hue angle = 120°

Fig. 4: Pseudocolor rendering of the principal components in Fig. 3 at two different hue angles. Undertext in the gutter region outlined by yellow dashed line is barely visible, if at all, in the image with hue angle equal to 0°. The visibility of the text in the gutter is much improved when rendered with a hue angle equal to 120°.

legible by PCA processing. The manuscript was illuminated by a narrow band of ultraviolet light centered approximately about $\lambda = 365$ nm. The ultraviolet illumination generated visible fluorescence in the manuscript, which was imaged by a sensor whose pixels were covered by visible color filters in the common Bayer pattern, so that each pixel in the array measures the amount of light in one of the three primary colors (red, green, and blue); half of the pixels in the sensor measure green light, and one quarter each measure red light and blue light. The camera had the capability to translate the sensor piezoelectrically relative to the image by increments of the pixel separation, so the amount of red, green, and blue light at each pixel could be measured from exposures at the different sensor positions. Three principal components are generated, matching the number of input bands. The contrast between the overtext and parchment dominates the first PC band because it spans the widest range of variation in gray level in the three-band image. Because the second and third PC bands are orthogonal to the first, the three-dimensional vectors of gray value of the overtext and parchment are projected onto the same location on the projection axis. In other words, the parchment and overtext are rendered in the same small range of gray values in PC bands #2 and #3, so that the overtext 'disappears' into the parchment. When rendered as black-and-white images, the much smaller range of gray values determined by the undertext and background parchment is rendered to the full range of eight available bits of gray value, so that the contrast between the undertext and parchment dominates these higher-order principal component bands.

Although the goal of PCA is to construct orthogonal renderings that will segment the feature classes, the typical overlap of class statistics means that the desired feature appears in more than one output principal component band. In a palimpsest, the variation of the erasure of the original text across the leaf generally means that the histogram of the image varies with the position on the leaf, as was the case for the Aristotle commentary. This means, in turn, that the undertext appears most clearly in different locations of the three principal-component images and that pseudocolor rendering of PCA bands is often useful. Three PC bands are selected and inserted into the red, green, and blue color channels of a pseudocolor image. This means that small variations in grey value may appear as large changes in color, which may improve the readability of the erased text. The user may also actively change the pseudocolor rendering by varying the hue angle (formally, this changes the mapping of the color tones without varying the saturation and luminance values at each pixel). In practice, perceived changes in luminance and saturation accompany hue rotation; it is this

triad of rendering adjustments that is exploited to enhance the visibility of the undertext at different locations on the leaf. A comparison of pseudocolor renderings on one leaf of the Aristotle commentary in the *Archimedes Palimpsest* for two different hue angles is shown in fig. 4. The visibility of the text in the gutter enclosed by the dashed square is noticeably improved in the second example where the hue angle has been rotated by 120º. The variation in color rendering is most valuable when performed interactively by the transcriber, who can often see features more clearly in a dynamic rendering than in a static image.

Hue-angle rotation proved to be essential in the different problem of recovering information from an illuminated armorial on the manuscript of the French epic poem *Les Echéz d'Amour* at the Saxon State Library in Dresden. The manuscript was a victim of water damage after the Allied bombing of the city in 1945, and the armorial had been so smudged and tarnished that almost no structure in the seal remained visible. Spectral images of the leaf were collected and analyzed by PCA in the standard manner. Rotation of the hue angle of a subset of three PCA bands rendered in pseudocolor resulted in the image in fig. 5, where the seal clearly shows two unicorns, the second one being on a shield. After a short search of an online archive, this image made it possible to identify that the book had been owned by the Waldenfels family in Bavaria.

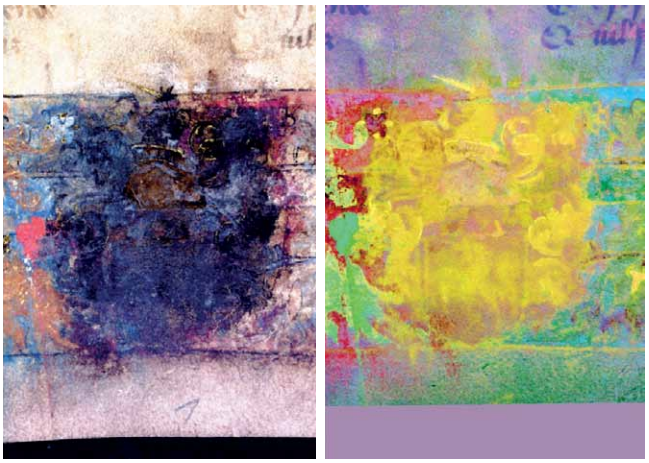
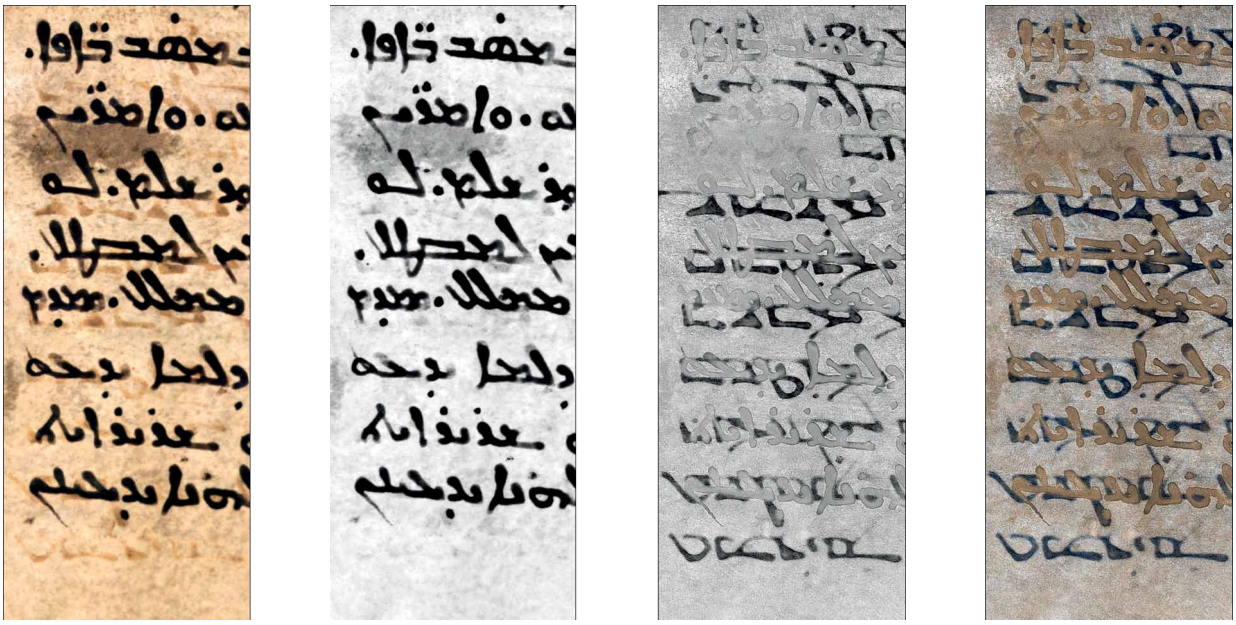


Fig. 5: Pseudocolor rendering of principal components of Armorial Achievement from f.1r of *Les Echéz d'Amour*, (OC66 from Saxon State Library, Dresden). The visual appearance on the left shows just a hint of a unicorn horn at the top. That unicorn is much more easily seen in the PCA pseudocolor on the right after rotation of the hue angle, which also shows a second unicorn on a shield and hints of a rampant lion or other animal on the left side of the shield. This image enabled identification of the family that owned the book.

6. Independent Component Analysis

The second similarly named statistical processing algorithm of 'independent component analysis' (ICA) also rearranges the set of N spectral images to make a different set of N images, but in this case the output bands are distinguished by statistical 'independence,' rather than the PCA criteria of being 'uncorrelated' and 'orthogonal.' The method is an example of 'blind source separation' applied to a mixture of input signals with the goal of determining the original independent components from different measurements of the combinations. A common example is the so-called 'cocktail party problem,' where independent conversations occur simultaneously at different locations in a crowded room. A single microphone positioned in the center of the room records the sum of independent conversations with weightings determined by the distances of each one from the microphone. The experience of the reader probably validates the difficulty of segmenting individual conversations from a single recording of this kind. Nonetheless, conversations could be separated from a recording by a single microphone if the frequency ranges of the voices were disjoint; a high-pitched voice could be separated from a simultaneous low-pitched voice in the same recording by applying the appropriate filter that passes one frequency range and blocks the other. The analogous situation for a manuscript would be the separation of two texts written in different colors of ink from images collected through two different bandpass filters. In the more realistic model of simultaneous conversations with overlapping frequency ranges, the required process for separating the components is less obvious; the corresponding imaging analogy is that the reflectance spectra of the different component writings will overlap.

The simultaneous voices may be recorded by multiple microphones placed around the room so that each one measures a different weighted sum of the conversations. The signals from the ensemble of microphones are analyzed and compared to segment the voices. At this point, it is useful to make an observation about the statistics of a meaningful conversation. The histogram of the spectrum of disordered random 'noise' tends to be uniform, with approximately equal populations at each frequency. The contrapositive statement is that histograms of 'ordered' components generally exhibit 'peaks' or 'clusters' at different frequencies, a feature that characterizes the statistics of a conversation as 'structured' or 'ordered.' From this observation, the process of independent



(a) (b) (c) (d)

Fig. 6: Example of ICA processing applied to Syriac 2A from St. Catherine's Monastery; the visual appearance is shown in (a) for reference. ICA rearranges the spectral image sets into uncorrelated output bands which effectively separate the overtext (b) and undertext (c). These output IC bands can then be combined into a pseudocolor rendering (d) that shows both texts.

component analysis is the search for a set of component conversations such that the histogram of each component is as 'clustered' as possible (with minimum 'disorder') and that the weighted sums of the hypothetical conversations match those of the individual recordings. The fact that the component signals are assumed to be independent means that the joint probability of all signals is the product of the probabilities of the individual components. This requirement for statistical independence is a stronger condition than the assumption in PCA that the signals are not correlated. The stronger condition invoked in ICA means that signals that are not well segmented by principal component analysis may be separable by independent component analysis.

When applying ICA to spectral images, the different feature classes (e.g. parchment, overtext, erased text, mold, etc.) are analogous to the individual simultaneous conversations, and the different spectral bands correspond to the individual recordings from the different microphones. Just as each sample of the recording from a microphone is a weighted sum of contributions from the different conversations, each pixel in a spectral image is a weighted sum of contributions from the individual object classes. ICA uses the multiband statistics of each pixel in the set of spectral images in an attempt to estimate the contributions of each object class at each pixel. The process is often combined with pseudocolor rendering and has been quite successful in recovering text from some palimpsests in the New Finds at St. Catherine's Monastery. The ICA tool available at ENVI was used in this analysis, although the algorithm is also available in other packages. An example of ICA applied to an image from St. Catherine's Monastery is shown in fig. 6.

Further, and often dramatic, improvements may be obtained from weighted combinations of images, including processed results from ICA and PCA and possibly original image bands. The choice of bands and the combination depends on the condition of the desired text. For example, erased text appears different under transmissive and fluorescence illumination; the process of scraping the text may thin the parchment so that the erased text is brighter than the surrounding parchment in transmission, while the erased text is darker than the surrounding parchment in fluorescence images. Statistical processing thus often produces different images that render different regions of the leaf better than other regions. It is often advantageous to combine the feature content from the two modes into a single image for viewing, while minimizing the visibility of the overtext. This can be done by combining multiple grayscale images into an intermediate pseudocolor image. Text features are thus

(a) (b) (c)

Fig. 7: ICA processing applied to Greek New Finds MG14 from St. Catherine's Monastery after subsequent processing to attenuate the visibility of the overtext: (a) visual appearance, (b) intermediate pseudocolor image showing different text features separated by color, (c) output grayscale image after modulating the brightness of individual colors in order to reduce the visibility of the overtext.

differentiated by color in the pseudocolor image, allowing precise control over the lightness and contrast of individual text features. In other words, color effectively separates different features and serves as a proxy for modulating the 'weight' of individual text components. The various features can be adjusted to optimize the visibility of the undertext in an output grayscale image. One such result is compared to the visual appearance and a pseudocolor PCA image in fig. 7.

It is important to note that all methods of statistical processing require an observer to make important selections from the image data, including the bands to be processed, the region(s) where the statistics are evaluated, the bands to be rendered in the final image, and the type and settings of any post-processing of the rendered image (such as the angle of hue rotation of a pseudocolor rendering). This observation indicates that the skill of the experienced observer remains an important factor in the success of the final result.

7. Conclusion

In summary, statistical processing techniques applied to spectral images of historical writings have been successfully applied to the task of distinguishing features that are not otherwise visible in images of manuscripts. These methods may be applied to a wide range of imagery. It seems quite certain that new developments in hardware and processing software in the near future will further enhance the value of statistical processing methods for spectral imaging. New sensors capable of imaging over a wider range of wavelengths may be anticipated, for example, and these techniques may be applied directly to that imagery.

ACKNOWLEDGMENTS

This work could not have been done without the assistance of many contributors, including Keith Knox, Michael Phelps, Executive Director of the Early Manuscript Electronic Library; Gregory Heyworth of the University of Mississippi; William Noel, Director of the Kislak Center and Schoenberg Institute for Manuscript Studies, University of Pennsylvania; Fenella France, head of the Preservation Research and Testing Division of the U.S. Library of Congress; Nigel Wilson of Lincoln College, Oxford University; and undergraduate and graduate students at the Chester F. Carlson Center for Imaging Science at the Rochester Institute of Technology, including Allison Bright, Kevin Bloechl, Elizabeth Bondi, Carolyn Houston, and Derek Walvoord.

REFERENCES

Albrecht, Felix (2012), 'Between boon and band: the use of chemical reagents in palimpsest research in the nineteenth century', *Care and Conservation of Manuscripts,* 13: 147–165.

Biot, J. B. (1840), reported in *Chronique* of Bibliothèque de l'École des Chartes I p. 408 (available from http://gallica.bnf.fr/ark:/12148/bpt6k123741.image).

Bloechl, K., Hamlin, H., and Easton, R. L., Jr. (2010), 'Text recovery from the ultraviolet-fluorescence spectrum for a treatise in the Archimedes palimpsest', *Proceedings of SPIE,* 7531-09.

Di Sarzana, Chiara Faraggiana (2006), 'La fotografia applicata a manoscritti greci di difficile lettura: origini ed evoluzione di uno strumento di ricerca e i principi metodologici che ne regolano l'uso', in Angél Escobar (ed.), *El palimpsesto grecolatino como fenómeno librario y textual*, 65-80 (available from http://ifc.dpz.es/recursos/publicaciones/26/54/_ebook.pdf).

Dold, Alban (1913), *Codex Sangallensis 193, Spicilegium Palimpsestorum Volumen I* (Leipzig: Harrassowitz).

Easton, R. L. Jr., Knox, K. T., Christens-Barry, W. A., Boydston, K., Toth, M. B., Emery, D., Noel, W. (2010), 'Standardized system for multispectral imaging of palimpsests', *Proceedings of SPIE*, 7531-12.

Edmonds, Tony E. (1998), 'An indicator of its time: two millennia of the iron-gall-nut test', *The Analyst*, 123: 2909–2914.

Gibson, Margaret D. (1893), *How the Codex Was Found* (Cambridge: MacMillan and Bowes) (available from https://archive.org/details/howcodexwasfound00lewi).

Harris, Helen B., and Harris, James R. (1891), *The Newly Recovered Apology of Aristides*, London: Hodder and Stoughton (available from https://archive.org/details/newlyrecoveredap00harr).

Haselden, Reginald Berti (1935), *Scientific Aids for the Study of Manuscripts* (Oxford: Oxford University Press for the Bibliographic Society).

Hyvärinen, Aapo and Oja, Erkki (2000), 'Independent Component Analysis: Algorithms and Applications', *Neural Networks*, 13: 411–430 (available from http://www.cs.helsinki.fi/u/ahyvarin/papers/NN00new.pdf).

Knox, K., Dickinson, C., Easton, R., Wei, L., and Johnston, R. (2001), 'Multispectral imaging of the Archimedes Palimpsest', *Proc. 54th IS&T Ann. Conf. PICS*, Montreal, 206–210.

Kögel, Raphael Gustav (1920), 'Die Palimpsestphotographie', *Enzyklopädie der Photographie*, 95 (available from http://goobipr2.uni-weimar.de/viewer/resolver?urn=urn:-nbn:de:gbv:wim2-g-2965569).

Kögel, Raphael Gustav (1914), 'Die Palimpsestphotographie', *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, 27: 974–978 (available from https://archive.org/details/sitzungsberichte1914deut).

Netz, Reviel, and Noel, William (2007), *The Archimedes Codex*, (London: Weidenfeld and Nicolson).

——, Noel, W., Tchernetska, N., and Wilson, N. (eds.) (2011), *The Archimedes Palimpsest* (Cambridge University Press).

Pringsheim, E., and Gradenwitz, O. (1894), 'Photographische Reconstruction von Palimpsesten', *Verhandlungen der Physikalischen Gesellschaft zu Berlin im Jahre 1893*, 58–60 (available from http://catalog.hathitrust.org/Record/000544812).

——, and Gradenwitz, O. (1901), 'Photographische Reconstruction von Palimpsesten', *Jahrbuch für Photographie und Reproduktionstechnik für das Jahr 1901*, 52–55 (available from https://archive.org/details/jahrbuchfrphoto04edergoog).

Rapantzikos, K., and Balas, C. (2005), 'Hyperspectral imaging: potential in non-destructive analysis of palimpsests', *IEEE International Conference on Image Processing (ICIP)*, 0-7803-9134-9/05/.

Rostagno, Enrico (1915), 'Della Riproduzione de' Palinsesti e d'un nuovo sistemo italiano ad essa applicato', *Rivista Delle Biblioteche,* 26, 58–67 (available from http://babel.hathitrust.org/cgi/pt?id=njp.32101073752162;view=1up;seq=68).

Schnauss, H. (1900), 'Photography as an Aid to Palaeography', *The American Amateur Photographer*, 12: 504–506 (available from Google Books).

Shlens, Jonathon (2009), *Tutorial on Principal Component Analysis* (http://www.snl.salk.edu/~shlens/pca.pdf).

Article

A Two-stage Approach to Segmentation-free Query-by-example Word Spotting

Leonard Rothacker, Marçal Rusiñol, Josep Lladós, and Gernot A. Fink | Dortmund – Barcelona

Abstract

With the ongoing progress in digitization, huge document collections and archives have become available to a broad audience. Scanned document images can be transmitted electronically and studied simultaneously throughout the world. While this is very beneficial, it is often impossible to perform automated searches on these document collections. Optical character recognition usually fails when it comes to handwritten or historic documents. In order to address the need for exploring document collections rapidly, researchers are working on word spotting. In query-by-example word spotting scenarios, the user selects an exemplary occurrence of the query word in a document image. The word spotting system then retrieves all regions in the collection that are visually similar to the given example of the query word. The best matching regions are presented to the user and no actual transcription is required.

An important property of a word spotting system is the computational speed with which queries can be executed. In our previous work, we presented a relatively slow but high-precision method. In the present work, we will extend this baseline system to an integrated two-stage approach. In a coarse-grained first stage, we will filter document images efficiently in order to identify regions that are likely to contain the query word. In the fine-grained second stage, these regions will be analyzed with our previously presented high-precision method. Finally, we will report recognition results and query times for the well-known George Washington benchmark in our evaluation. We achieve state-of-the-art recognition results while the query times can be reduced to 50% in comparison with our baseline.

1. Introduction

Text recognition in scanned documents usually refers to transcribing images showing text into a machine-based representation like ASCII or UTF-8.[1] Even though researchers have been investigating this topic for decades, only the recognition of clear machine-printed texts can be considered solved. For handwritten documents and historical manuscripts, results are far from satisfactory. Difficulties with handwritten documents are caused by the extreme variability in human writing. Historic documents often show severe degradation, such as ink bleed-through, bad contrast due to changes in paper color, text line deviations, old fonts and other artifacts caused by old technical standards and storage. Nevertheless, algorithmic methods for human assistance exist in order to explore these kinds of documents. They usually work in a relatively constrained scenario but will save a significant amount of work if they are applicable.

One of the most prominent techniques for this purpose is word spotting.[2] The task is not to transcribe entire images of text but to automatically detect regions in document images where the query word is likely to be found. The results are presented to the user in the form of a ranked list of these regions. In comparison with a full transcription of the document images, word spotting is much more robust against recognition errors. As long as the relevant regions are among the top ranks of the list, the user can eventually decide what he actually wants to use from the given results.[3]

Errors in a text transcription are not as easy to recognize and even small mistakes can corrupt further processing such as a full-text search.

In this scenario, the query is an exemplary occurrence of the respective word in the image. Query-by-example word spotting works only as long as the variability in the text is

[1] Cf. Doermann and Tombre 2014, chap. 8–14.

[2] Cf. Lladós et al. 2012.

[3] Rath and Manmatha 2007.

low, as in single-writer scenarios or documents printed in a single font. Apart from that, no prior knowledge of the problem domain is required.

A widely recognized approach to word spotting was presented by Rath and Manmatha. Their method follows a relatively classic pattern recognition pipeline.[4] A document image is first segmented into word regions. After skew and slant normalization, each region is represented by a sequence of feature vectors. Words that are similar to the query are then found by computing distances between feature vector sequences using Dynamic Time Warping.[5] An effect of a prior document segmentation is that subsequent processing steps can be specifically designed for working with word regions. Examples include specialized features, such as the upper and lower word contour or the number of ink background transitions along the word image's columns.[6]

However, while this may seem advantageous, a severe disadvantage is that such a system does not recover from segmentation errors. It is implicitly assumed that perfect segmentation is possible. If that assumption fails, these errors cannot be handled and all further steps will be based upon them. For instance, if the word segmentation fails, the information encoded in the feature representation will not be useful. In addition, the processing steps in such a pipeline are only able to work in a locally optimal manner and no information regarding the actual objective, i.e. the recognition, can be taken into account. When segmenting documents into lines or words, a recognition step is already incorporated because knowledge of their appearance is assumed. These steps are usually based on heuristics and, consequently, the recognition will also be based on heuristics. In challenging unconstrained handwriting recognition scenarios, it is impossible to rely on knowledge justifying those assumptions. Therefore, we propose to avoid early decisions and, rather, to integrate as much information as possible into the final recognition.

The first methods for segmentation-free word spotting were presented in Leydier et al.[7] and Gatos and Pratikakis[8]. Leydier et al. used a keypoint-based approach. Local zones-of-interest from the query word are matched with the most similar local zones in the document image. Regions are retrieved where the spatial configuration of matching zones in the query and the document image fit. Gatos and Pratikakis use a patch-based framework for segmentation-free word spotting. After preprocessing and normalization, they obtain text regions in a filtering step. Within these salient image regions, patches are sampled that are finally matched with the query. By over-segmenting the text regions, they are able to analyze all possible word locations.

In our previous work on segmentation-free word spotting,[9] we demonstrated how very accurate results can be achieved using a statistical sequence model. The model captures the spatial sequential structure of the query word in a dynamic probabilistic way. Similarity scores are then computed between the model and patches that were densely sampled over the document image.

However, a drawback of this method is the high computational effort that is required when processing a complete document image without focusing the recognition on potentially relevant regions as, for example, presented by Gatos and Pratikakis. A word spotting system that is not able to respond instantly is not likely to be accepted by potential users. In order to improve computational efficiency, we propose a two-stage approach. In the first stage, we identify potentially interesting, i.e. salient regions, in a fully segmentation-free manner. In this step, we do not incorporate any heuristic decisions but only identify regions in the document image that are visually similar to the query image. We implement this efficiently using an index structure for looking up small image patches in the document image that are similar to small image patches in the word region. This index is independent of a particular query and can be precomputed for any document image. In the second step, we investigate the salient regions in more detail. We apply our statistical sequence model for obtaining highly accurate matches. A major advantage of the approach is that both stages are based upon small image patches that can be found in the query and the document image. No additional preprocessing or feature extraction is required.

[4] Cf. Duda and Hart 2001, chap. 1.

[5] Cf. Rabiner and Juang 1993, 221-226.

[6] Marti and Bunke 2000.

[7] Leydier et al. 2009.

[8] Gatos and Pratikakis 2009.

[9] Rothacker et al. 2013; Rothacker, Rusiñol, and Fink 2013.

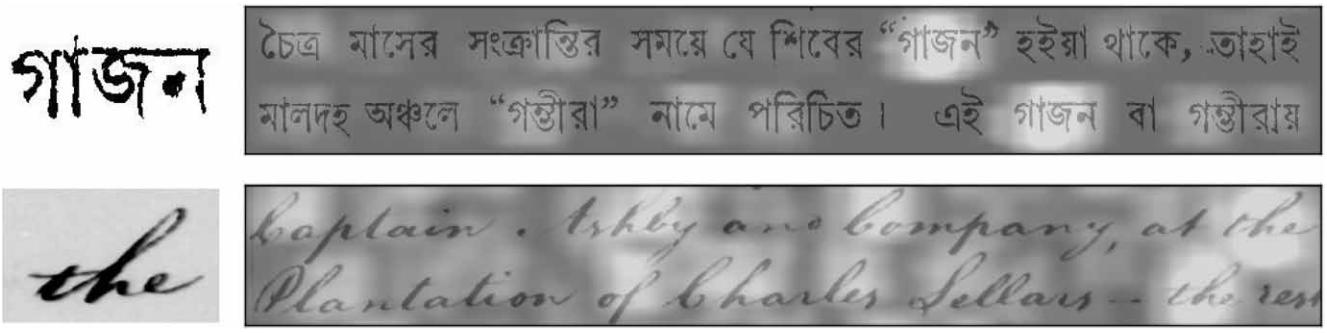


Fig. 1: Segmentation-free query-by-example word spotting. Top: Old printed Bangla book. Bottom: Historical George Washington dataset. Queries are shown on the left. On the right, the word occurrence probability is indicated by brightness.

2. Related work

A recent trend in document analysis is to adapt methods from computer vision. These algorithms usually have to work in completely unconstrained scenarios dealing with a huge variety of different objects and their appearances in natural scene images.[10] One of the best known approaches that has found its way into document analysis is Bag-of-Features (BoF).[11] These are statistically estimated feature representations that automatically integrate information from the problem domain, for example, document images. In contrast, a heuristically designed feature representation only captures the much more limited number of aspects that have been considered in the manual design process.

The basic idea behind BoF representations is to estimate feature representatives that are typical for the problem domain. When applied to images, gradient-based local image descriptors are very popular, for example SIFT descriptors,[12] which encode the local neighborhood at a given point in the image, i.e. a small image patch.

As the typical image descriptors are not directly available but have to be estimated statistically, three steps are required to compute a BoF. First, the typical image descriptors have to be estimated, usually by clustering a large set of descriptors from an image sample set.[13] Second, the descriptors in a given image have to be mapped to the most similar typical image descriptor.[14] Finally, the BoF is obtained as the histogram of typical image descriptor frequencies. One of the first and most widely accepted methods for BoF in computer vision was presented by Sivic and Zisserman.[15] They demonstrated how objects can be efficiently retrieved from a large number of movie frames. For this purpose, objects as well as movie frames are represented by BoF representations. Large-scale retrieval is possible because features from the query can be efficiently localized in each movie frame with an inverted file structure.[16] This index saves all locations of each feature in a frame.

With respect to word spotting, the methods of Shekhar and Jawahar[17], Rusiñol et al.,[18] and Almazán et al.[19] are especially relevant to our work. Shekhar and Jawahar use BoF representations for retrieving query words from large volumes of segmented word images. In analogy to Sivic and Zisserman, they use an inverted file structure for efficient indexing. The index stores pointers for each feature to all segmented word images that contain that specific feature. Given the features occurring in a query word image, the relevant word images in the database can be retrieved almost instantly. Rusiñol et al. were the first to apply BoF to segmentation-free word spotting. Almazán et al. showed how recognition accuracy and speed can be improved by using a Histogram-of-Oriented-Gradients in the same segmentation-free scenario. What both representations have in common is that no information about line and word locations is needed.

[10] Cf. Szeliski 2011, chap. 1.

[11] Cf. O'Hara and Draper 2011; Szeliski 2011, chap. 14.4.

[12] Lowe 2004.

[13] Cf. Gersho and Grey 1992, 362-370.

[14] Gersho and Grey 1992, chap. 10.

[15] Sivic and Zisserman 2003.

[16] Cf. Baeza-Jates and Ribeiro-Neto 1999, chap. 8.

[17] Shekhar and Jawahar 2012.

[18] Rusiñol et al. 2011.

[19] Almazán et al. 2012.

Instead, they are computed uniformly over the document image. While the BoF captures the simple occurrences of typical gradient-based local image features, the Histogram-of-Oriented-Gradients representation directly captures the image gradients in a grid of cells. For segmentation-free word spotting, the query image is transformed according to the respective feature description. This query feature description is then compared with patch feature descriptions that are densely extracted from the document image. This way, no prior assumptions about word positions have to be made but all possible locations are taken into account. An over-segmentation strategy of this kind is computationally very demanding. Rusiñol et al. reduced the search space by only considering a single patch size for all queries. Almazán et al. adapted the patch size to the query size. The huge amount of feature descriptions is efficiently compressed and stored in memory with product quantization.[20] Note, however, that limited variability of the script's visual appearance is assumed. Only a single patch size is used for spotting a certain query word. For that reason, it cannot be guaranteed that differently scaled instances of the query will be reliably found.

A drawback of the methods presented by Rusiñol et al. and Almazán et al. is limited flexibility with respect to modeling spatial gradient configurations. The more detailed the spatial information, the more specific becomes the feature representation of a query word image. While for single writer scenarios, a relatively explicit representation is advantageous, more abstraction will be needed once the variability increases. We, therefore, propose to integrate the BoF with a statistical sequence model, specifically Hidden Markov Models (HMM).[21] As shown in numerous examples, HMMs are able to model this spatial information in a dynamic probabilistic way.[22]

For a segmentation-free application, we adapted the patch-based framework presented in Rusiñol et al.[23] and Almazán et al.[24] but created a sequence of BoF representations from the query word image as well as from each patch. After the BoF-HMM was estimated from the query, we obtained a probabilistic similarity score for each patch position with Viterbi decoding.[25]

Fig. 1 illustrates this for two datasets that we used for evaluating the method: old printed Bangla documents[26] and historic documents handwritten by George Washington and his associates.[27]

The probabilistic score maps show that the variability in the printed document scenario is much smaller than in the handwritten document scenario. But even though the detections in the handwritten word spotting case are less prominent, their scores are still good enough to be considered as most important in the given example. In terms of word spotting accuracy, we clearly outperformed the results reported by Rusiñol et al. and Almazán et al. on the same benchmark. In terms of computational speed, however, the massive generation of BoF sequences from patch representations and their evaluation with the Viterbi algorithm was considerably slower.

In the remainder of this paper, we will present a two-stage method for improving computational efficiency by only applying the BoF-HMM at document image locations that are salient with respect to the query word. The query-specific saliency map is based on an inverted file structure. This way, locations of typical features in the document image can be efficiently indexed. The major difference in computing the saliency map compared with Gatos and Pratikakis[28] is that we already integrate information from the query instead of just looking for arbitrary text areas. This makes the salient regions specific to the query and we can apply the BoF-HMM in a more focused manner.

In our experimental evaluation of the George Washington benchmark,[29] we will show that speed-ups of more than 50% are possible while the word spotting accuracy is only marginally affected.

[20] Jégou, Douze, and Schmid 2011.

[21] Rothacker, Vajda, and Fink 2012; Rothacker, Rusiñol, and Fink 2013.

[22] Cf. Fink 2014, chap. 5.

[23] Rusiñol et al. 2011.

[24] Almazán et al. 2012.

[25] Cf. Fink 2014, chap. 5.6.

[26] Rothacker et al. 2013.

[27] Rothacker, Rusiñol, and Fink 2013.

[28] Gatos and Pratikakis 2009.

[29] Rusiñol et al. 2011; George Washington Papers at the Library of Congress.

3. A two-stage approach for segmentation-free word spotting

The sole application of the BoF-HMMs for segmentation-free word spotting is very costly with respect to its computational efficiency. Running such a high precision method for all densely sampled patches on the document image also does not always seem to be appropriate. Most patches are visually very dissimilar to the query, which raises the question if these patches can be rejected with a more efficient approach.

In this section, we will present a method that performs a segmentation-free coarse-grained analysis of the document image followed by a fine-grained application of the BoF-HMM. The coarse-grained analysis produces a saliency map identifying regions in the document image that are similar to the query. The detailed analysis in the second stage is only applied in the local neighborhood of these salient locations. For the overall segmentation-free property of the method, it is very important that both stages rely on the same local image features. This means that if there is no indication of the query word in some region in the saliency map, the BoF-HMM also produces low scores in this particular area. The same features that did not match with the query in the first stage will not start matching in the second stage. For this reason, we do not incorporate any explicit prior segmentation step based on heuristic assumptions, such as distances between words or lines in the document image.

The two-stage method for segmentation-free word spotting consists of three processing steps: document image representation, model estimation and model decoding. As the computation of local image features is independent of any particular query and the features are shared among the two stages, they have to be computed once for each document image. In order to search for a query, it has to be modeled with the BoF-HMM. This incorporates a statistical model estimation procedure. Once the query model is available, it can be used for retrieving relevant regions in document images. The decoding step consists of the coarse-grained and the fine-grained analysis stages. After efficiently identifying regions in the document image that are similar to the query, these salient areas are analyzed in more detail using the BoF-HMM. Finally, regions are ranked according to their similarity with the query and presented to the user. The overall process is visualized in fig. 2 for an exemplary document image section of George Washington's letters.[30]

3.1. Document image representation

Documents are represented by typical local image features that are extracted on a dense grid in the image. This is shown at the top of fig. 2. The descriptors that have been considered here (SIFT[31]) consist of histograms of oriented gradients and are intended for capturing the main directions of the pen stroke in the local neighborhood of a grid point. Each of the highly overlapping image features is an abstraction of the document's visual appearance representing mainly the information that is relevant for recognizing handwritten words. This has previously been demonstrated in many applications to document analysis.[32] The 'Dense Grid of Descriptors' in fig. 2 exemplarily shows the local neighborhood of a single image feature in the dense grid.

In the next step, typical image features are found with a cluster analysis using Lloyd's algorithm.[33] For this purpose, a sample set of local image features is required that is representative of the problem domain. In case of a word spotting system, all document images are known a priori and no unknown documents need to be considered. If new documents are added in the future, the cluster analysis can simply be repeated. The basic idea of the cluster analysis is to group similar image features into a given number of clusters. Each cluster has a representative that maximizes the average similarity to all its elements. Note that the representative is usually not among the set of local image features used in the cluster analysis. Finally, all descriptors in the dense grid are assigned to their most similar cluster representative, i.e., the typical image feature. This process is known as quantization. In fig. 2, the 'Descriptor Quantization' visualizes the dense grid with points. The points' colors indicate the typical image feature that they have been assigned to. Note how similar color patterns correspond to similar patterns of the pen stroke in the section of the document image. In the following, words will be spotted by the occurrence of typical image features in the document image that also appear in the query word region.

3.2. Model estimation

When users want to query the word spotting system, they must select an exemplary occurrence of the word in a

[30] George Washington Papers at the Library of Congress.

[31] Lowe 2004.

[32] Rusiñol et al. 2011; Rothacker, Rusiñol, and Fink 2013; Shekhar and Jawahar 2012; Lladós et al. 2012.

[33] Cf. Gersho and Grey 1992, 362–370.

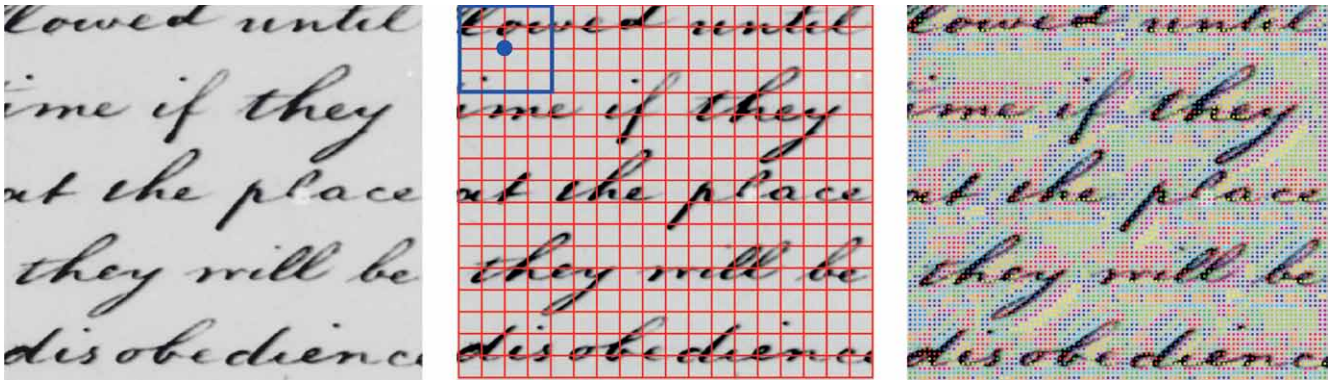


Fig. 2: Overview of the two-stage word spotting method. Top: Document image representation. Each document image is represented by typical local image features. In the dense grid these are indicated with different colors. Middle: Query word modeling and inverted file structure lookup & accumulation. Some accumulator cells show counts of how many features from the query model have been detected. Bottom: Coarse-grained saliency map and fine-grained analysis with the BoF-HMM. Saliency scores are indicated with blue to red colors. For the fine-grained analysis the BoF sequence extraction is visualized for all detected regions. The rank of the finally retrieved regions is indicated with blue to red colors.

document image. The respective region also specifies the typical image features that represent the query word. In order to capture the query word's spatial sequential structure, we extract a BoF representation for each column within the dense grid. The BoF only captures the relative frequency of each typical image feature, not their order.

This will be beneficial in the subsequent decoding step. The sequence of BoF representations is then modeled with a BoF-HMM. HMMs are generative finite state machines. At each point in time, one state is active and generates an output according to an underlying statistical process. In the estimation procedure, parameters are optimized with respect to the probability of generating this sequence of BoF representations from the query word. The process is referred to as the Baum-Welch algorithm.[34] In the query-by-example scenario, it is important that the BoF-HMM can be estimated with just a single sample.[35] This is usually impossible with continuous HMMs.[36] Fig. 2 exemplarily sketches the model estimation for the query word 'the'. Below the query word region, the extraction of the BoF sequence is visualized. A sliding window, shown in black, is moved over the dense grid in the direction of writing. At each window position, a histogram of typical image feature frequencies is created. For the window position, visualized in fig. 2, the purple feature has a higher frequency than the red and orange features. The BoF-HMM models feature probabilities for the query word directly within the states. In the given example, the states roughly represent the features occurring at the beginning, in the middle and at the end of the query word. For example, the cyan features have a higher probability in the middle, while the red features are rather probable at the beginning and the end.

3.3. Model decoding

After the model has been estimated, it can be used for retrieving regions in document images that are similar in terms of typical image feature occurrences. The overall process consists of two stages. In the first stage, a coarse-grained analysis is performed. It roughly identifies areas in a document image containing typical image features that also occur in the query word. In the second stage, these regions are explored in more detail by measuring their similarity to the BoF-HMM.

3.3.a First stage

The most important component in the first stage is an inverted file structure that we use for localizing typical features in the document image. A similar approach has been used for object detection and retrieving segmented word images. The inverted file structure index contains an entry for each typical image feature that has been determined in the cluster analysis. Each entry indexes all feature locations. The inverted file structure has to be computed only once and feature location lookups become very efficient afterwards. In our scenario, we want to look up features from the query word. Every feature with a non-zero probability in any of the states of the query model will be localized this way. For an actual occurrence of the query word in a document, we expect to find a larger number of detections in roughly the same area. We, therefore, accumulate detections in a cell structure over the document image. For each cell, the number of feature detections is counted. This is similar to the generalized Hough transform.[37]

In fig. 2, the inverted file structure lookup and accumulation is visualized for the exemplary query. All features occurring in the query model are localized in the section of the document image. Note that only a few features with respect to the total number of features will actually occur in a specific query word region. Also, not all features might reappear in a document. In fig. 2, this is indicated for the red feature. Inverted file structure lookups are very efficient for these reasons.

The resulting accumulator matrix can be interpreted as a coarse-grained saliency map. It is coarse-grained because no spatial relations between features have been taken into account yet. In fig. 2, saliency measures are visualized with blue to red colors. Due to the coarse-grained character, detections are usually not precisely located over the relevant regions.

Regions-of-interest that are positioned at the peaks in the saliency map are the final output of the first stage, i.e. a region is created for each locally optimal score. The regions' sizes are equal to the size of the query word. The locally optimal scores are determined in such a way that the regions do not overlap.

[34] Cf. Fink 2014, chap. 5.7.

[35] Rothacker, Rusiñol, and Fink 2013.

[36] Plötz and Fink 2011.

[37] Cf. Szeliski 2011, chap. 4.3.2.

Table 1: Evaluation of the two-stage approach for segmentation-free word spotting

Method	RoI dilation	Overlap threshold	mAP	mR	mQT	Speed-up
BoF	—	50%	30.4%	71.1%	340 ms	
HoG	—	50%	54.4%	—	15 ms	
VT	—	50%	69.7%	83.0%	4100 ms	Baseline
IFS	—	50%	50.1%	60.1%	380 ms	10x
IFS+VT	—	50%	60.4%	60.0%	540 ms	7x
IFS+VT	3x3	50%	64.9%	69.2%	980 ms	4x
IFS+VT	5x5	50%	69.6%	80.7%	1830 ms	2x
IFS+VT	7x7	50%	70.0%	82.5%	2750 ms	1.5x
VT	—	25%	71.4%	96.8%	4100 ms	Baseline
IFS	—	25%	48.0%	93.1%	380 ms	10x
IFS+VT	—	25%	53.9%	92.1%	540 ms	7x
IFS+VT	3x3	25%	64.7%	94.7%	990 ms	4x
IFS+VT	5x5	25%	70.9%	96.2%	1830 ms	2x
IFS+VT	7x7	25%	71.6%	96.3%	2760 ms	1.5x

3.3.b Second stage

The regions-of-interest from the first stage are analyzed in more detail in the second stage. A sequence of BoF representations is extracted from each region. This captures the spatial sequential structure in the writing direction. As no information about the feature order is encoded in a BoF, these representations are robust against smaller vertical displacements of the detected region with respect to the relevant region in the document. The horizontal displacements can be handled by the BoF-HMM, which models the sequential structure in a dynamic probabilistic way. The probability of generating the BoF sequence from a detected region can be computed with the Viterbi algorithm. Each region obtains a new similarity score with respect to the query model. However, even though the BoF-HMM is relatively robust against displacements of the detected regions, better scores can be obtained when these regions exactly fit the relevant regions in the document. For this reason, we dilate the search area locally in order to overcome the coarse detections of the first stage. Finally, we extract detected regions with locally optimal scores and rank them accordingly. Fig. 2 visualizes the fine-grained evaluation of the HMM and the retrieved regions that are presented to the user. The regions' scores are indicated with blue to red colors.

In the following evaluation, we will focus on performance measures in terms of the two stages. We will evaluate their individual and joint accuracy as well as their speed. The region-of-interest dilation will be of special interest with respect to the joint evaluation of both stages.

4. Evaluation

The effect of the two-stage method is evaluated using papers written by George Washington and his associates.[38] The full collection consists of over 65,000 documents and covers many aspects 'of colonial and early American history'.[39] Document types include correspondences, diaries, journals, military records, notes etc. that were collected by George Washington from 1741 to 1799. At the Library of Congress, the collection is organized accordingly into nine series. For the word spotting benchmark, a small dataset of 20 pages that are in overall good condition has been compiled from 'Series

[38] Rusiñol et al. 2011.

[39] George Washington Papers at the Library of Congress.

2: Letterbooks 1754–99' and consists of pages 270–279 and 300–309. The dataset was first used for evaluating a word spotting system by Rath and Manmatha.[40] In order to measure and compare segmentation-free word spotting performance, Rusiñol et al.[39] defined a benchmark containing 4,860 queries from these 20 pages. Ground truth annotations consisting of a bounding box in the document image and a word label are available for all words. The pages are written in an overall very similar style, thus the benchmark can be considered as a single writer scenario. Following this evaluation protocol, we use every word as a query without any further modification, like filtering short words or stemming words. For each query, we retrieve a ranked list of regions throughout all 20 pages. By comparing each of those regions with the ground truth annotations, we can decide if a single region is relevant with respect to the query or not. A region is considered relevant if it overlaps with a bounding box from the ground truth by more than a given threshold and the corresponding word label matches the query word. The choice of the overlap threshold is critical for performance measures and we will report results for two different values in the following. In table 1, we refer to this as the 'Overlap threshold'.

Given the list of relevant and non-relevant regions, two aspects are of prime importance to users of word spotting systems. All relevant regions should be ranked first and the list should contain all relevant words.[41] The first requirement is measured by average precision, while the second requirement is measured by recall. An average precision of 100% refers to a list where all relevant regions are listed first. A recall of 100% refers to a list that is complete, i.e., no relevant regions are missing. In order to report results over all queries, we compute means over the individual results, thus mean average precision and mean recall. In table 1, we refer to these measures as 'mAP' and 'mR'.

The motivation for the two-stage approach is to improve the computational speed. Mean runtimes for retrieving a single query on a single page are reported as well as the relative speed-ups with respect to our baseline method.[42] Results have been measured on a Xeon 3.0 GHz. In table 1, we refer to the mean query time per page as 'mQT' and to its relative improvement as 'Speed-up'.

Our experiments focus on the results obtained individually and jointly with the stages. Additionally, we report results for different overlap thresholds. As the first stage is based on an inverted file structure, it is referred to as 'IFS' in table 1. Analogously, scores in the second stage are computed with the Viterbi algorithm, thus referring to it as 'VT'. For the individual application of the first stage (IFS), we directly use the regions-of-interest as the output of the word spotting system. For the individual application of the second stage (VT), we use the same approach as presented in Rothacker, Rusiñol, and Fink,[43] i.e. we densely sample patches that are all decoded with the Viterbi algorithm for obtaining similarity scores. Please note that the results reported differ due to some parametric optimizations. In the joint evaluation of both stages, the region-of-interest dilation is important. In table 1, the different dilation masks can be found in the column 'RoI dilation'. A dilation mask of 3x3 refers to extending the search area to all neighboring cells in the accumulator matrix from a detected region in the first stage. Larger masks extend the search area further.

Starting with the results obtained for our baseline system (VT),[44] we observe an upper boundary for retrieval accuracy. This is expected as we are applying a fine-grained analysis throughout entire document images. When, in contrast, only the coarse-grained analysis in first stage is applied, a significant drop in mean average precision can be observed. As the spatial sequential structure of the query word is not modeled in the coarse-grained analysis, the simple occurrence of features from the query words is already a strong indication for a relevant region. Problems occur, for example, when single or multiple characters from the query word appear within other, typically longer words. When putting both stages together, we first rerank the results with the Viterbi algorithm without dilating the search area. While the mean recall stays constant, there is a considerable increase in mean average precision. This nicely demonstrates the effect of adding spatial information to the query word modeling. The constant mean recall is due to the fact that the search space has not been extended in the second stage. The list of relevant regions stays the same.

When increasing the search area, improvements can be observed for both mean average precision and mean recall. Now, the regions-of-interest can be positioned better over

[40] Rath and Manmatha 2007.

[41] Cf. Baeza-Yates and Ribeiro-Neto 1999, chap. 3; Lladós et al. 2012, 13–15.

[42] Rothacker, Rusiñol, and Fink 2013.

[43] Ibid..

[44] Ibid.

the relevant regions in the document image and the BoF-HMM produces better matches when being evaluated with the Viterbi algorithm. Regions-of-interest that overlap only slightly with relevant regions in the document are discarded for better matches in the local neighborhood. Additionally, more regions will be regarded as relevant if the detections from the first stage were not precise enough to produce sufficient overlap percentages.

Regarding the overlap threshold, a reduction from 50% to 25% shows that over 90% of the relevant words can roughly be located in the first stage. The localizations are just too imprecise to produce over 50% overlaps. This nicely motivates the application of the regions-of-interest dilation.

Finally, we measured strong improvements in the mean query time with respect to our baseline system. The sole application of the inverted file structure is more than ten times faster. Using the two-stage approach, the size of the search area (RoI dilation) is strongly related to the mean query time and recognition accuracy. When matching the accuracy of the baseline system, a speed-up by a factor of two is still possible.

When comparing our results with BoF-based results reported in Rusiñol et al.[45] and the Histogram-of-Oriented-Gradient (HoG)-based results reported in Almazán et al.,[46] we clearly outperform their recognition accuracy with our integrated approach. Low recognition scores in Rusiñol et al. are due to the uniform patch size used for all queries in the retrieval stage. Small words that are hard to detect because patches contain a lot of context in the document image besides the relevant word. Results reported in Almazán et al. show good recognition accuracy and very fast retrieval times. The HoG features nicely encode handwritten words in the single-writer scenario. In comparison with our two-stage approach, we see two important aspects. Encoding full-page HoG representations in memory does not scale for very large collections of document images. By contrast, the inverted file structure in our coarse-grained stage is able to rapidly reduce the search space over all document pages in the collection. This capability has already been demonstrated for segmented word images.[47] Finally, the application of Hidden Markov Models offers more flexibility, especially when more than a single sample of a query is available. Please note that the query times for neither Rusiñol et al. nor Almazán et al. are directly comparable because different machines have been used in the evaluation. However, they are suitable for showing the general order of magnitude in which the methods operate.

5. Conclusion

In this paper, we presented a two-stage approach for segmentation-free word spotting based on the George Washington benchmark. With respect to our baseline system, considerable improvements in computational speed have been observed. With the sole application of the coarse-grained first stage, speed-ups of more than ten times are possible. This comes at the cost of decreased recognition accuracy. When adding the fine-grained word spotting stage, the benefits of integrating a statistical sequence model can be demonstrated. The recognition accuracy increases substantially. Computational speed can still be reduced to 50% without loosing any precision. In comparison with Rusiñol et al.[48] and Almazán et al.,[49] more accurate results can be achieved at the cost of higher query execution times.

Furthermore, we present the effect of different overlap thresholds for the evaluation. We can show that the fine-grained stage handles regions-of-interest having smaller overlaps with the ground truth very well. For larger dilation masks, very high mean recall scores can be achieved without any loss in mean average precision. We doubt that an overlap threshold of 25% or 50% will make any difference for users of word spotting systems as detected regions will usually be presented in a larger context in the document image.

In our future research, we will work on better region localizations in the coarse-grained stage. This could be achieved by taking some limited amount of the query word's spatial structure into account.

ACKNOWLEDGMENTS

This work is partially supported by the German Academic Exchange Service on the basis of a DAAD-Doktorandenstipendium and by the Spanish project TIN2012-37475-C02-02.

[45] Rusiñol et al. 2011.

[46] Almazán et al. 2012.

[47] Shekhar and Jawahar 2012.

[48] Rusiñol et al. 2011.

[49] Almazán et al. 2012.

REFERENCES

Almazán, J., Gordo, A., Fornés, A., and Valveny, E. (2012), 'Efficient exemplar word spotting', *Proceedings of the British Machine Vision Conference*, 67.1–67.11.

Baeza-Yates, R., and Ribeiro-Neto, B. (1999), *Modern Information Retrieval* (Addison Wesley).

Doermann, D., and Tombre, K. (2014), *Handbook of Document Image Processing and Recognition* (Springer).

Duda, R. O., Hart, P. E., and Stork, D. G. (2001), *Pattern Classification*, 2nd edition (Wiley).

Fink, G. A. (2014), *Markov Models for Pattern Recognition, From Theory to Applications*, 2nd edition (Springer; Advances in Computer Vision and Pattern Recognition).

Gatos, B., and Pratikakis, I. (2009), 'Segmentation-free word spotting in historical printed documents', *Proceedings of the International Conference on Document Analysis and Recognition*, 271–275.

George Washington Papers at the Library of Congress, 1741–1799, Manuscript Division, Library of Congress, Washington, D.C. http://memory.loc.gov/ammem/gwhtml/gwhome.html.

Gersho, A., and Grey, R. M. (1992), *Vector Quantization and Signal Compression* (Kluwer Academic; Communications and Information Theory).

Jégou, H., Douze, M., and Schmid, C. (2011), 'Product quantization for nearest neighbor search', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33: 117–128.

Marti, U. V., and Bunke, H. (2000), 'Handwritten Sentence Recognition', *Proceedings of the International Conference on Pattern Recognition*, 3: 463–466.

Leydier, Y., Ouji, A., LeBourgeois, F., and Emptoz, H. (2009), 'Towards an omnilingual word retrieval system for ancient manuscripts', *Pattern Recognition*, 42.9: 2089–2105.

Lladós, J., Rusiñol, M., Fornés, A., Mota, D. F., and Dutta, A. (2012), 'On the influence of word representations for handwritten word spotting in historical documents', *International Journal of Pattern Recognition and Artificial Intelligence*, 26.5.

Lowe, D. (2004), 'Distinctive Image Features from Scale-Invariant Keypoints', *International Journal of Computer Vision*, 60.2: 91–110.

O'Hara, S., and Draper, B. A. (2011), *Introduction to the Bag of Features Paradigm for Image Classification and Retrieval*, (Cornell University Library, http://arxiv.org/abs/1101.3354).

Plötz, T., and Fink, G. A. (2011), *Markov Models for Handwriting Recognition* (SpringerBriefs in Computer Science, Springer).

Rabiner, L. R., and Juang, B.-H. (1993), *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs).

Rath, T., and Manmatha, R. (2007), 'Word spotting for historical documents', *International Journal on Document Analysis and Recognition, 9.2–4: 139–152.*

Rothacker, L., Fink, G. A., Banerjee, P., Bhattacharya, U., and Chaudhuri, B. B. (2013), 'Bag-of-Features HMMs for Segementation-Free Bangla Word Spotting', in I*nternational Workshop on Multilingual OCR*, article no. 5.

——, Rusiñol, M., and Fink, G. A. (2013), 'Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents', *Proceedings of the International Conference on Document Analysis and Recognition*, 1305–1309.

——, Vajda, S., and Fink, G. A. (2012), 'Bag-of-Features Representations for Offline Handwriting Recognition Applied to Arabic Script', *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 149–154.

Rusiñol, M., Aldavert, D., Toledo, R., and Lladós, J. (2011), 'Browsing heterogeneous document collections by a segmentation-free word spotting method', *Proceedings of the International Conference on Document Analysis and Recognition*, 63–67.

Shekhar, R. and Jawahar, C. (2012), 'Word image retrieval using bag of visual words', *International Workshop on Document Analysis Systems*, 297–301.

Sivic, J., and Zisserman, A. (2003), 'Video Google: A text retrieval approach to object matching in videos', *Proceedings of the International Conference on Computer Vision*, 2: 1470–1477.

Szeliski, R. (2011), *Computer Vision, Algorithms and Applications* (Springer).

Article

The Evolution of Imaging Techniques in the Study of Manuscripts

Athina Alexopoulou and Agathi Kaminari | Athens

1. Introduction

This paper outlines the evolution of non-destructive testing in the study of manuscripts. This overview starts with the time when manual SLR cameras, photographic films and huge infrared analogue tubes were the cutting edge. It arrives at current practice, in which extremely high-resolution DSLR cameras and other digital camera systems coupled with hyper-spectral imaging approaches are used for acquiring image sequences in different spectral regions. The term 'non-destructive' refers to those techniques – mainly imaging – that provide information without invasive activities, as opposed to chemical analysis, which requires sampling the object.

Moreover, modern technology enables the collection and distribution of large volumes of information that can now be analysed intuitively and quickly and assessed with the help of continuously developing and improving IT systems and software packages. It has been known since 1972 that certain old manuscripts become more legible through infrared photography. Earlier use of infrared photography and current use of multi-spectral imaging[1] have become important tools in non-destructive testing of all those objects of an artistic, archaeological or forensic nature that have the characteristics of writing or sketch work, e.g. various inscriptions and manuscripts made with specific types of inks, pencils, coal, etc. on different substrates from papyrus to industrial paper. Although primarily used for detecting underdrawings in paintings, these methods have been successfully applied to the documentation and scientific investigation of documents as well as the evaluation and monitoring of conservation treatments.[2] The term 'hyperspectral imaging' was coined in response to the continuous increase in the number of spectral bands available to multi-spectral cameras.

Representative case studies, selected from a large number handled by the Laboratory of Physical and Chemical Methods for Diagnosis and Documentation at the Department for Conservation of Antiquities and Works of Art, TEI, Athens, during its 25 years of operation, are presented here to illustrate and highlight the effectiveness of these methods. The comparison does not aim to pinpoint the best instrumentation for any given case, but rather to more generally consider the overall advantages, disadvantages and difficulties currently encountered and to suggest the best means of overcoming them.

The documentation of Heinrich Schliemann's copybooks, the investigation of Nikolaos Gyzis' oil sketches on paper and, most recently, the deciphering of the papyrus text from the 'Musician's Tomb' in Daphne, Greece, will be presented as examples.

The use of traditional colour photography (in normal and macro mode), ultraviolet reflection, fluorescence photography and infrared reflectography will be presented as basic steps in an artefact's documentation. Finally, false colour infrared imaging and hyperspectral imaging in the range of 420–1000 nm under normal and raking light combined with simple subtraction algorithms and principal component analysis (PCA) will be assessed as one of the next technological steps in the non-destructive examination of works of art.

2. Case study of Heinrich Schliemann's Copybook Archive

Heinrich Schliemann (1822–1890) was a well-known archaeologist who is particularly celebrated for his excavations at Troy and Mycenae. He was one of the first archaeologists to keep a meticulous archive of data from his excavation sites. His interest in technological advances led him to adopt James Watt's method to facilitate copying his extensive correspondence in a more efficient manner.

[1] Liang 2011; Fisher and Kakoulli 2006.

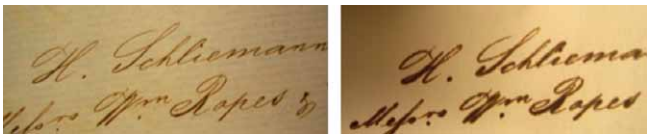
[2] Padoan et al. 2008; Chabries et al. 2003; Banou et al. 2010.

James Watt's copying method, which he patented in 1780, consisted of off-setting the ink from an original document onto a thin, unsized, dampened tissue paper after being pressed with a screw press. The original text was written with a variation of iron-gall ink, which contained Aleppo galls, green vitriol, gum arabic, roach alum, spring water as well as a medium such as glycerin or sugar to prolong drying of the ink to enable copying. The text was transferred as a mirror image to the tissue paper on contact, while the ink penetrated the core of the paper to provide a readable copy on the reverse side (fig. 1).

In situ research carried out in 2004 within the framework of the Archimedes I research programme in public and private archives in Greece[3] traced and examined more than 180 letter copybooks[4] spanning the period 1845–1929. These were used by institutions, shops, businesses and banks in Athens and its surrounding region as well as by private individuals with systematic and extensive outgoing correspondence, among them Heinrich Schliemann, Ioannis Valaoritis, the Greek Olympic Committee, the Byzantine Museum, the National Bank of Greece, the mines of Serifos island as well as shops in Athens and on the island of Kastelorizo. The use of letter copybooks was progressively abandoned after 1915, although examples can be found as late as 1930.

Among the various collections of letter copybooks located in Greek archives, research focused on the Schliemann Archive housed in the Gennadius Library, a part of the American School of Classical Studies in Athens. Only 43 volumes of this extensive copy-book archive survived the First and Second World War and are now included in the comprehensive collection of his private papers. This archive covers the period from 1845 to 1890, and the copies display a wide variety of distinct reproduction techniques and materials that reflect the German archaeologist's special interest in this type of technological development.[5]

From 1845 to 1867 Schliemann copied his letters using loose numbered sheets of copy and accumulated them until they were of sufficient quantity to be bound into books. From 1867 to 1890 he used commercial letter copybooks for this purpose.



original copy

Fig. 1: H. Schliemann's signature in an original and a copy letter.

The study of selected pages from the copybooks was carried out using imaging techniques to record the optical behaviour of the inks at different wavelengths (ultraviolet, visible and near-infrared) and, in this way, document the condition of the copy letters, especially their clarity and readability, and identify the different types of iron-gall copy inks. The methods applied were visible light photography (Vis), ultraviolet reflectance photography (UVR) at 365 nm emission wavelength, ultraviolet fluorescence photography in both black-and-white (UVF) and colour (UVFC), as well as infrared reflectography (IR-Ref) using a digital infrared detector CCD with sensitivity up to 1,250 nm.[6]

Methods were implemented according to protocols based on instrumentation and procedures available at that time, i.e. SLR cameras, optical filters, films and lighting.[7] For better assessment and data management, negatives and colour slides were scanned into digital images in .tiff format[8] immediately after their development.

The great advantage of the equipment used at that time was the capability of black-and-white films to record surface reflectance at 365 nm, monochromatic UV radiation emitted by black-light Philips lamps. Furthermore, film characteristics

[3] Alexopoulou et al. 2006.

[4] The copybooks discussed in this paper should not be mistaken for the copybooks employed by calligraphers and teachers for step-by-step exercises to form capital and small letters of the alphabet.

[5] ASCSA 2005.

[6] Alexopoulou et al. 2012.

[7] The instrumentation for the UV photography included a Canon T70 photographic camera with programmable back for the exposure time and a Canon 50 mm macro lens with an extension tube in order to boost magnification up to 2:1, two Philips HPW 125W 120V and Philips MLW 125W 220V UV black-light lamps as light sources together with the appropriate reflectors, plus the barrier filters Kodak 2E for fluorescence and Kodak 18A for reflection. Images were captured with Kodak Technical Pan film, FujiFilm Provia 100 and Kodak Ektachrome EPR 64 respectively. Visible light photographs were acquired using a Nikon F80 camera with a Nikon AF micro Nikkor 60 mm lens, two Osram 500 Watt tungsten lamps as light sources, two stands with reflectors, Kodak Ektachrome 64T films and the Cokin 82B colour-compensating filter. Exposure time was determined after bracketing. All black-and-white films were developed under the same conditions by the researchers. Infrared reflectography was carried out with a digital infrared camera (CCD, ARTI S.p.A. with a Pentax 50 mm lens) and two Osram 500 Watt tungsten lamps as light sources and two stands with reflectors. The B+W 093 IR-transmitting filter and the B+W 489 heat-absorbing filter were used for visible light and infrared imaging respectively.

[8] Scanning was performed with a Mikrotek ScanMaker 4900 4,800 x 2,400 dpi CCD, 48-bit colour scanner at 1,200 dpi.

in combination with exposure parameters made it possible to control the quality (contrast, value and sharpness) of the primary image in order to achieve optimum results. Even though DSRL cameras now offer flexibility in recording parameters, they still yield poor results for ultraviolet reflection as they are not sensitive to this type of radiation.

The study of Schliemann's letter copy archive mainly focused on the condition of the inked areas, the related phenomena of oxidation and diffusion, and the quality of the copying process.

Ink diffusion is a physical phenomenon that results from blurred ink strokes and can occur during the copying process. It relates to properties of the copy ink and paper, but is mainly due to poor application of the method, usually because of excess moisture, which causes lateral migration of the ink. Oxidation, on the other hand, is a chemical phenomenon due to the migration of the iron ions present in the ink around the ink strokes, which results in degradation of the paper matrix. The stages of degradation range from the formation of a halo around the lines of the ink strokes to expanding discolouration with gradual loss of the paper's mechanical strength around the inked areas and, finally, extensive loss of structural support within the writing itself.

Successful application of the copying method should produce a result comparable to authentic writing while leaving the original document intact. However, copies present differences in quality, characteristics and aging behaviour. In general, copies fall into four categories: (a) excellent, i.e. almost indistinguishable from the originals, (b) good, clear writing with satisfying legibility, (c) bad, faint and uncertain legibility, and (d) poor, with blurred letters and/or seriously pronounced diffusion of the ink, rendering the texts illegible. The structural sensitivity and chemical instability of the copybook materials have challenged archivists, curators and conservators concerning the assessment of their condition as well as the preservation and stability of their data over time.

After the visual comparison of all images obtained at 365 nm (UV), in the visible range of 380–760 nm and in the near-infrared range of 760–1,200 nm, four main categories were identified (table 1).

The first category consists of copies in which the ink and writing appear to be in very good condition. The original writing characteristics have been maintained. Neither intense oxidation nor diffusion in the inked areas nor fading are observed. It should be noted that the phenomena of oxidation and diffusion appear similar in visible light.

The second category consists of copies in which the halo surrounding the inked areas is so extended that the original ink strokes are indistinguishable from the halo. None of the non-destructive methods used practically aided differentiation of the letters from the halos. The IR reflectogram shows high transparency. Only UVFC records a light fluorescence in the perimeter of the letter. This may be the case where the halo is due to ink diffusion during the copying process.

The third category consists of copy letters with thick writing where an intense halo around the writing can be observed. This is probably due to oxidation because the core of the letter can be distinguished clearly from the circumferential oxidation in the visible range and in UV fluorescence colour photography. Only the core is visible in the infrared range.

This last category consists of copy letters in which several areas of diffusion and oxidation may be observed around the letters. A diffusion zone, a thin borderline and a second zone due to oxidation are observed beyond the distinct letter core.

In this case, visible and ultraviolet image capture can make a notable contribution to identifying a distinction between the original ink strokes and the diffusion zones. In the infrared reflectogram, the core of the letter and the pale diffusion around it are evident, while oxidation is not recorded. This, then, is the most helpful category for distinguishing oxidation from diffusion since it consistently presents different and separable visual results in the infrared range.

The methodology which was implemented using the technical facilities of that time helped in reading the texts, distinguishing between oxidation and diffusion phenomena that are quite often present at the same time in the writing, and in providing indications of the quality of ink, its components and, in some cases, the types of inks employed.

3. The Ultraviolet Fluorescence Colour Imaging problem

An important problem, which has arisen from films currently falling into disuse, is the difficulty of obtaining reproducible and reliable images with UV fluorescent stimulation using DSLR cameras. This problem will be presented in more detail in the following discussion.

In the above research, all the ultraviolet fluorescence colour photographs were obtained on film. This method has been studied thoroughly and is thus understood and predictable in terms of its consistent repeatability, so the resulting images maintain a high degree of consistency.

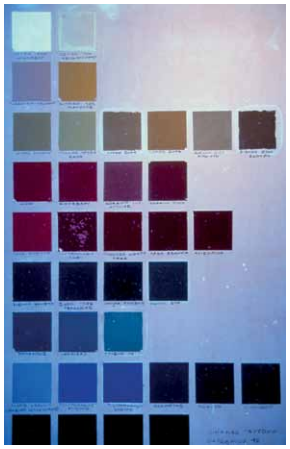
Table 1: Images of the four categories of copy letters at different wavelengths.

As technology advanced and digital cameras became commonplace in visible-light photography, the question arose whether they could be used as reliably and predictably for UV fluorescence colour images. To settle this question, the Laboratory of Physical and Chemical Methods for Diagnosis and Documentation, TEI, Athens, which emphasizes non-destructive testing, conducted a series of experimental comparisons of conventional versus digital captures. These tests were carried out in co-operation with the National Gallery – Alexandros Soutzos Museum (NGA) and the General State Archives (GSA). The aim of these experiments was to closely compare digital images produced by DSLR cameras to matching ones on film produced by conventional SLRs. The question was whether one could ensure the same quality with the DSLR output regarding predictable reproducibility and reliability without any further image post-processing as one could do with film images taken by SLRs.

For this preliminary experiment, three different light sources were tested: Philips MLW125 light bulbs at 365 nm [Philips 2005], Sylvania black-light blue tubes 36W and Osram black-light blue tubes 36W. As for digital cameras, the experiments had to be carried out using the DSLRs already available: Nikon D70s with an AF micro NIKKOR 60 mm lens, Nikon D70s with an AF-S DX Zoom-Nikkor 18–70 mm IF-ED lens, Canon EOS 5D Mark II with a Canon EF100 USM 100 mm lens, and Pentax *istDL with an smc Pentax DA 18–55mm lens. Auto-focus was chosen for all cameras.

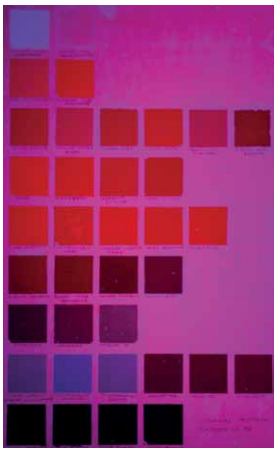
A specific test panel was employed as the common target for the shootings: a wooden substrate with a layer of inorganic material (gesso) mixed with organic binding medium (rabbit-

Table 2: Best images for each detector (f5.6, varied speed, Kodak 2E, black-and white 489, white balance [Kodak white card]).



a) Film + SLR

b) Nikon D70s



c) Pentax *istDL

d) Canon EOS 5D MarkII

skin glue) bearing thirty-six tiles 3 cm x 3 cm in size, each with a different pigment. The pigments were chosen based on the fact that they were not only widely employed in works of art, but also well understood and well documented in the international literature as yielding a known and consistent result on film when made to fluoresce in colour by known light sources such as the ones chosen above.

Different white balance settings (cameras' pre-sets of incandescent, fluorescent and daylight conditions or custom setting on different surfaces, i.e. photographic white and grey cards) were tested to see if they influenced the image quality in any way.

These conditions were tested using the Kodak 2E filter, which cuts off ultraviolet radiation as well as excess blue radiation and is considered a standard procedure for conventional analogue images on film. A black-and-white 093 (equivalent to Kodak Wratten 87C) visible-radiation (up to 780 nm) cut-off filter was also employed to test whether infrared radiation affected the images. Both CCD and CMOS sensors in digital cameras are considered sensitive in the near-infrared range despite there being filters incorporated in their main body to cut off unwanted IR radiation from being recorded. It has often been reported that these cameras are not completely free from infrared interference. The outcome necessitated further testing with an IR cut-off filter, the black-and-white 489, to improve image quality.

It soon became clear that, even when using extra filters, a small amount of infrared radiation could not be avoided, which therefore affects the final image to a greater or lesser extent, depending on the camera. It should be noted that reversal (i.e. slide transparency) films, when employed in conventional analogue (film) cameras, are not sensitive to infrared radiation and thus produce IR-free images.

After thorough experimentation and readjustment of the image acquisition parameters, the DSLR camera equipment (Nikon D70s/AF micro NIKKOR 60 mm lens, Nikon D70s/AF-S DX Zoom-Nikkor 18-70 mm IF-ED lens, Canon EOS 5D Mark II/Canon EF100 USM 100 mm lens and Pentax *istDL/smc Pentax DA 18-55 mm lens) exhibited different behaviour in UVFC photography than the equivalent SLR-and-film system (table 2). Furthermore, even though the DSLR cameras were used under the same conditions, the resulting images were not reproducible. Nonetheless, by using the final experimental combination,[9] it is possible to study fluorescence by achieving an image quality that at least differentiates the fluorescent colour between pigments.

4. Case study of Nikolaos Gyzis' sketches and signatures

Nikolaos Gyzis is considered to be one of the most important Greek painters of the 19th century. He initially studied painting in Athens and later in Munich, where he worked under Hermann Anschütz and Alexander von Wagner at the Academy of Fine Arts. In 1868, he was accepted into Karl von Piloty's class. Gyzis is a dominant figure in the 'Munich School' movement. He not only influenced the course of Greek art with his painting, but also occupies an important place in the history of 19th century German art. His versatile artistic personality promoted the creation of a

[9] Best results for DSLR detectors: Nikon D70s, f5.6, varied speed, Kodak 2E, black-and white 489, white balance (Kodak white card).

Fig. 2: 'Spirit', Nikolaos Gyzis, bilateral, head and torso of a young boy, possibly an angel (both sides). Late 19th c.

complex opus that transcended the level of mere narration, with his true merit showing in his idealistic, allegorical and religious work – fields in which he was in step with the aims of the *Jugendstil* avant-garde movement. In addition to genre painting, he worked on still lifes and portraiture as well as producing idealistic allegorical subjects, ethnographical scenes and poster art. He proved that he could master a broad range of artistic expression to further his personal ideas and visions.

Two aspects of his work were studied in our imaging analysis of Gyzis' painting: his sketches and his signatures.

Nikolaos Gyzis' oil painting commissions were usually quite large in size, so it was customary for him to prepare sketches using oil-on-paper substrate in order to visualize what he had in mind. Oil-on-paper is very delicate and requires careful conservation treatment in order to ensure an optimal state of preservation. As part of the Costopoulos Foundation Research Programme (2010) and the Archimedes III Research Funding Programme (2012–14), oil-on-paper paintings of Nikolaos Gyzis were examined in order to study the behaviour of oil in interaction with the paper substrate and whether oxidation of the paper is enhanced or suppressed by the oil medium.[10] Several non-destructive examination techniques were employed, but the following example merits particular attention.

The oil sketch 'Spirit', catalogue number Π588, which belongs to the main collection of the National Gallery – Alexander Soutzos Museum, attracts special interest for two reasons: not only does it show links to other important works of the artist, but it is also bilateral in presentation (fig. 2). In this work, Gyzis depicts the head and torso of a young boy on both sides of the paper support. On the recto, he sketched a child figure, which resembles an angel, exclusively by applying coloured oils, mostly red, blue and brown, both in thin (diluted) and thick layers. On the verso, he depicts much the same figure using some kind of pencil. The head is detailed while the torso is executed in a somewhat more simplified and impressionistic fashion. Comparative study of the two sides of the work has shown that the painted figure corresponds to the black-and-white sketch. Although there

[10] Banou et al. 2010.

Fig. 3: 'Spirit', Nikolaos Gyzis, detail, images at different wavelengths.

are a few differences in the movement of the hand as well as the position of the head and the hair, the two images seem to be intentionally connected. However, each image exhibits a number of specific stylistic characteristics. Based on observations of published pictures of Gyzis' works, charcoal and chalk portraits are usually detailed, expressive and atmospheric with emphasis on light denoted by the use of white chalk. By contrast, oil sketches are more freestyle, without details or strictly defined forms. They are executed naturally and unconstrained by stricter artistic boundaries.

Hyperspectral imaging and false colour infrared imaging[11] were applied to both sides of the painting (fig. 3). These images, combined with UVR and UVFC recordings, permitted differentiation of the different types of materials the artist used for the sketch: the intense black strokes on the eyelids, the mouth and the left side of the head (relative to the viewer) exhibit strong ultraviolet absorption (black) and infrared reflectance (soft grey to white) in contrast to the grey-black drawing material the artist employed in the remainder of the sketch, which presents absorption (dark grey) at all wavelengths. According to Attas,[12] black and dark drawing materials exhibit different behaviour in the infrared range. For example, carbon black absorbs light strongly and is therefore darker than graphite, which exhibits medium

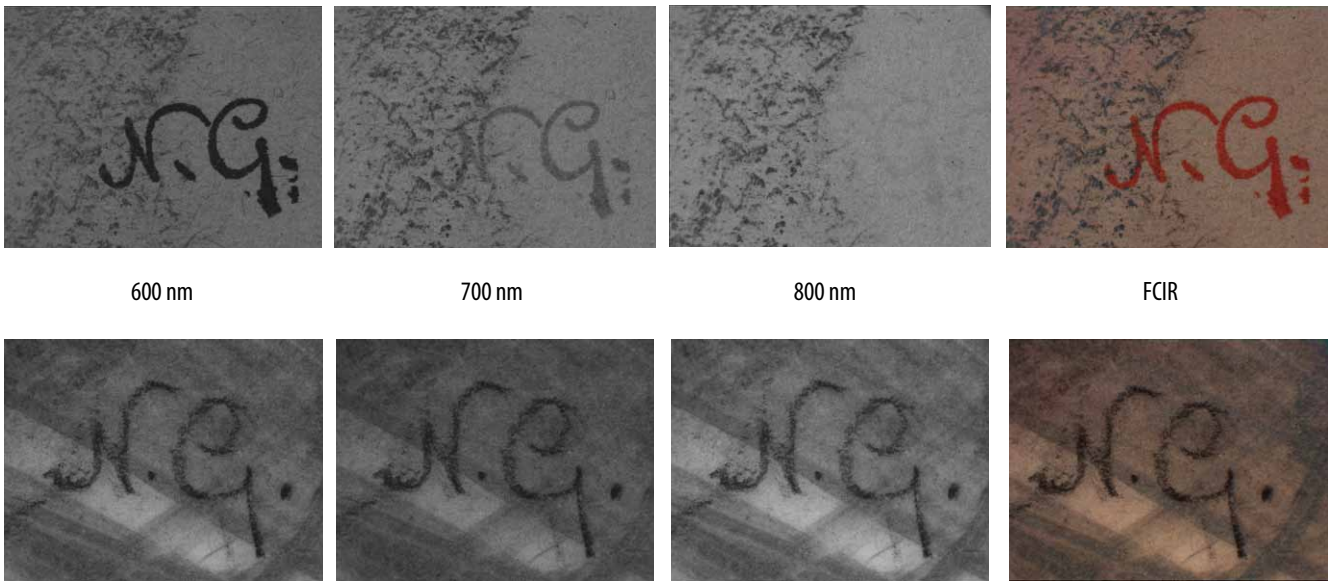
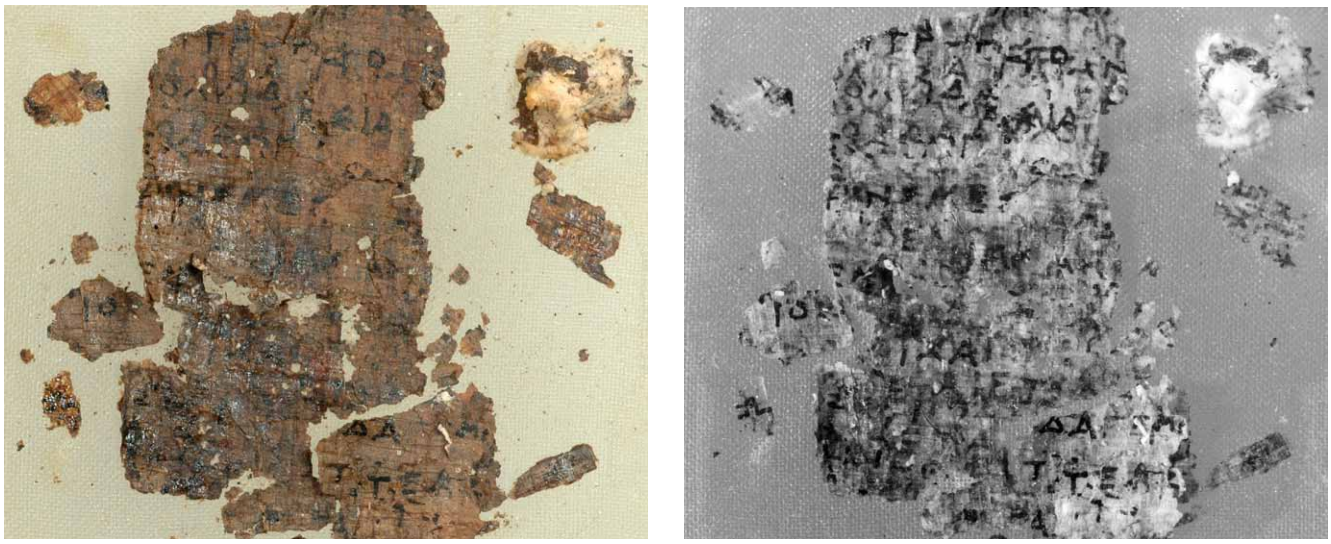


Fig. 4: Nikolaos Gyzis' signatures from 'Spirit' (above) and 'Study Π581_1' (below) at specific wavelengths.

[11] MuSIS HS by Forthphotonics (now DySIS) 400–1,000 nm with 20 nm interval and 34 spectral bands, Schneider-Kreuznach XENOPLAN 1.4/23 CCTV lens, Osram 250W tungsten light sources.

[12] Attas et al. 2003.



Visible Infrared at 1,000 nm

Fig. 5: The Musician's Tomb papyrus MP8523 fragment, approx. 2.3 cm x 3.5 cm.

absorption. This differentiation of materials can also be observed in the false colour infrared image (FCIR), in which the former material presents a greenish hue, but the latter a sepia one. Another interesting aspect that can be observed in the ultraviolet fluorescence colour image (UVFC) is that the sketch is free of brushstrokes. The brushstrokes are on the reverse side and were only visible because of the transparency of the paper. They disappeared because fluorescence is a phenomenon that only pertains to an object's surface.

Gyzis' signatures were examined to study and document the different media the artist used to sign his work. Hyperspectral imaging was used for this reason. Black-and-white images of the signatures in the visible and infrared range were acquired in narrow bandwidths to track the inks' reflective behaviours. As can be seen in fig. 4, the selected images from hyperspectral imaging at 600 nm, 700 nm and 800 nm clearly present the gradual decrease in absorption of the first ink and the unchanged absorption of the second ink. While infrared reflectography produces images pertaining to the entire near-infrared range (760–1,000 nm), in hyperspectral imaging, wavelengths are recorded at specific intervals. This enables changes to be observed and tracked that take place in the infrared range. Furthermore, hyperspectral imaging cameras provide important information on the materials through false colour infrared images, which were obtained and compared with each other for the signatures. Red false colour can be typical of iron-gall inks, while black can be for carbon-based inks or pencils.

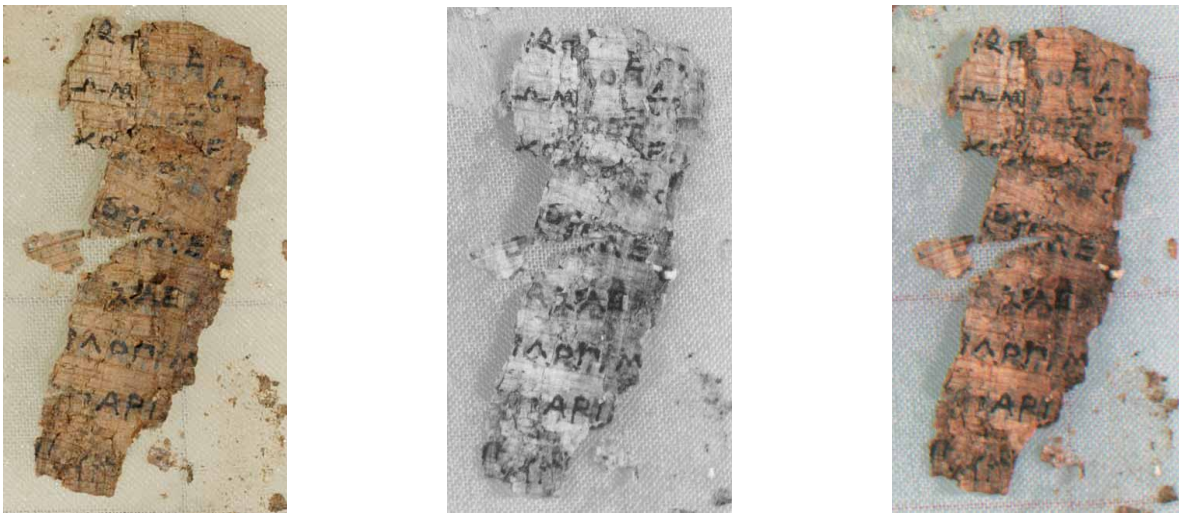
The same apparatus enabled the acquisition of a reflectance spectrum. This revealed the inks' visible and near-infrared characteristics, which proved to be very effective in tracking and distinguishing data produced from different depths with greater precision and resolution when compared with classic reflectography. Hyperspectral imaging combined with the raking light technique provided information-enriched images due to the surface texture, thus making discrimination between the signing media possible.

5. The papyrus found in the 'Musician's Tomb'

'Greek artist's grave yields rare papyrus' – this was the headline of an article by the Athenian correspondent of the British national daily *The Times* in May 1981 announcing that two tombs dating to 420/430 BC had been found during an emergency excavation in Daphne, Athens.[13] One of the tombs contained, among other artefacts, parts of musical instruments, writing tablets and a papyrus, thus leading the archaeologists to name it the 'Musician's Tomb'. Although the papyrus was at first considered completely destroyed and incapable of any reclamation, special significance had to be attributed to it and to the *polyptychon* tablets as they are the oldest Greek text testimonials found in Greek territory to date.[14]

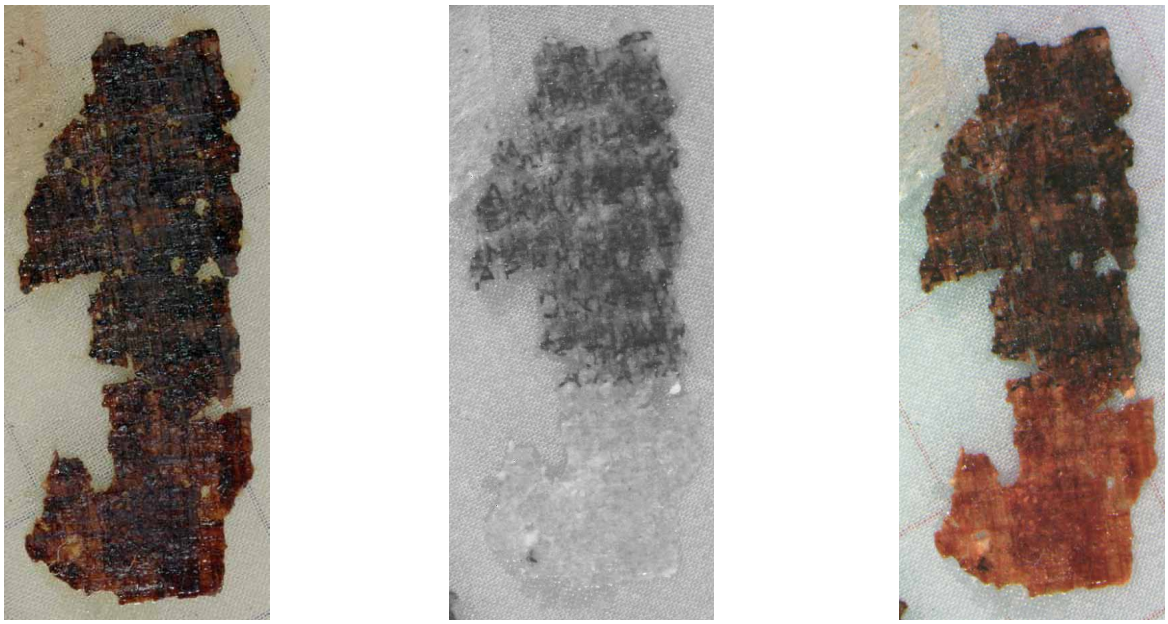
[13] Alexopoulou et al. 2013; Alexopoulou and Kaminari 2013.

[14] West 2013.



visible | infrared at 800 nm | FCIR

Fig. 6: the Musician's Tomb papyrus MP8520 fragment 1, approx. 3.1 cm x 1.4 cm.



visible | infrared at 1,000 nm | FCIR

Fig. 7: The Musician's Tomb papyrus MP8520 fragment 2, approx. 1.6 cm x 4.8 cm.

The papyrus, decayed into an amorphous mass of nearly rotten fibres,[15] comprised several leaves pasted together. A white material resembling plaster had destroyed most of the upper portion. The conservator in charge carefully separated as many fragments as he could from the formless mass, placed them onto an organic silk fabric in wooden frames and consolidated and restored them. The papyrus and all the other finds are now kept at the Archaeological Museum of Piraeus.

In some cases, one could observe more letters on the surface in various places with orientations or further letters appearing between the lines. In the inventory, these were assumed to be probable musical notes, and this was one of the critical questions to investigate since this element would be very important in the interpretation of the finds (fig. 5).

The main concern in this research, carried out in 2012, was to somehow revisualize and thus reclaim written data from

[15] Pöhlmann 2013.

underneath layers of papyrus, which were found mashed together in the tomb before being transferred together to their current holding place, as well as to facilitate making the dark fragments of papyrus legible.

Hyperspectral imaging combined with principal component analysis (PCA) and image segmentation techniques were employed for these reasons. Spectral image cubes were acquired at visible and infrared wavelengths to differentiate the script from elements pertaining to the condition of the object, e.g. scratches, deposits, etc. False colour mode also enhanced our ability to collect more information on the script reclaimed from the papyrus as well as on issues related to the investigation of the writing both in terms of overall legibility and discerning individual letters of the text. Fragments that had turned dark due to chemical consolidants applied during conservation were rendered clearer and the script again became more readable (figs. 6 and 7). As images were difficult to interpret by the unaided eye since they contained information from different layers, image segmentation, image subtraction and the well-known application of PCA were used to optimize the extraction and differentiation of information. The resulting images showed clearly distinguishable areas of text, the layers of the papyrus were differentiated from one another better and letters that belong to the same sheet were able to be grouped together with greater confidence.

To complete the investigation of the Musician's Tomb papyrus, an effort was made to capture digital photographs in macro mode in the visible spectrum using the newly developed digital Nikon D800, which employs a Nikon FX-format CMOS sensor with 36.3 effective megapixels' resolution. This photographic apparatus was employed in order to collect the maximum amount of readable information from the fragmentary remnants of words and letters on the papyrus, which initially seemed to resist successful imaging.

In conclusion, for objects whose significance and fragility demands special care and which do not allow for physical sampling, these non-destructive test methods assisted by image-processing techniques are the best approach as these reveal information from a certain depth near the surface of the object as well as from the surface itself. False colour imaging differentiates relevant features from scratches and yields better results for distinguishing them when compared with single wavelength recording. Currently, a photographic apparatus with extremely high resolution and optical quality seems to be a low-cost tool for macro-mode investigation.

6. Conclusions

The case studies presented here are just part of the research activities carried out using equipment available at various times. Each case had its own needs and presented a unique set of problems and limitations. This paper has shown that it is possible to study artefacts and obtain useful images capable of yielding significant scholarly results regardless of the restrictions encountered.

These case studies have also made it clear that even though technology has facilitated image acquisition from manuscripts and works of art by exploiting a wide range of wavelengths, the comparative interpretation of the results requires a very careful approach to be taken based on the structure and chemistry of the materials. Optical data must always be correlated with physical and chemical properties, morphology, aging and construction technology for individual objects.

ACKNOWLEDGEMENTS

Part of this research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Programme 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF) – Research Funding Programme: ARCHIMEDES III. Investing in knowledge society through the European Social Fund.

REFERENCES

Alexopoulou, A., Banou, P., Gerakari, K., Kaminari, A., Stassinou, A. (2012), 'Non-destructive documentation of the H. Schliemann copy letters archive using imaging techniques', *European Research Center for Book and Paper Conservation, Newsletter,* 2.

——, Banou P., Stassinou A. (2006), 'Βιβλία αντιγραφής επιστολών. Τεχνολογία κατασκευής – Η χρήση τους στην Ελλάδα', *Archival News. The Society of Greek Archivists' Newsletter,* 23: 55–59.

—— and Kaminari, A. (2013), 'Multispectral Imaging Documentation of the Findings of Tomb I and II at Daphne', *Greek and Roman Musical Studies,* 1: 25–60.

——, ——, Panagopoulos, A., and Pöhlmann, E. (2013), 'Multispectral Imaging Assisted by Image Processing: A Useful Tool for the Study of Ancient Writing and Sketches on Different Substrates', *Journal of Archaeological Science,* 40.2: 1242–1249.

ASCSA (2005), *The American School of Classical Studies at Athens* (www.ascsa.com [accessed 2005]).

Attas, M., Cloutis, E., Collins, C., Goltz, D., Majzels, C., Mansfield, J. R., and Mantsch, H. (2003), 'Near-infrared spectroscopic imaging in art conservation: investigation of drawing constituents', *Journal of Cultural Heritage*, 4.2: 127–136.

Banou, P., Kaminari, A., Moutsatsou, A., Alexopoulou, A., and Singer, B. (2010), 'Investigating the conservation problems of oil paintings on paper supports', in *International Symposium 'Works of Art and Conservation Science Today', 26–28 November, Thessaloniki, Greece*.

Chabries, D. M., Booras, S. W., and Bearman, G. H. (2003), 'Imaging the past: recent applications of multispectral imaging technology to deciphering manuscripts', *Antiquity,* 77 (296): 359-372.

Fischer, C., and Kakoulli, I. (2006), 'Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications', *Reviews in Conservation*, 7: 3–16.

Liang H. (2012), 'Advances in multispectral and hyperspectral imaging for archaeology and art conservation', *Applied Physics A: Materials Science and processing*, 106.2: 309–323.

Padoan, R., Steemers, T. A. G., Klein, M. E., Aalderink, B. J., and de Bruin, G. (2008), 'Quantitative Hyperspectral Imaging of Historical Documents: Technique and Applications', *9th International Conference on Nondestructive Testing (NDT) of Art* (Jerusalem, Israel; CD-ROM), 1–10.

Pöhlmann, E. (2013), 'Excavation, Dating and Content of Two Tombs in Daphne, Odos Olgas 53, Athens', *Greek and Roman Musical Studies,* 1: 7–24.

'Special HID lamps, blacklight Blue HPW and MLW, Optical lamps', *Philips Technical Data Sheet* (accessed 2005), 28–29.

West, M. (2013), 'The Writing Tablets and Papyrus from Tomb II in Daphne', *Greek and Roman Musical Studies,* 1: 73–92.

Article

DivaDesk: A Holistic Digital Workspace for Analyzing Historical Document Images

Nicole Eichenberger, Angelika Garz, Kai Chen, Hao Wei, Rolf Ingold, and Marcus Liwicki | Fribourg

Abstract

In this article we present the concept of DivaDesk – a Virtual Research Environment (VRE) for scholarly work on historical documents inspired by the shift toward working with digital facsimiles. The contribution of this article is three-fold. First, a review of existing tools and projects shows that a holistic workspace integrating the latest outcomes of computational Document Image Analysis (DIA) research is still a desideratum that can only be achieved by intensive interdisciplinary collaboration. Second, the underlying modular architecture of the digital workspace is presented. It consists of a set of services that can be combined according to individual scholars' requirements. Furthermore, interoperability with existing frameworks and services allows the research data to be shared with other VREs. The proposed DivaDesk addresses specific research with historical documents, as this is one of the hardest cases in computational DIA. The outcomes of this paradigmatic research can be transferred to other use cases in the humanities. The third contribution of this article is a description of already existing services and user interfaces to be integrated in DivaDesk. They are part of ongoing research at the DIVA[1] research group at the University of Fribourg, Switzerland. The labeling tool DivaDia, for example, provides methods for layout analysis, script analysis, and text recognition of historical documents. These methods build on the concept of incremental learning and provide users with semi-automatic labeling of document parts, such as text, images, and initials. The conception and realization of DivaDesk promises research outcomes both in computer science and in the humanities. Therefore, an interdisciplinary approach and intensive collaboration between scholars in the two research fields are of crucial importance.

1. Introduction

The technological developments seen in the last few decades have triggered a shift in how scholars in the humanities work when viewing historical documents in their repositories whenever research questions require an analysis of the original documents. Thanks to digitization, they are now able to work with digital facsimiles, so access to visual representations of historical documents has become much easier. The increasing amount of digital data available in virtual libraries, such as *e-codices*[2] and *manuscripta mediaevalia*,[3] provides new research possibilities, such as comparing digital facsimiles of different repositories and annotating digital images. In order to handle the data and to perform research tasks on digital facsimiles, scholars need usable tools. In addition to viewing facsimiles, direct searches for specific text passages in the digitized data are also needed. Furthermore, linking research data and annotations with corresponding text passages would be highly beneficial. Another desideratum is a tool for viewing all the samples of a specific text phrase, initial, or decoration contained in a document (or set of them).

While tools for specific tasks[4] do exist, none of them serve scholars in the humanities in all aspects of their work on digital facsimiles, i.e. the generation and presentation of item descriptions, content representation, and research data. Our vision is to realize DivaDesk, a VRE for scholars in the humanities that includes semi-automatic state-of-the-art methods from computer science. This undertaking requires an interdisciplinary approach and intensive collaboration between scholars in the humanities and computer science. On the one hand, DivaDesk should be appropriate for

[1] This stands for Document, Image, and Voice Analysis.

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

[3] See http://www.manuscripta-mediaevalia.de.

[4] Libraries: presentation of digitized manuscripts; research: edition tools, annotation tools; computer science: automatic DIA methods.

scholars' work, therefore the conception of its generic architecture and functions belongs to the domain of the humanities. On the other hand, it builds on current evolution in the field of computational DIA and is also meant to be the prototypical area of application for ongoing computer science research in that field. The realization and application of DIVADesk, therefore, promise fruitful research outcomes both in the humanities and computer science. The structure of the present paper reflects this interdisciplinary approach. It contains specific information and deals with specific problems in the humanities as well as in computer science in order to address readers in both research communities, while also striving to make the specifics comprehensible for readers in either community.

The rest of this paper is organized as follows. Section 3 summarizes existing tools and projects for historical document viewing and analysis as well as the state of the art of computational methods. Section 4 discusses the current situation in the humanities and conceptualizes the virtual workspace for scholars in the humanities. Existing tools and services from our group are presented in section 5, where we also evaluate several automated DIA tasks. Finally, section 6 concludes the paper and highlights the challenges we are currently facing in interdisciplinary research in the digital humanities.

2. State of the art

Historical document image analysis is considered one of the hardest tasks in the digital humanities for several reasons. For one thing, there are many research questions in the humanities related to individual documents as well as extensive corpora and historical collections. For another, the properties of the physical documents pose many research challenges to automatic DIA. Past research on historical DIA can be categorized in two main tiers: tools developed at institutes in the humanities, typically tailored to the needs of a specific use case, and automatic tools developed by computer scientists in the context of Pattern Recognition (PR). While the border between these categories is not clearly defined and several projects have been conducted jointly by scholars in the humanities and PR researchers (see section 3.1), there is a general tendency toward two divergent movements, as was observed at a recent Dagstuhl Seminar on Digital Humanities.[5]

In the following, we will first review select tools and projects that make use of computational methods in the context of humanity research and, second, summarize important outcomes of PR research in automatic DIA. Finally, toolkits for Ground Truth (GT) generation will be discussed, as GT is an essential requirement for the development of automated DIA methods.

2.1 Historical document viewing and analysis

Many virtual manuscript libraries with different mandates and different functionalities exist. For example, the virtual library called *e-codices* assembles medieval manuscripts in Swiss repositories. It provides a viewer, allows nuanced searches, and enables annotations to be made to specific digitized manuscripts. Close collaboration with libraries ensures sustainable availability of the manuscript images (under a Creative Commons license). Active development to enhance the user interface makes *e-codices* a highly renowned and extensively used virtual library. Other online libraries, such as the *Bibliothèque virtuelle des manuscrits médiévaux*[6] and *manuscripta mediaevalia*, offer similar functionality. Several libraries provide their manuscript collections in useful digital viewers of their own, e.g., the British Library (London),[7] the Bibliothèque nationale de France (Paris),[8] the Houghton Library at Harvard University,[9] the University Library of Heidelberg,[10] and the Herzog August Bibliothek in Wolfenbüttel.[11] One example of a thematically oriented library is the *New Testament Virtual Manuscript Room*,[12] which allows manuscripts to be viewed, indexed, and transcribed. These libraries do not include any automated methods for layout analysis or OCR (Optical character recognition), however.

Tools and projects targeting the transcription of texts appearing in document images have been implemented for specific text corpora. The *Transcribe Bentham*[13] initiative is

[5] Hassner et al. 2013.

[6] http://bvmm.irht.cnrs.fr/.

[7] http://www.bl.uk/manuscripts/.

[8] http://gallica.bnf.fr/.

[9] http://hcl.harvard.edu/libraries/houghton/collections/early_manuscripts/.

[10] http://www.ub.uni-heidelberg.de/helios/digi/handschriften.html.

[11] http://dbs.hab.de/mss/?list=browse&id=project.

[12] http://ntvmr.uni-muenster.de/home.

a collaborative transcription initiative for manuscripts by the philosopher Bentham that relies on a novel crowdsourcing strategy. *Die Rätoromanische Chrestomathie*,[14] another crowdsourcing approach for printed books from the 19th and 20th century, was recently finalized. While the former project does not take advantage of automated recognition by any means, the latter used OCR as initial seed. It has been shown that a recognition error rate lower than 15% is already enough to significantly speed up the transcription process[15] given an appropriate transcription tool when compared with manual transcription without any assistance.

While the tools and projects mentioned above only target the generation of transcriptions, the EU project known as *tranScriptorium*[16] goes one step further. In addition to utilizing DIA results for crowdsourcing applications (e.g., *Transcribe Bentham*), the results are intended to be included in digital archives and e-research portals.[17] The specific objective of *tranScriptorium* is the integration of automated methods into platforms for DIA, text recognition, and keyword spotting. Comparison and presentation aspects are beyond the scope of *tranScriptorium*, however. In the earlier project, IMPACT (*IMProving Access to Text*),[18] a set of automatic tools for specific tasks was developed, which will be reviewed in sections 3.2 and 3.3.

Another initiative targeting automated support is the *Genizah project*,[19] which deals with fragments of Jewish manuscripts collected in a digital library (the *Genizah platform*). An automated search for similar fragments can be performed using local features description to assemble fragments belonging to the same manuscript. As this project deals with the specific case of fragments, OCR and semantic information retrieval functionalities are not included. The ORIFLAMMS project[20] aims at a computer-based paleographic analysis of single characters or parts of characters. Its ultimate goal is to develop an ontology of characters and graphs with the help of automated clustering. The *Monk* system[21] is a platform for word searches in archive collections. It allows for storage and annotation of scanned page images and provides automated methods for text recognition, which deal with problems such as different writing styles for specific words in order to perform search queries on the archival material.

Many tools have been developed for digital editions in recent decades.[22] In the field of Middle High German literature, the *Parzival Project*[23] is one of the most ambitious digital edition projects, dealing with a voluminous text that is transmitted in different textual versions and in more than 80 manuscripts. The *Parzival Project* provides a synoptic edition of different text versions including the possibility of viewing the manuscript image next to the transcribed text as well as a critical edition and digital editions of individual manuscripts on CD or DVD. An interesting feature of this project is the use of the (originally biological) concept of phylogeny to determine and visualize interrelationships between manuscripts.[24] Another powerful tool dealing with textual criticism, phylogeny, and automated collation is CollateX.[25] This tool covers modern machine-readable texts, but does not deal with digitized images.

For digital representation of annotations and comments on document images, the *Shared Canvas platform*[26] has become very popular. It is employed for several use cases (e.g., the Archimedes Palimpsest[27]) and as the underlying architecture for virtual libraries, such as *e-codices*. With well-organized web interfaces, users can easily annotate images and select individual layers for the presentation of the original page and its annotations. The *System for Annotation and Linkage in Arts and Humanities* (SALSAH)[28] provides a VRE for

[13] Causer et al. 2012.

[14] Neuefeind et al. 2011.

[15] Vilar et al. 2010.

[16] Gatos et al. 2014.

[17] http://transcriptorium.eu.

[18] http://www.impact-project.eu.

[19] Wolf et al. 2011.

[20] http://www.irht.cnrs.fr/fr/recherche/les-programmes-de-recherche/oriflamms. See also Stutzmann 2013.

[21] http://www.ai.rug.nl/~lambert/Monk-collections-english.html.

[22] For a survey of the development of digital editions and their theorization, refer to Robinson 2013.

[23] http://www.parzival.unibe.ch/englishpresentation.html.

[24] Viehhauser and Chlench.

[25] http://collatex.net/.

[26] Sanderson et al. 2011.

[27] http://www.archimedespalimpsest.org/.

[28] Schweizer and Rosenthaler 2011.

working with digitized material of all kinds, e.g., manuscripts, books, videos, and even tape recordings. *Shared Canvas* and SALSAH are powerful tools for working with research data and meta-data, but computational DIA methods and automated recognition processes are not included.

Note that this review of tools and projects is far from exhaustive; for a more complete review of other projects and initiatives, such as ENRICH,[29] *manuscriptorium*,[30] and TEI, the reader should refer to Ciocoiu 2012. For a position paper raising general issues in the digital humanities, but focusing on computational methods for paleography, refer to the recent report by the Dagstuhl seminar.[31]

2.2 Computational document image analysis

The majority of DIA methods use binarization, i.e. they separate a given color or greyscale input image into its foreground and background (more than 90% of the recent DIA methods published at ICDAR and ICPR apply binarization at some point). It is apparent that this procedure leads to information loss, while errors made at this stage are inherited in subsequent automated processing steps. It is noteworthy that several recent approaches work without any binarization,[32] but while many text extraction or layout analysis methods require binarized input, the original input image (which should always be kept[33]) can be used again for further processing to ensure that no information is lost.

A set of heuristic methods have been proposed in the literature for the task of binarization, but typically focus on a certain class of documents. Global methods,[34] for example, are extremely fast. Local methods,[35] on the other hand, use different threshold values adaptively deduced for different image regions based on local information. They are strongly dependent on image resolution, the signal-to-background ratio, and local context, thereby making the window size a crucial parameter for a specific document set. Hybrid binarization methods are a straightforward extension of the above methods. They consider both local and global information in the binarization process. Apart from heuristic methods for binarization, machine-learning-based methods also exist. These are applied in two different ways: either (i) by automatically learning the parameters of a given binarization method[36] or (ii) by dividing the image into different regions and learning to select the appropriate method for each region.[37]

Surveys of document layout analysis are found in Mao et al. 2003 and Baird et al. 2011. Document layout analysis is performed in two stages, which are referred to as physical and logical layout analysis respectively. In physical layout analysis, the document image is divided into homogeneous regions depending on their content, e.g. text, graphics, and background. In the succeeding logical layout analysis, these regions are then assigned a specific label, e.g. 'title', 'heading', or 'main text'. A performance analysis of several page segmentation algorithms is presented by Shafait et al. 2008. Text-line detection methods are typically performed with an error rate of around 5%.

Research on text recognition (also often referred to as OCR for printed text, and Handwriting Recognition (HWR) for handwritten text) has been carried out for more than five decades.[38] The current state of the art is performing recognition using Recurrent Neural Networks (RNN) with error rates as low as <1% for printed historical documents and 6% for medieval manuscripts.[39] For more information on these methods, refer to section 5.3.

2.3 Ground Truth generation tools

An indispensable prerequisite for developing reliable semi-automated methods is to incorporate experts' knowledge in such systems ('learn from the expert'). This knowledge needs to be provided in the form of correct labels for the information extractable from digital facsimiles, called 'Ground Truth' or 'GT' in DIA research. GT facilitates the development of robust automated DIA approaches by enabling a machine to learn by example and, furthermore, allows assessing its performance by referring to the correct labels, i.e. how close

[29] http://enrich.manuscriptorium.com/.

[30] http://www.manuscriptorium.com/.

[31] Hassner et al. 2013.

[32] Chen et al. 2014; Garz et al. 2012.

[33] E.g., by building on the International Image Interoperability Framework (IIIF); see http://iiif.io.

[34] Otsu 1975; Yang et al. 2006.

[35] Niblack 1990; Trier, and Jain 1995; Sauvola and Pietikäinen 2000.

[36] Chou et al. 2010; Chamchong and Fung 2010.

[37] Sari et al. 2012.

[38] Plamondon et al. 2000; Bunke 2003.

[39] Fischer et al. 2012.

the prediction of the system is to the real data as labeled by a human expert. Hence, the primary aim of our project is to facilitate fast and uncomplicated generation of GT for large amounts of digitized documents.

Several labeling tools have been developed in recent years. AGORA[40] segments (historical) document images into two maps in order to divide them into the fore- and background. A user can then label, merge, and remove computed regions. *PixLabeler*[41] is a pixel-level labeling tool for binarized (bi-tonal) document images, where foreground pixels are assigned a color, with each color representing a label such as 'handwriting', 'machine print', 'graphics', etc. *Aletheia*[42] follows a top-down approach (iteratively splitting regions into smaller entities) for labeling binarized document images. Regions automatically detected by splitting and shrinking (fitting the boundary of a region to the entity) can be modified by the user, while low-level elements can be aggregated into more complex structures.

3. DivaDesk, a digital workspace for analyzing historical documents

3.1 Discussion of the current situation

As highlighted in the introduction, technical developments have had a considerable impact on the humanities. Working with historical documents has been simplified by the availability of digitized manuscripts and prints, while the way of working has also changed dramatically. Digitized manuscripts are more easily and readily accessible than microfilm or original documents, but this facilitated access also has its drawbacks. Obviously, the impression of the whole physical object is not available in a digital viewer and specific properties, e.g., binding and watermarks, can only be investigated on the original document. More crucially, the presentation of digitized material leads to a change in the behavior of researchers when investigating the material. Instead of considering the manuscript as a whole, in most cases they selectively access only a few pages, parts of texts, and other points of interest. As a result, relations the viewed page has to other pages or parts of the manuscript or general properties of the whole manuscript remain unnoticed. During an examination of the physical object, such discoveries are often made unintentionally or by chance and lead to novel insights. A digital tool for scholars in the humanities should, therefore, take the aforementioned workflow into account and facilitate fast investigation of whole manuscripts by including functionalities to automatically identify potential regions of interest and notify the scholar, while equally allowing for new relationships the system did not come up with.

Another obstacle is the cumbersome way of accessing documents through diverging interfaces and function modes of different viewing tools employed by libraries or manuscript databases. If a scholar works on manuscripts held in different repositories, he has to adapt his working procedure to the specifics of different viewing tools; this reinforces the tendency to mainly perform specific, short-sighted searches in order not to lose too much time in adapting to unknown interfaces.

It is therefore crucial to build a digital workspace for scholars in the humanities that meets professional standards and allows intuitive handling of digitized data at the same time. The conceptualization of such a workspace is the main purpose of this article. The workspace should support the scholar in his daily work and in the digital presentation of research outcomes by making use of computational methods without forcing him to adapt his working procedures to implementation-specific requirements that different tools have for specific tasks.

3.2 Interdisciplinary approach

The aim of our work is to create a workspace that allows scholars to perform research in their accustomed way while being supported by state-of-the-art computational DIA methods. Therefore, an interdisciplinary approach is of crucial importance. Computer scientists and scholars in the humanities will collaborate intensively during the entire development and evaluation process of the tool. This will enable relevant functionality and good usability to be provided for the tool and thus ensure its acceptance by the research community.

One example of the interdisciplinary approach we are taking is the projected workflow for (semi-)automatic script analysis (for further information on the technical issues of this task, see Garz et al. 2014). After the development of a prototype framework by computer scientists, scholars in the humanities will assist them in finding appropriate test data (manuscript images) along with corresponding GT. Initial experiments will then be performed on this data, and the results of these will then be reviewed and evaluated by scholars in the humanities by means of GT and feedback

[40] Ramel et al. 2006.

[41] Saund et al. 2009.

[42] Clausner et al. 2011.

on the result representation. This iteratively performed dual evaluation is a very important step in the improvement of the tool.

3.3 Conception

In this section, we present our conception of DivaDesk, which addresses specific research with historical documents. It is noteworthy that these concepts can be transferred to other areas in the humanities as well. In the VRE, the scholar is able to gather digital facsimiles of manuscripts he is working on and arrange them in virtual libraries in order to reconstruct historical collections (i.e., manuscripts having once belonged to the same library, but now dispersed in different repositories) or to assign a manuscript to one or more of his research topics, such as different testimonials of the same text for an edition project. The scholar can describe the manuscripts or import existing catalogs, transcribe texts or parts of the text from the primary source, and make local annotations for specific spots inside a manuscript or extensive annotations on a more general level. All these tasks are not fundamentally different from the usual procedure of scholars working with historical documents, as they generate and store their research outcomes as text files on a computer or as paper files in a physical folder. The scholar is therefore not forced to adapt his procedure to a standard interface developed according to common practice in computer science, but the virtual workspace offers essential advantages with respect to usability, connection, and presentation of the research outcomes when compared with a real writing desk.

The fact that the whole working process – from the description of the manuscript to the finished article on the topic – can take place in the same VRE and that outcomes of previous working processes are also accessible in this environment allows for synergistic effects. By means of a (semi-)automated or manual search, similarities or relations between different documents or research outcomes can easily be discovered or retrieved. Let us illustrate this with the example of automated watermark retrieval, which we plan to integrate into DivaDesk. If scholars work on an inventory of several dozen manuscripts, they are very unlikely to remember the watermarks of the first manuscript when describing the 30th manuscript. Yet a relation between two manuscripts could be discovered through an identical watermark. An automatic image-retrieval process can point out similarities between watermarks in different manuscripts and hence help scholars to cope with large collections of research material.

Furthermore, DivaDesk will support scholars in visualizing and presenting research outcomes. The joint presentation of research outcomes in textual and visual form, e.g., transcribed text together with an image of the original document and visualizations of relations between manuscripts or contents, is particularly useful for research inspired by the theory of New Philology[43] or guided by the interest in the materiality of the documents.[44]

3.4. Architecture

The DivaDesk workspace currently focuses on the analysis of medieval manuscripts, but its modular architecture allows for the integration of services for other kinds of documents, e.g., prints, (early) modern handwritten documents, pictures, and objects. Its specification allows for the integration of particular services or sets of services into other frameworks as well as the integration of external services into DivaDesk using specified protocols. The interoperability of DivaDesk with other frameworks and platforms is thus guaranteed (see section 5 for technical details).

The architecture of the workspace for medieval manuscripts builds on three main modules, as depicted in fig. 1, and contains two different interaction modes. The input mode supports the working tasks of the scholar, e.g., cataloging, transcribing, and arranging digitized data. The presentation mode allows data to be displayed, either as an intermediate stage of the working process or as a research outcome.

The first module addresses the description of items — medieval manuscripts in our case. It contains information about each item in the form of a catalog. In accordance with the *TEI-P5 format for manuscript description*,[45] it is composed of three submodules:

1. the physical description of the manuscript;
2. the history of the manuscript containing its origin and provenance;
3. an overview of the content.

This information is generated by the scholar while working on the manuscript or is imported from existing manuscript catalogs. In the submodule *Physical Description*, several (semi-)automated processes can be used: layout analysis for

[43] Gleßgen and Lebsanft 1997.

[44] Nichols et al. 1996; Nichols 1997; Ortlieb 2013.

[45] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html.

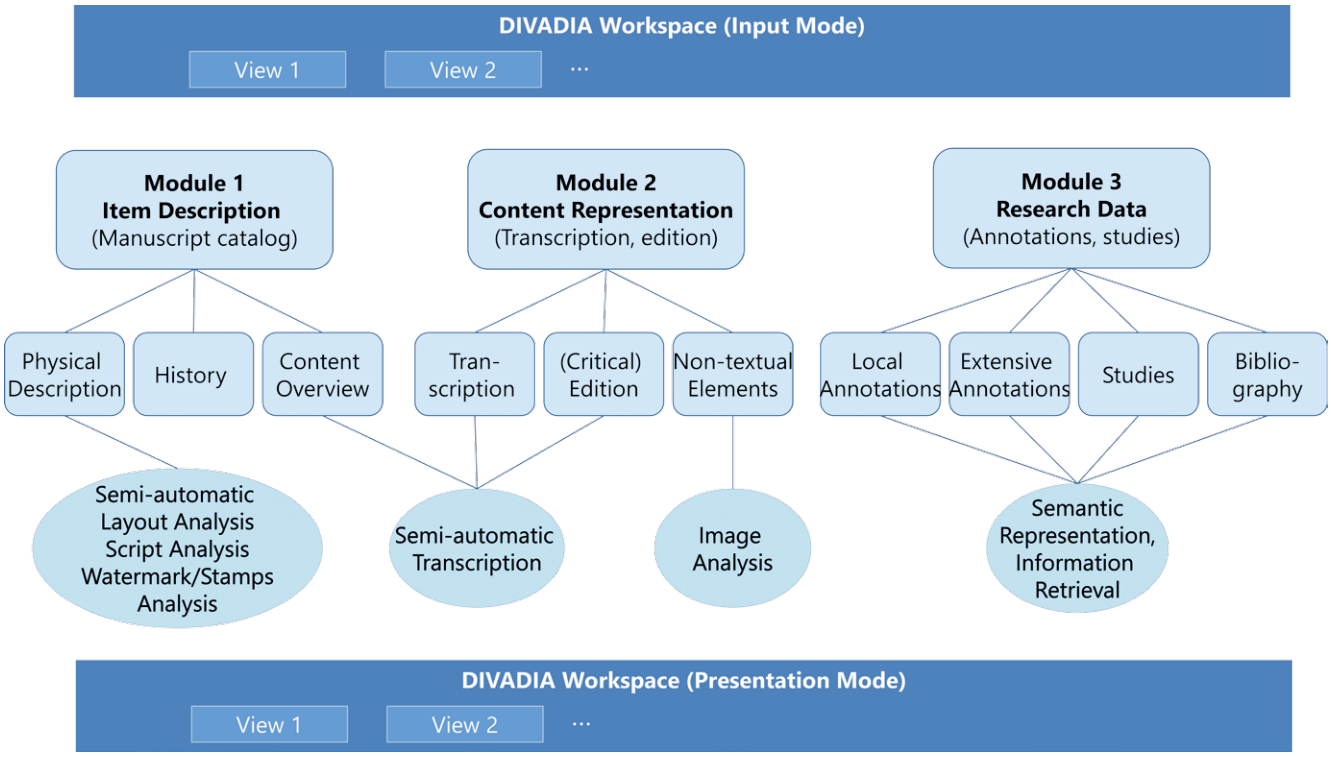


Fig. 1: Overview of DivaDesk's architecture.

structuring the pages and finding points of interest; script analysis for identifying scribe changes; and watermark and binding stamp analysis for identifying and comparing similar watermarks/stamps.

The second module is concerned with the representation of the content. For medieval manuscripts, it contains three submodules:

1. a submodule for transcription built according to the *TEI-P5 Format Representation of Primary Sources*[46] and linked to the digital facsimile;
2. a submodule that allows for (critical) editions of several textual witnesses/versions and is built on the *TEI-P5 format Critical Apparatus*;[47]
3. a submodule for the representation of non-textual elements, i.e., decorations, miniatures, or diagrams. (Semi-)automatic text-recognition processes will be used for the transcription.

The third module contains research data created by the scholar that is neither a description/catalog of the manuscript nor a representation (transcription/edition) of the primary source. In this way, we are building on the foundations laid out in DARIAH.[48] The four submodules of this module allow for different ways of linking the research data to the digitized data:

1. local annotations are linked to a specific spot in the digital facsimile, e.g., a word, text block or image;
2. extensive annotations are not bound to a specific local area within the facsimile having a larger/more abstract scope, e.g., summaries, plot patterns, or motifs;
3. in contrast to the annotation submodules, the third submodule contains finished (and/or published) studies discussing one or several manuscripts that are part of the virtual workspace. For the semantic annotation, XML-based methods and tools are already available, e.g., the project *Sharing Ancient Wisdoms*,[49] which can be integrated and adapted in the DivaDesk workspace;
4. a bibliography of secondary literature mentioning manuscripts that are part of the virtual workspace. In this module, information retrieval methods will be used, e.g., for finding shelfmark mentions in secondary

[46] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html.

[47] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html.

[48] http://dariah.eu/.

[49] Jordanous et al. 2012.

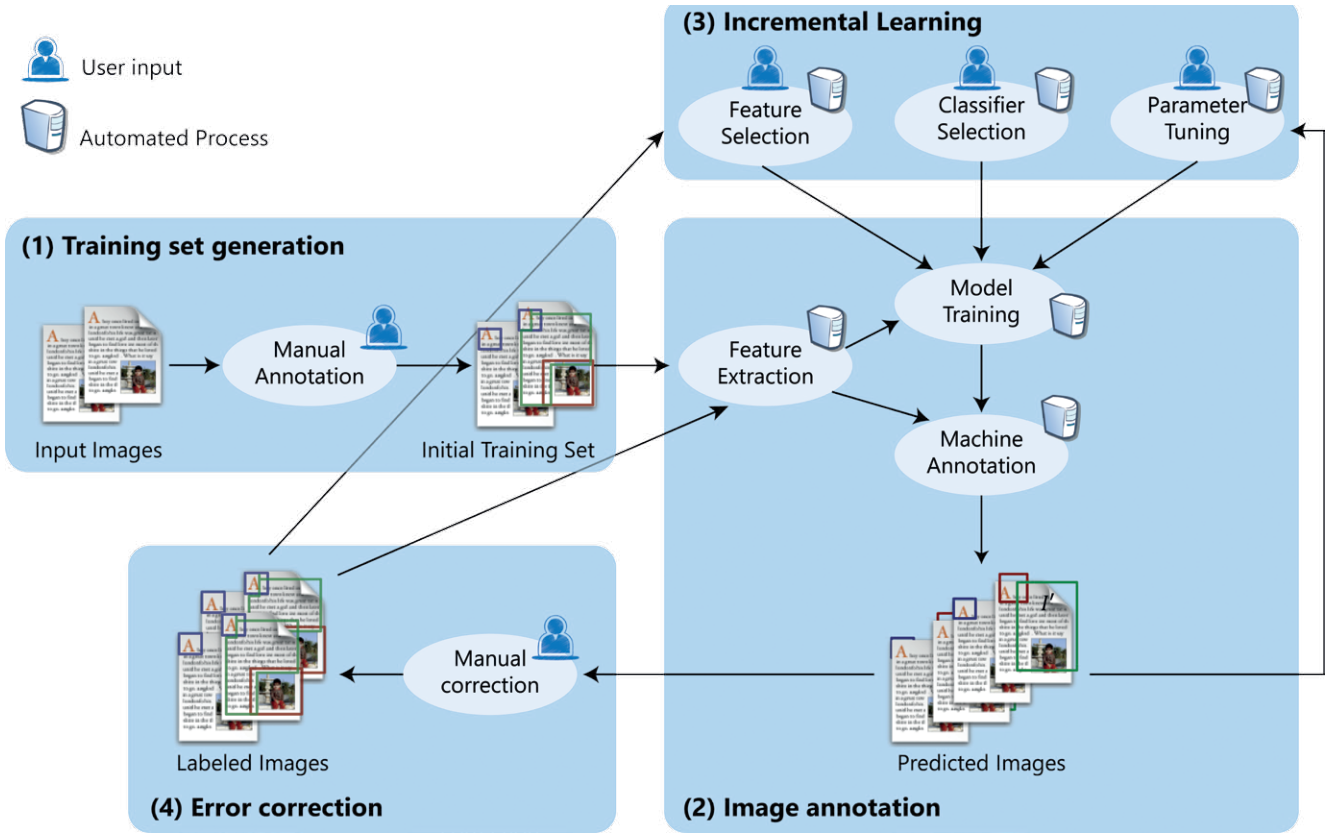


Fig. 2: Workflow of the document-labeling process.

literature and linking this information to the respective facsimiles in the workspace.

4. Technical realization and implemented tools & services

DIVADESK is a modular software system implementing several state-of-the-art methods for the modules illustrated in fig. 1 as well as views for input and presentation. The system can easily be expanded by plugging in new modules offering additional functionality. Currently, a set of services called DIVADIA Services has already been realized. These services are used by the DIVADIA labeling tool.[50] While DIVADIA is based on Java, any other programming language can be used; for *Item Description and Transcription*, for example, DIVADIAWT, an HTML5 view (based on the AngularJS framework), has been implemented (see section 5.4).

The modular design of our system allows for seamless integration of existing tools for specific tasks into the overall framework as well as the integration of components into other platforms. The computational methods generated during the IMPACT and tranScriptorium project, for example, enhance the image-processing functionality of Module 1. Furthermore, the open-source OCR engine OCROPUS[51] provides state-of-the-art layout analysis and text recognition. Since we are building on the Shared Canvas framework and use TEI for the representation of document meta-data, data created in our tools can be shared with other platforms, such as *e-codices*. Our target-oriented web service (e.g., text-line segmentation) can be used by other transcription tools, and even the web-based user interface for the transcription of documents can be integrated into other online platforms, such as the new version of *e-codices* and SALSAH. Such interaction with other projects allows for a holistic framework that serves all aspects of scholarly work with historical manuscripts and research data, incorporating the recent outcomes of research in automatic DIA.

While at present the holistic workspace is still in the conceptualization phase, several web services have already been implemented as DIVADIA Services, which will be integrated

[50] A preliminary Java version is available at http://diuf.unifr.ch/hisdoc/divadia. The Web-Interface and RESTful Webservices are only accessible inside the local network of the University of Fribourg, but will be opened as soon as the administrative process at our University has been completed. The web services have preconfigured settings for common workflows, such as image and text retrieval in the facsimiles.

[51] http://code.google.com/p/ocropus/.

into the three modules of the DIVADesk architecture. In this section, we describe the functionality of these services. Subsequently, we provide an overview of existing views/interfaces for specific workflows. Finally, we conclude with evaluation results on publicly available data sets and an outlook on future work.

4.1 Module 1: Item Description

The main concept used in this module is incremental learning, i.e. the system adjusts what it has learned previously according to new examples. We illustrate this with an example: Given that an automated DIA method works well on one set of historical documents, the recognition performance might still not be optimal for a specific unseen manuscript. In such a case, the user labels a few samples (text lines or pages). These labels are then used as GT for adapting and improving the automated DIA methods. The detailed process of incremental learning is described in the following and illustrated in fig. 2.

In order to start a new labeling process, a user manually annotates several[52] representative lines or page images of a document (1). These images in conjunction with their annotations compile the GT for the generation of an adapted prediction model, which is computed on the basis of a feature set and a machine-learning (ML) algorithm (2). Having computed the adapted model, the user tests it by selecting another set of images (3), which is automatically annotated based on the model. The predicted result is presented to the user, who can manually correct and/or accept it (4a) or try to improve the model (4b) by changing the ML algorithm, feature set, or parameters, for instance. If the user accepts a result, the GT is extended (5) by those newly labeled images, and a refined prediction model is computed. Starting from the third step, the process is pursued until the entire document has been annotated.

In DIVADIA Services, the following DIA methods are currently available:

- Image processing: standard binarization methods, local filters such as edge detectors, smoothing, Laplacian of Gaussian (LoG), Difference of Gaussians (DoG), and other non-linear filters that help enhance the image and remove noise.
- Feature extraction:[53] color and coordinates, incorporating information on the neighborhood of the considered pixel, Local Binary Patterns (LBP) that focus more on the textual structure of the image, Scale-Invariant Feature Transform (SIFT), which considers 'interesting' regions in the image, and Gabor features, which describe the dominant orientation of a pixel. Currently, unsupervised feature-learning methods are implemented as well.
- Feature selection: greedy forward/backward selection, sequentially floating forward selection, linear forward selection, genetic selection, and hybrid feature selection.[54] Feature selection methods help to retrieve the best set of features by automatically testing several combinations and systematically searching for the best combination.
- Machine-learning algorithms that automatically learn to classify given patterns after receiving sample patterns with GT: Support Vector Machines (SVMs), Modified Quadratic Discriminant Function (MQDF), k-nearest neighbor algorithm (k-NN), Naïve Bayes classifier (NB), Gaussian Mixture Models (GMMs), Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM),[55] and Markov Random Fields (MRF).
- Evaluation: Presently, our evaluation metrics are based on precision and recall. Additional metrics will be incorporated to enable a user to assess the quality of the prediction results. The annotation information is saved in XML format.[56]

These methods and all future versions of DIVADIA Services are available as open-source projects. Note that the methods are generally able to deal with texts written in any scripts and orienting any direction, but the trained methods only work on Latin script. For a description of technical details, refer to Wei et al. 2013.

4.2 Module 2: Content Representation

The concept of the automated DIA methods for content representation is similar to the one described for Module 1. Furthermore, most of the approaches described above can be used for recognition and analysis of non-textual

[52] Note that the exact number of text lines or page images to be manually annotated depends on the specific difficulties of a given manuscript.

[53] Wei et al. 2013.

[54] Wei et al. 2014.

[55] Graves et al. 2009.

[56] Pletschacher and Antonacopoulos 2010.

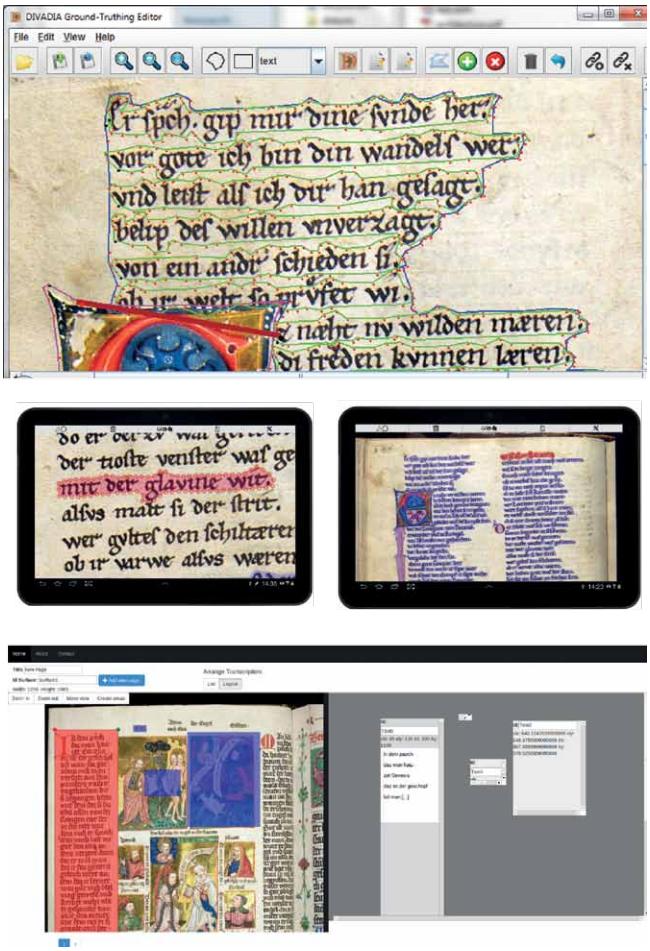


Fig. 3: DIVADIA labeling tool (top row), Android application (second row) and DIVADIAWT (bottom row).

images as well, especially SIFT and the ML methods. We have developed a set of methods in the HisDoc project for recognizing texts (OCR/HWR) and for editing.[57] These methods rely on handwritten text lines as input. After normalization of skew (inclination of the text line), slant (inclination of the characters), and width and height, one out of two state-of-the-art text-line recognition modules can be applied: Hidden Markov Models (HMM)[58] and LSTM[59] neural networks. The results of these methods as well as the manually entered information are stored in TEI format in order to ensure compatibility with other existing and future toolkits and frameworks.

4.3 Module 3: Research Data

Currently, our services allow researchers to find specific text passages in manuscripts even if the automatic recognition did not perform perfectly or the query is ambiguous, e.g., due to orthographic or dialectal varieties, similar to Google search. In order to support generation and access research data, we will integrate state-of-the-art Natural Language Processing (NLP) and Information Retrieval (IR) methods in the future.[60] Furthermore, we plan to integrate the identification of manuscript shelfmarks in order to automatically link existing annotations or published studies concerning the same manuscripts.

4.4 Views

The concept of DIVADESK incorporates several views (or interfaces) for the input and (re-)presentation of item descriptions, content, and research data. Note that most of the views can be used for input as well as for presentation of the data, depending on current needs, thereby reducing the learning curve for a new user. For example, a view for semi-automatic transcription of manuscript pages can be used for viewing the transcriptions as well. In the following, three existing prototypical views are presented.

The DIVADIA labeling interface allows users to manually or (semi-)automatically label a document image. They can display and enhance an image using several image-manipulation methods in order to make details visible, for example. Drawing tools similar to those known from image-editing software (e.g., Adobe Photoshop, Gimp) allow manual annotation of regions. In order to modify the automatic prediction of the system, a user can change system parameters after visually inspecting the previewed results or directly modify the size, boundaries, position, or category of predicted regions. A screen shot of this interface in its current form is presented above in fig. 3. In addition, we have implemented an interface for mobile Android devices supporting *touch & write* input[61] (see the second row in fig. 3), providing a natural user interface for scholars in the humanities, as annotations can be directly drawn and written with a pen. Finally, DIVADIAWT is a web-based interface that presents the transcriptions in the layout of the original manuscript image (see bottom row in fig. 3). It includes

[57] Fischer et al. 2012.

[58] Ploetz and Fink 2009.

[59] Graves et al. 2009.

[60] Naji and Savoy 2011.

[61] Liwicki et al. 2010; Dengel et al. 2012.

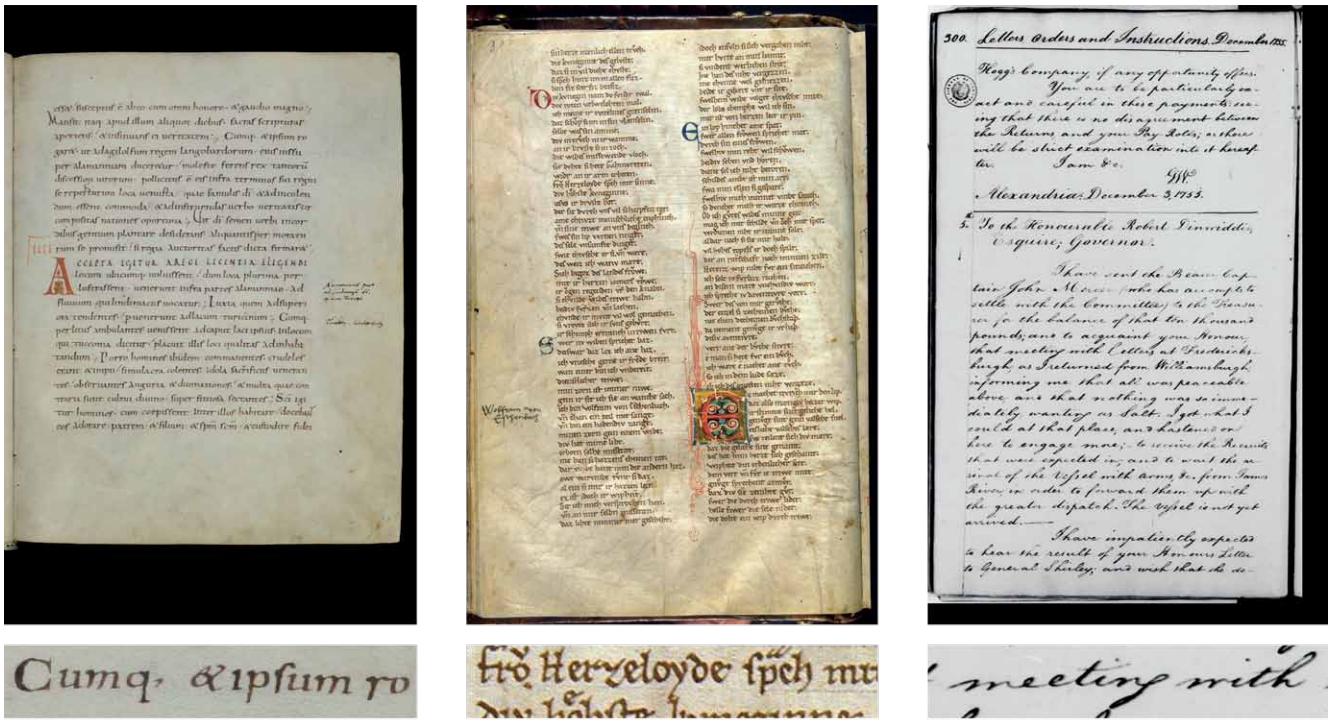


Fig. 4: From left to right: exemplary images of the Saint Gall data set (St. Gallen, Stiftsbibliothek, Cod. Sang. 562, p. 7), the Parzival data set (St. Gallen, Stiftsbibliothek, Cod. 857, p. 36), and the George Washington data set (Washington, Library of Congress, Letterbook 1, p. 300).

automatic support for line segmentation and the generation of XML files in TEI without having to install any software.

4.5 Evaluation of automated methods

The methods developed throughout the completed HisDoc project,[62] the ongoing project HisDoc 2.0,[63] and methods implemented in the DivaDia Services (see sections 5.1 and 5.2) have been constantly evaluated on data sets of a diverse nature. More specifically, the three data sets of the IAM-HistDB[64] have been used for evaluation: the Saint Gall data set,[65] the Parzival data set,[66] and the George Washington data set.[67] For sample images, refer to fig. 4. Note that for purposes of brevity, most technical details (exact sizes of training and test set, algorithm- and data-specific parameters, and detailed analyses) do not appear in this article, but are available in the cited publications. In this article, we focus on the most significant outcomes of these experiments.

The performance of our layout analysis methods is measured at pixel level. When considering a four-class classification problem, i.e. categorizing the pixels into periphery, background, text block, and decoration, our best method achieves an error rate of less than 9% on the Parzival data, around 4% on the Saint Gall data, and 11% on the George Washington data[68] with the Naïve Bayes ML algorithm, which runs very fast. These results are already useful for practical applications, since errors only appear on the border of the lines and, typically, a perfect border is not required for recognition.

For text recognition, i.e. OCR on complete text lines, we achieved word-level error rates of around 3.5% on the two medieval data sets when text lines were segmented manually. When applying our fully automated system, which first detects text lines and then recognizes the text, the error rate increases to 7%,[69] but even this result is useful in practice. As mentioned in section 3.1, integrating an automatic system with an error rate of less than 15% into the transcription

[62] Fischer et al. 2012.

[63] Garz et al. 2014.

[64] http://www.iam.unibe.ch/fki/databases/iam-historical-document-database.

[65] St. Gallen, Stiftsbibliothek, Cod. Sang. 562.

[66] St. Gallen, Stiftsbibliothek, Cod. Sang. 857.

[67] Washington, Library of Congress, Letterbook 1.

[68] Wei et al. 2014.

[69] Fischer et al. 2014.

process would speed up the annotation process significantly. To assess the performance of Information Retrieval (IR), we simulated 60 possible user queries searching for text appearances in manuscripts of the Parzival data set, e.g., 'dem man dirre aventivre giht', 'iwer oder decheines man', and 'als man von siner helfe saget'. System performance is measured by the mean reciprocal rank, a measure that is high if the text lines of interest appear in the top ranks. We have observed that the performance loss of IR on automatically recognized text is less than 1% compared with the performance on perfect transcription.[70] This result confirms that the OCR is already useful in practice.

4.7 Future work

Apart from the technical realization of DivaDesk, the quality of source material and legal issues are going to be major challenges in future. In order to achieve high-recognition performance, the quality of the digitized images needs to be sufficiently high (300 dpi is considered to be the minimum resolution for our current methods, and lossless image formats are indispensable). We are currently studying methods that work on low-resolution images in order to integrate images from other sources than recent high-quality digitization, e.g., digitized microfilm or historical photographs of lost or destroyed manuscripts.

Legal issues include copyright on image data, which typically remains with the repositories of the original documents. In the present development phase, we have not started focusing on those issues yet, given that our primary aim is to build an individual workspace for personal research data. As long as the data is used in the individual workspace only, legal issues are no major problem. We are aware of the importance of these issues for later stages of the realization of our workspace when functionalities for publishing and sharing research data will be added. Those issues can only be solved by all the stakeholders concerned cooperating with each other.

The finished HisDoc project and the ongoing project, HisDoc2.0, focus on the development of computational methods for DIA and OCR mainly from the computer scientist's point of view. The development of usable tools and a workspace for scholars in the humanities is a long-term goal to which these projects will lead; DivaDesk's conception and design are fine for the current project phase, but its realization is not included in the planned outcomes of the HisDoc 2.0 project. In order to realize a functional open-access version of the DivaDesk VRE, we envisage follow-up projects with an interdisciplinary focus.

5. Conclusion and outlook

In this paper, we proposed a novel conception of a holistic digital workspace called DivaDesk. It allows scholars working with historical documents to continuously make full use of new possibilities arising from technological development and digitization. The workspace provides a set of services that are interoperable with other frameworks and platforms. Therefore, individual combinations of different VREs are possible. The architecture of DivaDesk consists of three main modules: Item Description, Content Representation, and Research Data. It provides several state-of-the-art computational methods for supporting scholars in the humanities in their daily work. Current implementations provide automated DIA methods that achieve cutting-edge performance for layout analysis and text recognition. The envisioned workspace will provide further functionalities, supporting high-performance searches, comparison of editions and texts, and seamless connections to diverse research data. DivaDesk will be useful for research in the humanities and will push the current limits of the art in DIA.

[70] Fischer et al. 2012.

REFERENCES

Baird, H. S., Bunke, H., and Kazuhiko, Y. (1992/2011), *Structured Document Image Analysis* (1st ed.), (Berlin – Heidelberg: Springer), (DOI: 10.1007/978-3-642-77281-8).

Bunke, H. (2003), 'Recognition of cursive Roman handwriting: past, present and future', *Seventh International Conference on Document Analysis and Recognition. Proceedings*, 448-459.

Causer, T., Tonra, J., and Wallace, V. (2012), 'Transcription maximized; expense minimized? crowdsourcing and editing *The Collected Works of Jeremy Bentham'*, *Literary and Linguistic Computing,* 27.2: 119–137.

Chamchong, R., and Fung, C. (2010), 'Optimal selection of binarization techniques for the processing of ancient palm leaf manuscripts', in: *Systems Man and Cybernetics (SMC)*: 3796–3800.

Chen, K., Wei, H., Liwicki, M., Hennebert, J., and Ingold, R. (2014), 'Robust Text Line Segmentation for Historical Manuscript Images using Color and Texture', in *22nd International Conference on Pattern Recognition*, 2978-2983.

Chou, C.-H., Lin, W.-H., and Chang, F. (2010), 'A binarization method with learning-built rules for document images produced by cameras', *Pattern Recognition*, 43.4: 1518–1530.

Ciocoiu, A. M. (2012), *International collaboration in digital libraries: an analysis of the Manuscriptorium digital library case study*, Master thesis: International Master in Digital Library Learning, University of Tallinn/University of Parma (http://hdl.handle.net/10642/1266 [last accessed 13 Jan. 2015]).

Clausner, C., Pletschacher, S., Antonacopoulos, A. (2011), 'Aletheia — An Advanced Document Layout and Text Ground-Truthing System for Production Environments', in *International Conference on Document Analysis and Recognition*, 48–52.

Dengel, A., Liwicki, M., and Weber, M. (2012), 'Touch & Write: Penabled Collaborative Intelligence', in *Knowledge Technology,* 1–10 (Berlin – Heidelberg: Springer).

Fischer, A., Bunke, H., Naji, N., Savoy, J., Baechler, M., and Ingold, R. (2012), 'The HisDoc Project: Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries', in *The Proceedings of InterNational and InterDisciplinary Aspects of Scholarly Editing*.

——, Baechler, M., Garz, A., Liwicki, M., and Ingold, R. (2014), 'A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents', in *International Workshop on Document Analysis Systems*, 71–75.

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

——, Eichenberger, N., Liwicki, M., and Ingold, R. (2014), 'HisDoc 2.0 – Towards Computer-Assisted Paleography', *manuscript studies,* 7.

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J. A., Toselli, A. H., and Vidal, E. (2014), 'Ground-truth production in the transcriptorium project', in *International Workshop on Document Analysis Systems*, 237-241.

Gleßgen, M.-D., and Lebsanft, F. (eds.) (1997), *Alte und neue Philologie* (Tübingen: Niemeyer).

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J. (2009), 'A novel connectionist system for improved unconstrained handwriting recognition', in *IEEE Transactions on Pattern Analysis and Maschine Intelligence,* 31: 855–868.

Hassner, T., Rehbein, M., Stokes, P. A., and Wolf, L. (2013), 'Computation and Palaeography: Potentials and Limits (Dagstuhl Perspectives Workshop 12382)', *Dagstuhl Reports*, 2.9: 14-35.

Jordanous, A., Lawrence, K. F., Hedges, M., and Tupman, C. (2012), 'Exploring manuscripts: sharing ancient wisdoms across the semantic web', in *International Conference on Web Intelligence, Mining and Semantics*, 44.

Liwicki, M., Rostanin, O., El-Neklawy, S. M., and Dengel, A. (2010), 'Touch & write: a multi-touch table with pen-input', in *International Workshop on Document Analysis Systems*, 479–484.

Mao, S., Rosenfeld, A., and Kanungo, T. (2003), 'Document structure analysis algorithms: a literature survey', in *Electronic Imaging* (International Society for Optics and Photonics), 197–207.

Naji, N., Savoy, J. (2011), 'Information retrieval strategies for digitized handwritten medieval documents', in *Asia Conference on Information Retrieval Technology*, 103–114.

Neuefeind, C., Rolshoven, J., and Steeg, F. (2011), 'Die Digitale Rätoromanische Chrestomathie – Werkzeuge und Verfahren für die Korpuserstellung durch kollaborative Volltexterschließung', in *Multilingual Resources and Multilingual Applications: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology*.

Niblack, W. (1990), *An Introduction to Digital Image Processing* (Upton Saddle River, NJ: Prentice Hall , Inc.).

Nichols, S. G. (1997), 'Why material philology?', in Tervooren, H., und Wenzel, H. (eds.), *Philologie als Textwissenschaft. Alte und neue Horizonte* (Berlin: Schmidt), 10–30.

——, et al. (eds.) (1996), *The Whole Book. Cultural Perspectives on the Medieval Miscellany* (Ann Arbor: Univ. of Michigan Press).

Ortlieb, Cornelia (2013), 'Materialität', in R. Borgards et al. (eds.), *Literatur und Wissen. Ein interdisziplinäres Handbuch* (Stuttgart: Metzler), 41–45.

Otsu, N. (1975), 'A threshold selection method from gray-level histograms', in *Automatica*, vol. C, no. 1: 62–66.

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.

Pletschacher, S. and Antonacopoulos, A. (2010), 'The PAGE (Page Analysis and Ground-Truth Elements) Format Framework', in *International Conference on Pattern Recognition*, 257–260.

Ploetz, T., Fink, G.A. (2009), 'Markov models for offline handwriting recognition: a survey', *International Journal on Document Analysis and Recognition*, 12: 269–298.

Ramel, J. Y., Busson, S., Demonet, M. L. (2006), 'AGORA: the Interactive Document Image Analysis Tool of the BVH Project', in *International Conference on Document Image Analysis for Libraries*, 145–155.

Robinson, P. (2013), 'Towards a Theory of Digital Editions', *Variants,* 10: 105–131.

Sanderson, R. Albritton, B. Schwemmer, R. Van de Sompel, H. (2011), 'SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination', *ACM/IEEE Joint Conference on Digital Libraries*, *Ottawa, Canada, June 2011,* 175-184.

Sari, T., Kefali, A., and Bahi, H. (2012), 'An MLP for binarizing images of old manuscripts', in *International Conference in Frontiers in Handwriting Recognition*, 247–251.

Saund, E., Lin, J., Sarkar, P. (2009), 'PixLabeler, User Interface for Pixel-Level Labeling of Elements in Document Images', *International Conferenceon Document Analysis and Recognition*, 646–650.

Sauvola, J., and Pietikäinen, M. (2000), 'Adaptive document image binarization', *Pattern Recognition,* 33.2: 225–236.

Schweizer, T., and Rosenthaler, L. (2011), 'SALSAH – eine virtuelle Forschungsumgebung für die Geisteswissenschaften', *Electronic Visualisation and the Arts Berlin* (Elektronische Medien & Kunst, Kultur, Historie. 9.-11. November 2011), 147–153.

Shafait, F., Keysers, D., and Breuel, T. (2008), 'Performance evaluation and benchmarking of six-page segmentation algorithms', *Pattern Analysis and Machine Intelligence*, 30.6: 941–954.

Stutzmann, D. (2013), 'Système graphique et normes sociales: pour une analyse électronique des écritures médiévales', in N. Golob (ed.), *Medieval Autograph Manuscripts* (Bibliologia, 36), 429–434.

Trier, O., and Jain, A. (1995), 'Goal-directed evaluation of binarization methods', in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17.12: 1191–1201.

Viehhauser, G., and Chlench, K. (2014), 'Phylogenese und Textkritik. Bioinformatische Anregungen zur Lösung genealogischer Klassifizierungsprobleme in der Editionsphilologie', in M. Stolz (ed.), *Internationalität und Interdisziplinarität der Editionswissenschaft,* (Berlin: de Gruyter; Beihefte zu Editio, 34), 57-82.

Vilar, J. M., Castro-Bleda, M. J., Zamora-Martínez, F., España-Boquera, S., Gordo, A., Llorens, D., Marzal, A., Prat, F., Gorbe, J. (2010), 'A flexible system for document processing and text transcription', in *Current Topics in Artificial Intelligence* (Berlin – Heidelberg: Springer), 291–300.

Wei, H., Baechler, M., Slimane, F., Ingold, R. (2013), 'Evaluation of SVM, MLP and GMM Classifiers for Layout Analysis of Historical Documents', in *International Conference on Document Analysis and Recognition*, 1252–1256.

——, Chen, K., Ingold, R., and Liwicki, M. (2014), 'A Hybrid Feature Selection Method for Historical Document Image Analysis', in *International Conference on Frontiers in Handwriting Recognition*.

Wolf, L., Dershowitz, N., Potikha, L., German, T., Shweka, R., Choueka, Y. (2011), 'Automatic Palaeographic Exploration of Genizah Manuscripts', in *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*, ed. by F. Fischer, C. Fritze, and G. Vogeler (Norderstedt Schriften des Instituts für Dokumentologie und Editorik, 3), 157-179.

Yang, Z. Ma, and M. Xie (2006), 'A novel binarization approach for license plate', in *IEEE Conference on Industrial Electronics and Applications*, 1–4.

Article

Multispectral Imaging, Image Enhancement, and Automated Writer Identification in Historical Manuscripts

Ana Čamba, Melanie Gau, Fabian Hollaus, Stefan Fiel, and Robert Sablatnig | Vienna

1. Introduction

This article deals with multispectral imaging (MSI) and image-processing techniques for the analysis of damaged historical manuscripts. Ancient manuscripts are often in a poor condition: the ink can be faded and there is often degradation caused by mold, water stains, and humidity as a result of bad preservation conditions. In order to decipher the latent texts, philologists are dependent on the support of experts from other disciplines. As part of the Sinaitic Glagolitic manuscripts project (http://www.caa.tuwien.ac.at/cvl/research/sinai/), our interdisciplinary team consisting of philologists and image-processing specialists is examining damaged parchment manuscripts and developing computational means for the digitization, image enhancement, and automated document analysis of historical manuscripts in order to facilitate philological research.

The manuscripts studied originate from Mt. Sinai, Egypt, and were written between the 10th and 12th centuries in Glagolitic, the oldest Slavonic script. They show the typical characteristics of ancient manuscripts as listed above. Several of the manuscripts also contain palimpsests, i.e., underwritten text that has been washed or scraped off so the page can serve as new writing material — and then be overwritten with new text. The readability of these palimpsests is of extraordinary value for philologists, since they are older than the visible text on the top and may contain historical information that would be lost otherwise.

MSI is used in order to enhance the legibility of such manuscripts. By applying post-processing techniques on multispectral images, the contrast and legibility of degraded manuscripts can be increased. One potential post-processing technique is Fisher's linear discriminant analysis (LDA), which belongs to the group of supervised dimension reduction methods. The results of this approach are described in the following.

Our project also explores image-processing methods employed in the field of document analysis. One of the methods developed comprises automated writer identification. We propose an approach that uses SIFT features (scale-invariant feature transform) and Gaussian mixture models (GMM). This was the first time that automated writer identification was applied on Glagolitic manuscripts. In experiments, the method achieved a writer identification rate of 98.9% on a dataset containing 363 Glagolitic folios.

This article is structured as follows: In section II, the general concept of MSI is presented as well as a detailed description of the MSI acquisition system used. Section III contains an introduction to post-processing techniques with a focus on LDA. Section IV outlines the proposed method for writer identification. The final section provides a conclusion and an outlook.

2. Multispectral imaging of ancient manuscripts

A. Concept of multispectral imaging

MSI is an imaging technique that produces images in selected narrow spectral ranges. A multispectral image can be described as the same image of one scene in different spectral ranges, i.e., at different wavelengths of the electromagnetic spectrum. The wavelengths can either be isolated by filters or using instruments that are sensitive to specific wavelengths, including light from frequencies above and beyond the visible light range such as infrared (IR) and ultraviolet (UV). MSI can reveal information in a document that cannot be seen by the human eye, since the latter is sensitive only to radiation from about 380 nm to 700 nm. Another asset of MSI is that it is a non-invasive technique and can therefore be used on any manuscript, even if it is in a fragile condition.

MSI was originally developed for remote sensing applications. In the early 1990s, the technique started to be applied in the fields of art conservation and art history.

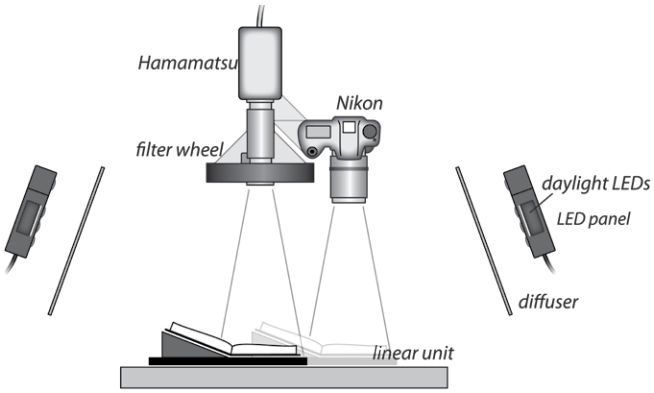


Fig. 1: Illustration of the MSI acquisition setup.

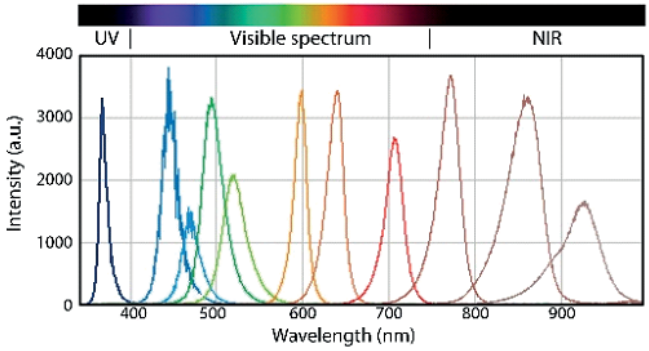


Fig. 2: Wavelengths of the LEDs used in the MSI system.

The imaging technique has also proven to be an effective method for the analysis and preservation of ancient manuscripts, as shown in the famous Archimedes palimpsest project.

The results of that project demonstrate that imaging in the UV spectral range in particular succeeds in providing additional data.

Several research studies have been carried out in the field of MSI for the enhancement of readability in ancient documents, of which just a few are mentioned in the following: Easton, Knox, and Christens-Barry introduce an MSI acquisition system that makes use of narrow-band LEDs. Bianco et al. describe an MSI apparatus that uses a filter wheel consisting of eight different optical filters and a monochromatic camera. Lettner et al. introduce a similar MSI system with an extra single-lens reflex camera. Finally, Rapantzikos and Balas make use of a system with optical filters for imaging in 34 narrow spectral bands.

B. MSI acquisition system

Since the manuscripts investigated are stored in different places and often even in different countries, our aim was to develop an apparatus that was robust, easily portable, and allowed fast setup and imaging. We constructed a portable MSI system consisting of two multispectral LED panels and two different cameras (illustrated by fig. 1). The setup takes approximately one hour. The two cameras are a Hamamatsu C9300-124 and a Nikon D4. The Hamamatsu camera is a near-infrared (NIR) grayscale camera with a cooled CCD chip and provides a resolution of 4000 x 2672 pixels. It possesses a spectral sensitivity ranging from UV to NIR (330 nm – 1000 nm). The Nikon camera is a traditional RGB camera with a resolution of 4928 x 3280 pixels and is utilized for UV fluorescence and visible light images.

The objects being investigated are placed on a board which is attached to a linear unit. The linear unit permits automatic shifting between the two camera positions. Thus, once positioned, there is no additional interaction with the manuscript apart from turning the folios.

Concerning the lighting system, we have made certain improvements in recent years. In the beginning we used optical filters built in a filter wheel which was placed in front of the Hamamatsu camera. Since the different filters within a filter wheel are not aligned exactly parallel, a shift occurs between single images. The filter influence is described by Brauers, Schulte, and Aach.[1] The shift has to be corrected for the statistical combination of different spectral bands. An additional drawback was the illumination system, where the illumination had to be switched manually between UV and white light illumination. In order to avoid these drawbacks, we constructed an image acquisition system where the multispectral images are gained by the lighting. For this purpose, two Eureka!Light™ (Archimedes project) LED panels were acquired, providing 11 different wavelengths (see fig. 2). Four white LED panels are attached to the left and right of the Eureka!Light™ panels. Additionally, two diffusers are situated between the lighting and the object in order to distribute the illumination uniformly.

Using LED panels with different wavelengths means that a filter wheel is redundant. This has the welcome side effect that optical distortions caused by filters can be avoided. Two filters are still required, however: the SP400 (400 nm short-pass filter) for UV fluorescence imaging and the LP400 (400 nm long-pass filter) for UV reflectography. An example of the results of multispectral imaging is shown in fig. 3.

[1] Brauers, Schulte, and Aach 2008.

Fig. 3: Vienna Folia. White light image (left) and UV fluorescence image (right).

3. Post-processing methods: dimension reduction techniques focusing on LDA

As described in the preceding chapter, the contrast and readability of historical texts can be enhanced with the aid of MSI. However, some sections of the manuscripts still remain illegible. At this point, post-processing methods such as dimension reduction techniques may assist in reconstructing these sections. Dimension reduction techniques reduce the dimensionality of the multispectral scan with the objective of extracting the relevant information, i.e. the text. Regarding manuscripts with one text, this means that the output from the MSI scans is at best a single image showing the writing. For palimpsests containing two layers of text, the reduction of dimensionality results in two images, one showing the overwriting and the other containing the underwriting.

Basically, dimension reduction techniques can be grouped into supervised methods (such as LDA) and unsupervised methods (such as PCA and ICA). For the former, class information is necessary to select different features, whereas this is irrelevant for the latter, since the relevant information in the MS scans is extracted by the approaches.

Previous studies[2] have proven that the application of unsupervised dimension reduction techniques such as principal component analysis (PCA) and independent component analysis (ICA) can enhance the contrast of degraded writings. Recently, we proposed a new enhancement technique based on a supervised dimension reduction approach, namely Fisher's linear discriminant analysis (LDA).[3] To the best of our knowledge, this was the first time that the LDA approach had been applied to historical documents. In a qualitative analysis (where the images were assessed by philologists familiar with the script of the relevant manuscripts), it was shown that the LDA-based approach produces partially better results than the unsupervised methods.

As a supervised dimension reduction tool – in other words, a method demanding additional information apart from the MSI data – the LDA-based method requires prior labeling of a subset of the multispectral data. The labeling is performed in a semi-automated label-generation step in which a subset of the multispectral data is labeled as belonging to the text or the background. The method was tested on two Glagolitic manuscripts, namely 'Missale Sinaiticum' and 'Glagolitic Fragments'. Since the text in the first manuscript is only partially visible in the multispectral images, the labeling is applied on PCA images where an enhancement in readability is achieved. Due to the bad condition of the writings, however, labeling by applying a binarization algorithm would still not be effective. Additionally, in the case of 'Missale Sinaiticum', it is not predefined in which wavelength the text is best visible, and hence it is not known a priori to which image of the multispectral scan the binarization algorithm should be applied.

Instead we noticed that the text line scheme is more recognizable in the PCA images than the single characters. Therefore, the samples are labeled as belonging either to text line or intermediate regions. In order to identify these regions, a text line detection algorithm similar to the one proposed in

[2] For example Easton, Christens-Barry, and Knox 2011; Salerno, Toazzini, and Bedini, 2007; Hyvärinen and Oja 2000.

[3] Hollaus, Gau, and Sablatnig 2013.

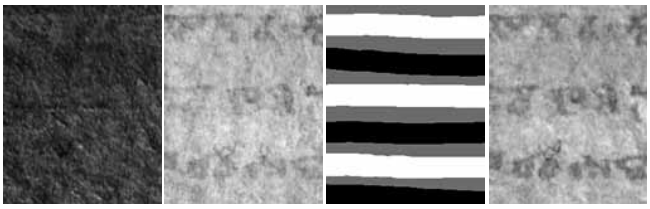


Fig. 4 (from left to right): UV reflectography image, PCA input image, text line detection result obtained on the PCA image, LDA output image.

Bar-Yosef et al.[4] is applied on the PCA image. The output of the method comprises lines passing through text lines and lines passing through intermediate regions between text lines. These lines are then dilated with a disk-structuring element in order to increase the number of samples. The labeled pixels are then utilized for training an LDA classifier. The hyperplane found by the classifier is subsequently used for the dimension reduction of the entire multispectral scan. The resulting image is in turn used in a further labeling step. A sample of the output from this stage is shown in fig. 4. The writing is most visible in the UV reflectography image in this case, but it should be noted that the writing is still barely visible. This can be attributed to a chemical reaction causing the ink to discolor from black to white.

The example shows that the text is usually more visible at this stage in comparison to the unprocessed multispectral images, and hence a binarization algorithm can be successfully applied. A binarization approach[5] is therefore used for a further labeling step. The binarization technique is specially designed for historical manuscripts and is mainly dependent on a single parameter that describes the average stroke width. The stroke width is identified manually and is similar within each folio of the manuscripts investigated, except for strokes which belong to initials. Unlike labeling based on text lines, this labeling step is used to label characters instead of text line regions. The output of the binarization approach is then used again in order to label a subset of the multispectral data and perform LDA-based dimension reduction.

Figs. 5 and 6 show images from a multispectral scan and the corresponding results of enhancement. The writing is most visible in the UV fluorescence image compared to the other multispectral images. Nevertheless, the legibility of the writing is limited, since the text is affected by bleeding degradations and corrupted by background clutter. These degradations are best restored using the LDA-based approach.

Furthermore, the examination showed that the LDA-based technique yields better results if it is applied to a region with a similar contrast. If there are both very dark and very bright areas in the background, the text line detection algorithm will fail.[6]

4. Automated writer identification

Automated writer identification for ancient, degraded manuscripts is a desirable technique for paleographers to use. The aim of a paleographer's work is to date, localize, and verify historical documents and – in the best-case scenario – to identify the author of the work. A great number of manuscripts have been digitized in recent years in order to protect them from abrasion and make the documents available to researchers quickly and easily (and to anyone else with an interest in them, for that matter).

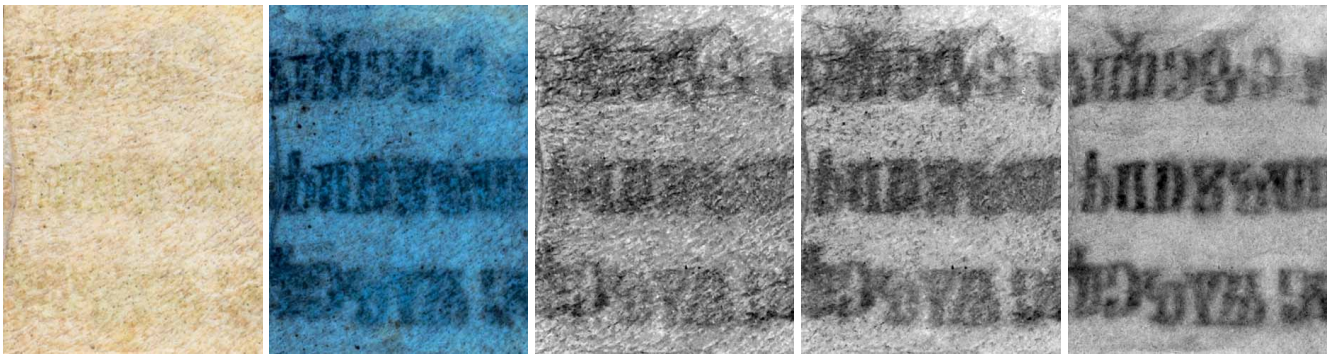


Fig. 5 (from left to right): white light image, UV fluorescence image, PCA result, ICA result, LDA result.

[4] Bar-Yosef et al. 2009. For a detailed description of the procedure applied, see Hollaus, Gau, and Sablatnig 2013.

[5] Su, Lu, and Tan 2010.

[6] This issue is discussed in more detail in Hollaus, Gau, and Sablatnig 2013.

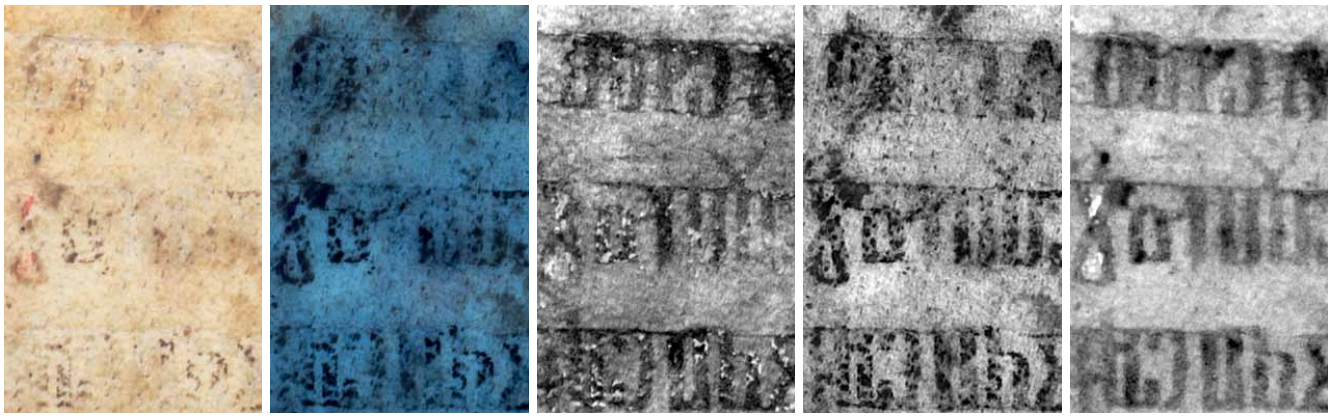


Fig. 6 (from left to right): white light image, UV fluorescence image, PCA result, ICA result, LDA result.

This trend is set to increase in the future. An automated method that can 'learn' to recognize and discriminate the typical writing styles of different authors within the vast jungle of data would therefore greatly facilitate the work of the paleographer. At present, paleographers almost exclusively use manual methods to solve this task. Thus, it was an essential objective of our project to create an effective model for writer identification.

Most of the writer identification models that can be found in literature are concerned with modern handwritings. Particularly over the past decade, however, several attempts have been made to design methods of automated writer identification for medieval manuscripts. Among the researchers, the following apply their methods on binarized images: Bensefia, Paquet, and Heutte[7] work with grapheme features and apply their method both on modern handwritings and on handwritings from the 19th century. The same approach is used for modern handwritings, whereby the results gained for modern writings are significantly superior to the results obtained for ancient writings.

Bulacu and Schomaker's[8] writer identification method is a combination of textual and allographic features. The authors show that combining the features obtains better results than with single features. The manuscripts investigated were 70 medieval English documents. Brink et al.[9] introduced a method where the width and the directionality of the ink trace serve as features. This system was tested on historical English and Dutch documents.

Bar-Yosef et al.[10] have criticized the fact that binarization does not show proper results when it is applied to damaged historical manuscripts. They utilized a multi-step binarization method themselves. As a result, particular letters are recognized automatically and applied for writer identification. In order to manage the classification and attribution of the different texts to correct scribes, k-nearest neighbors and Bayes linear classifiers are evaluated. The authors note that the latter classifier achieves better performance than the k-nearest neighbors classifier.

Contrary to the above approaches, which are all based on binarized images, Bres, Eglin, and Volpilhac-Auger[11] used a model that works with grayscale images. They applied the Hermite transformation for denoising and identification of scribes and investigated 1438 historical manuscripts written by 189 different scribes.

Wolf et al.[12] approached the topic of writer identification in a broader sense. They designed a semi-automatic clustering method based on graphical models for the classification and identification of document fragments that once belonged to the same codex. They used the bag-of-words model to find image similarities. Using their technique, approximately 1000 previously unidentified relations could be detected.

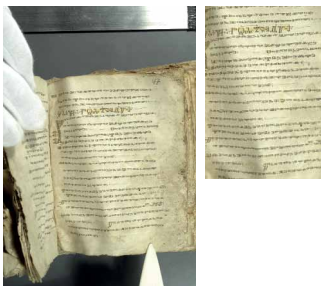


Fig. 7: Input and output images of the cropping procedure.

[7] Bensefia, Paquet, and Heutte 2003.

[8] Bulacu and Schomaker 2007.

[9] Brink et al. 2012.

[10] Bar-Yosef et al. 2007.

[11] Bres, Eglin, and Volpilhac-Auger 2006.

[12] Wolf et al. 2011.

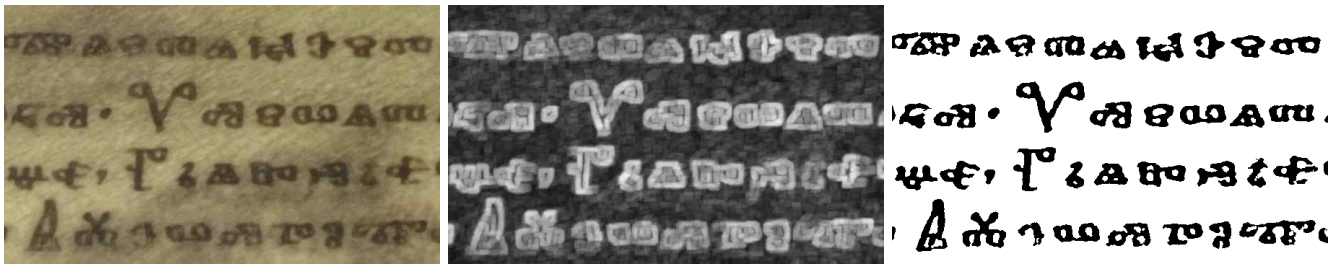


Fig. 8: binarization steps. Input image (left), contrast image (center), binarization result (right).

We propose a writer identification approach which is based on Fisher kernels and the utilization of scale-invariant feature transform (SIFT) features and Gaussian mixture models (GMM). This approach was originally developed to work on grayscale images of modern handwritings, i.e., with a uniform background. What happens when this method is applied on damaged historical documents is that SIFT features are also recognized on background clutter. In order to avoid this negative side effect, a binarization step is required as a form of pre-processing – a method that has already provided good results on binarized images of modern handwritings.

We applied our writer identification model on five Glagolitic manuscripts originating between the 10th and 11th centuries. These manuscripts had already been well-examined by scholars, so there were already paleographic results regarding the scribes.

Before the writer identification method is applied, the documents are cropped to images containing only the text of the manuscript. This is done by applying the text line detection method, which was mentioned in the preceding section.[13] Since this method only allows the extraction of rectangular regions, decorative elements such as initials can also be found in the cropped image (see fig. 7).

In the next step, the images are binarized using a binarization method designed specifically for historical documents by Su, Lu, and Tan.[14] First, a contrast image is calculated which encodes local gray-value differences. These differences are especially high at stroke boundaries, and the algorithm is based on the detection of stroke boundaries. An example of this type of contrast image is given in fig. 8 (center). After computation of the contrast image, high-contrast pixels are identified by applying Otsu[15] thresholding on the contrast image. These high-contrast pixels are usually located at stroke boundaries and are used in the final step of the binarization process. At this stage, each pixel of the input image is considered and is classified as a foreground pixel if the following conditions are met: firstly, the pixel must be near a sufficient number of high-contrast pixels, whereby the required number of high-contrast pixels is a user-defined parameter; secondly, the gray value of the pixel considered must be smaller than or equal to the average gray value of pixels in its local neighborhood, which are marked as being high-contrast pixels. A sample output of the algorithm is shown in fig. 8.

The writer identification model proposed is based on the Fisher kernels.[16] The starting point is to calculate the SIFT features using the scale-invariant feature transform (SIFT)[17] on a training dataset. In order to recognize Glagolitic characters correctly, the features were adopted in our examination, as suggested by Diem and Sablatnig:[18] if an angle occurs which is larger than 180°, the orientation of the key-point is mirrored. This is because the upper and lower level of features within the writing is a discriminative feature for Glagolitic character recognition. Other modifications we performed were heightening the contrast threshold (due to a lot of noise) and reducing the edge threshold (due to a high number of edge-similar features).

The calculation results of two Glagolitic letters and the corresponding histograms are depicted in fig. 9.
Where the two Glagolitic characters are concerned, the relevant features are found in the circles and at the corners. The histograms clearly show that if the SIFT features are calculated based on rotational invariance, both letters are impossible to distinguish.

[13] For a detailed description, see Fiel et al. 2014 (forthcoming).

[14] Su, Lu, and Tan 2010.

[15] Otsu 1979.

[16] Perronnin and Dance 2007; improved by Perronnin, Sanchez, and Mensink 2010.

[17] Lowe 2004.

[18] Diem and Sablatnig 2010.

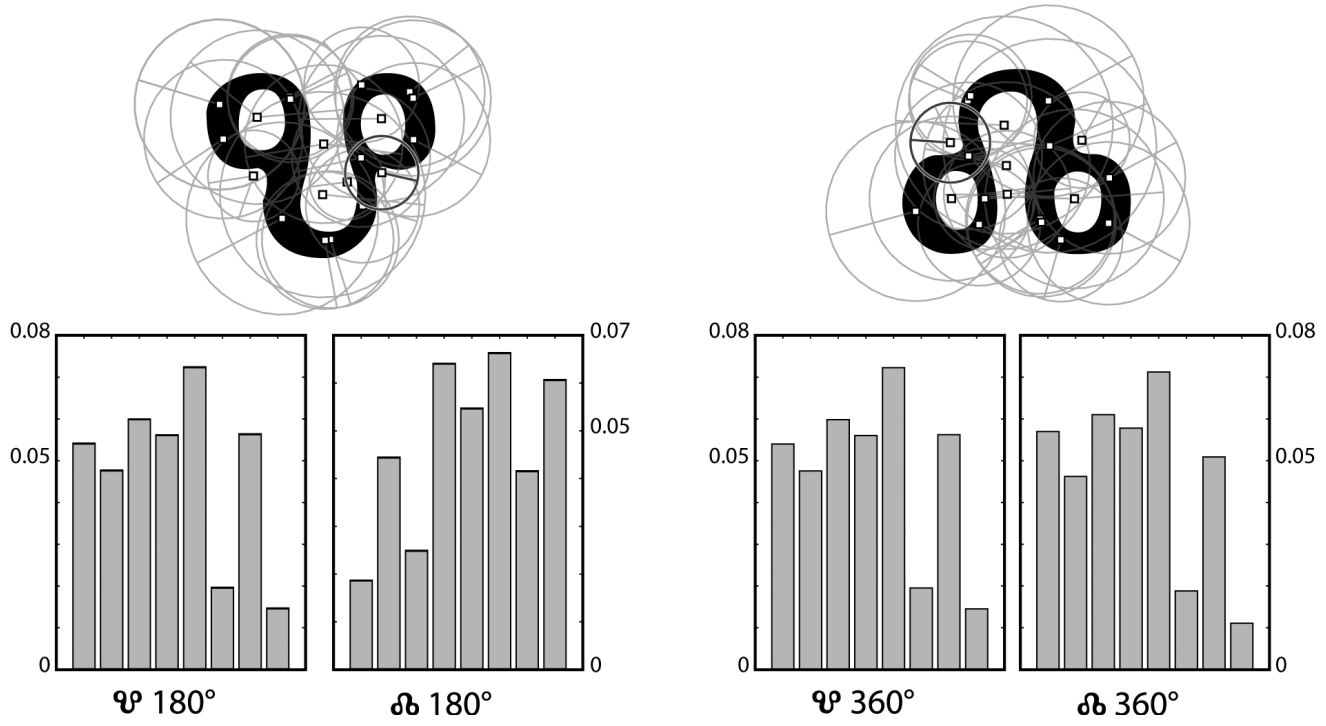


Fig. 9: the top row shows two Glagolitic characters with their SIFT features, while the bottom row shows generated histograms of the two marked features (blue). In the bottom row on the right, the calculation of the SIFT features is rotation-invariant, thus they both generate a similar histogram. On the left, the calculation is rotation-dependent, which makes the two characters distinguishable.

The objective of the next step is to create a visual vocabulary. For evaluation reasons, a different training dataset is used than in the SIFT procedure. First, a PCA is employed on the features to reduce the dimensionality from 128 to 64. Each SIFT feature in the training set can be seen as an observation of a Gaussian mixture, and the parameters of this density function can be estimated using an expectation maximization algorithm. The number of Gaussian has to be set in advance, and experiments have shown that 100 distributions have the highest performance. When identifying the writer of a new page, the SIFT features of the relevant page are calculated and the Fisher kernel for each GMM is applied. The resulting vectors for each distribution are then concatenated. The cosine distance of the resulting vector to the features of known writers is calculated, and the writer with the smallest distances is assigned as the writer of the particular document.

The documents are then ordered by their similarity to a reference document, and the evaluations are carried out using the k-nearest neighbors algorithm. The evaluation method is based on the model of the ICDAR 2011 Writer Identification Contest.[19] Each document in the database is used as a reference document for the evaluation, and the distances to the remaining documents are calculated. These distances can then be used to compute several criteria, particularly the Top 1 criterion, meaning that the document with the smallest distance to the reference document has to be by the same writer as the reference document. For this criterion, the method gained a rate of 98.9% on binarized images. The identification rate on grayscale images was 87.6% for the Top 1.[20]

5. Conclusion and outlook

Interest in the digital imaging of ancient documents has significantly increased in the past decade. There can be no doubt that study and processing of this cultural heritage in the future will be done predominantly by means of those image products. MSI is a useful technique for the non-invasive investigation and preservation of ancient documents. It has been proven that MSI is a successful method for enhancing the legibility of damaged historical documents, and applying UV light is especially effective for deciphering faded text or palimpsests. Since the approach is still relatively young, we also plan in future to focus on MSI research within our

[19] Louloudis, Stamatopoulos, and Gatos 2011.

[20] For the interested reader, a detailed evaluation can be found in Fiel et al. 2014.

interdisciplinary project. This will enable us to collect further qualitative data about the imaging of degraded historical manuscripts and palimpsests in order to make new scientific findings.

This article has also dealt with the enhancement of degraded manuscripts that have been captured by MSI. The supervised dimension reduction technique of LDA was shown to be able to further heighten text legibility. The method is organized as follows: first, the data is labeled in text line regions and intermediate regions using a text line detection technique, since the condition of the manuscript investigated is extremely bad. We then apply a procedure to label a subset of the multispectral data as foreground and background. Finally, a finer labeling method is employed by applying a binarization technique. The results show that this enhancement method is partially superior to the unsupervised methods of PCA and ICA. However, our method has not been successful in cases where there were dark and bright areas in the background, since the text line detection algorithm fails to work. We are currently working on a solution to this issue by manual refinement of the text line detection output.

In this article, we also reported on the first application of automated writer identification on Glagolitic manuscripts. Our approach uses Fisher kernels on visual vocabularies. The documents are initially cropped to images containing only text. In the next step, the SIFT features are calculated on the image. The Fisher kernel is then generated with the aid of the GMM. The documents are organized by their similarity into a reference document, and the evaluation uses the k-nearest neighbors algorithm. The best performance was achieved by applying binarization to the image. We plan to improve the cropping method in future. Since the current method only allows the extraction of rectangular regions, decorative elements such as initials are also found in the cropped image. This means that the SIFT features are calculated on non-text regions, too, which makes a correct identification of the writer harder.

ACKNOWLEDGMENTS

The research mentioned in this article was funded by the Austrian Science Fund (FWF): P23133.

REFERENCES

Bar-Yosef, I., Beckman, I., Kedem, K., and Dinstein, I. (2007), 'Binarization, Character Extraction, and Writer Identification of Historical Hebrew Calligraphy Documents', *International Conference on Document Analysis and Recognition* (ICDAR), 9.2–4: 89–99.

——, Hagbi, N., Kedem, K., and Dinstein, I. (2009), 'Line Segmentation for Degraded Handwritten Historical Documents', *International Conference on Document Analysis and Recognition* (ICDAR), 10: 1161–1165.

Bensefia, A., Paquet, T., and Heutte, L. (2003), 'Information Retrieval-based Writer Identification', in *International Conference on Document Analysis and Recognition* (ICDAR), 7: 946–950.

Bianco, G., Bruno, F., Salerno, E., Tonazzini, A. (2010), 'Quality Improvement of Multispectral Images for Ancient Document Analysis', in *Digital Heritage. Third International Euro-Mediterranean Conference* (Berlin: Springer; Lecture notes in computer science, 6436), 29–34.

Brauers, J., Schulte, N., and Aach, T. (2008), 'Multispectral Filter-Wheel Cameras: Geometric Distortion Model and Compensation Algorithms', *Transactions on Image Processing,* 17.12: 2368–2380.

Bres, S., Eglin, V., and Volpilhac-Auger, C. (2006), 'Evaluation of Handwriting Similarities Using Hermite Transform', in *Tenth International Workshop on Frontiers in Handwriting Recognition*, 664–673.

Brink, A., Smit, J., Bulacu, M., and Schomaker, L. (2012), 'Writer Identification Using Directional Ink-trace Width Measurements', *Pattern Recognition,* 45.1: 162–171.

Bulacu, M. and Schomaker, L. (2007), 'Automatic Handwriting Identification on Medieval Documents', in *Proc. of the 14th Int. Conf. on Image Analysis and Processing*, 279–284.

Diem, M. and Sablatnig, R. (2010), 'Recognizing Glagolitic characters', in *Proc. of the 32nd Workshop of the Austrian Association for Pattern Recognition*, 39–46.

Easton, R., Knox, K., Christens-Barry, W. (2003), 'Multispectral Imaging of the Archimedes Palimpsest', in *32nd Applied Image Pattern Recognition Workshop*, 111–118.

——, Knox, K., Christens-Barry, W. (2011), 'Spectral Image Processing and Analysis of the Archimedes Palimpsest', in *19th European Signal Processing Conference*, 1440–1444.

Fiel, S., Hollaus, F., Gau, M., Sablatnig, R. (2014) (forthcoming), 'Writer Identification on Historical Glagolitic Documents', in *21st Document Recognition and Retrieval Conf.* (San Francisco).

Hollaus, F., Gau, M., Sablatnig, R. (2012), 'Multispectral Image Acquisition of Ancient Manuscripts', in *Progress in Cultural Heritage Preservation. Fourth International Euro-Mediterranean Conference*, 30–39.

——, ——, —— (2013), 'Enhancement of Multispectral Images of Degraded Documents by Employing Spatial Information', in I*nternational Conference on Document Analysis and Recognition* (ICDAR), 12: 145–149.

Hyvärinen, A., and Oja, E. (2000), 'Independent Component Analysis: Algorithms and Applications', *Neural Networks,* 13.4–5: 411–430.

Lettner, M., Diem, M., Sablatnig, R., Miklas., H. (2007), 'Registration of Multispectral Manuscript Images as a Prerequisite for Computer-Aided Script Description', in *12th Computer Vision Winter Workshop*, 51–58.

Liang, H. (2012), 'Advances in Multispectral and Hyperspectral Imaging for Archaeology and Art Conservation', *Applied Physics A: Materials Science & Processing,* 106: 309–323.

Louloudis, G., Stamatopoulos, N., and Gatos, B. (2011), 'ICDAR 2011 Writer Identification Contest', *International Conference on Document Analysis and Recognition*, 11: 1475–1479.

Lowe, D. G. (2004), 'Distinctive Image Features from Scale-Invariant Keypoints', *International Journal of Computer Vision*, 60.2: 91–110.

Miklas, H., Gau, M., Kleber, F., Diem, M., Lettner, M., Vill, M., Sablatnig, R., Schreiner, M., Melcher, M., Hammerschmid, E-G. (2008), 'St. Catherine's Monastery on Mount Sinai and the Balkan-Slavic Manuscript-Tradition', in H. Miklas and A. Miltenova (eds.), *Slovo. Towards a Digital Library of South Slavic Manuscripts* (Sofia: Bulgarian Academy of Science, Institute of Literature), 13–36.

Otsu, N. (1979), 'A Threshold Selection Method from Grey-level Histogram', *Transactions on Systems, Man and Cybernetics*, 9: 62–66.

Perronnin, F., and Dance, C. (2007), 'Fisher Kernels on Visual Vocabularies for Image Categorization', in *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR '07), 1–8.

——, Sanchez, J., and Mensink, T. (2010), 'Improving the Fisher Kernel for Large-scale Image Classification', in *11th European Conference on Computer Vision*, 143–156.

Rapantzikos, K. and Balas, C. (2005), 'Hyperspectral Imaging: Potential in Non-destructive Analysis of Palimpsests', in *IEEE International Conference on Image Processing*, 2: II, 618–621.

Salerno, E., Tonazzini, A. and Bedini, L. (2007), 'Digital Image Analysis to Enhance Underwritten Text in the Archimedes Palimpsest', *International Journal on Document Analysis and Recognition*, 9.2: 79–87.

Su, B., Lu, S., and Tan, C. L. (2010), 'Binarization of Historical Document Images Using the Local Maximum and Minimum', in *International Workshop on Document Analysis Systems*, 159–166.

Wolf, L., Litwak, L., Dershowitz, N., Shweka, R., and Choueka, Y. (2011), 'Active Clustering of Document Fragments Using Information Derived from both Images and Catalogs', in *International Conference on Computer Vision*, 1661–1667.

Article

Interdisciplinary Perspectives from Material and Computer Sciences on the Dead Sea Scrolls and Beyond

Daniel Stökl Ben Ezra | Paris

Abstract

This paper will present the current state of the application of material and computer sciences to the Dead Sea Scrolls of Qumran and point out possible implications for other fields. It will also focus on the importance of working in interdisciplinary approaches to achieve further synergies.[1] After a brief introduction to the corpus, it will discuss present and future possibilities related to the four classical tasks of editing fragmentary papyri and will then address three further topics in which IT and material sciences could advance research.

1. The corpus

The Dead Sea Scrolls encompass about 2,000 manuscripts found during the second half of the last century, mostly in caves in the area near the western shore of the Dead Sea in Israel and the Palestinian territories.[2] The oldest manuscripts come from about the seventh century BC, the youngest from Islamic times. They were written in different forms of Aramaic, Hebrew, Greek, Latin and Arabic. These are the Dead Sea Scrolls in a broad sense.

When speaking of the Dead Sea Scrolls, many people refer to a specific group: the Qumran scrolls, doubtlessly the most important group, which consists of about 1,000 scrolls. They are actually the feeble remains of what was once a huge ancient Jewish religious library and were discovered between 1946/47 and 1956 in eleven caves about 12 km south of Jericho.[3] The Qumran scrolls date from roughly the third century BCE to the first century CE.[4] So far, it is the largest exclusively religious library known from the ancient Mediterranean world whose remains have been unearthed.[5] The scrolls have revolutionised scholarly perspectives in Biblical Studies, ancient Judaism, early Christianity, Hebrew palaeography, Hebrew language, Jewish book studies and many other fields.[6]

However, speaking of scrolls may create a false impression: 98% of the 'scrolls' consist of tiny fragments that can be as small as a fingernail (or smaller still). Altogether, there must be about 15,000 of them, but nobody has ever counted them. Furthermore, most parts of the majority of scrolls are now lost. Deciding which fragment once belonged to which scroll and where it used to be on that scroll has been one of the greatest puzzles in the history of human culture, not to mention the questions regarding genre, content, authorship and scribal origin of each reconstructed scroll.[7] Answering the questions posed by this extremely complex cultural puzzle will undoubtedly help solve analogous problems with fragmentary finds from many other cultures.

You may compare the task with the following parable. There are two versions of it. Here is the simple version: Your mother-in-law gives you a 'present': She takes eleven big boxes (the eleven caves) and 1,000 puzzles (the 1,000 texts) of 10 to 10,000 pieces each (the fragments). Each puzzle is put in one of the eleven boxes. She puts about 650 of them into the biggest one (Cave 4). Then the contents of each box

[1] For previous surveys, see especially Tov (2011) and Broshi 2004. Many studies of varying quality have been published in Humbert, and Gunneweg 2003; Galor, Humbert, and Zangenberg 2006; Gunneweg, Greenblatt, and Adriaens 2006.

[2] See the catalogue by Tov 2010, *Revised Lists of the Texts from the Judaean Desert*.

[3] See Stökl Ben Ezra 2011.

[4] Paleographical dating schemes have been developed by F. Cross for the book scripts and by J. Milik and by A. Yardeni for the cursive scripts and for Nabatean; Cross, 1961 and 1998; Yardeni 2000. For the so-called Palaeo-Hebrew script of this period, see McLean 1982.

[5] See Stökl Ben Ezra 2009.

[6] See Stökl Ben Ezra, forthcoming.

[7] See Tigchelaar 2012.

are thoroughly stirred. Finally, she throws 95% of the pieces away. This is not the only challenge; many of the pieces are extremely darkened, so one often cannot distinguish ink from parchment with the unaided eye. Your task is to reconstruct the original puzzle for each piece, its size and artist, the image depicted and as many details as possible on it.

For some texts you have images (the known texts, such as the Bible), but for most you do not. Some puzzles show the same image as others, and you can use one partial image to reconstruct another one. This was the mission of a very select group of about eight scholars in the scriptorium.

2. The four classical tasks

The first part of this paper focuses on the four main tasks of the first two generations of scholars (which are in fact tasks of any publication project connected with a major discovery):

1. Transcription of fragments
2. Reconstruction of hypothetical manuscripts with the help of paleography, codicology (e.g., shape matching) and contents
3. Study of textual parallels to derive compositions and their recensions
4. Ideological provenance: study of contents of compositions to discern the author group and web of groups who authored, copied and/or transmitted the scrolls.

One can imagine this as a pyramid of increasing hypotheticality, as shown in fig. 1:[8] The smallest physical unit is a fragment. The next level consists of manuscripts tentatively reconstructed from a group of fragments that join directly or indirectly. Each manuscript is in fact a hypothesis, a construction rather than a reconstruction. Several manuscripts may have so many parallels that it becomes clear that they are different copies of more or less the same composition. Several compositions can in certain lucky cases be attributed by their ideology, rules and sociological descriptions to a specific group.

The task of making a preliminary transcription and establishing initial tentative associations that fragments have with one or another manuscript took researchers about ten years, as we know from the compilation of a handwritten concordance, but this did not result in the publication of most fragments.[9] Then, for several reasons, work slowed down

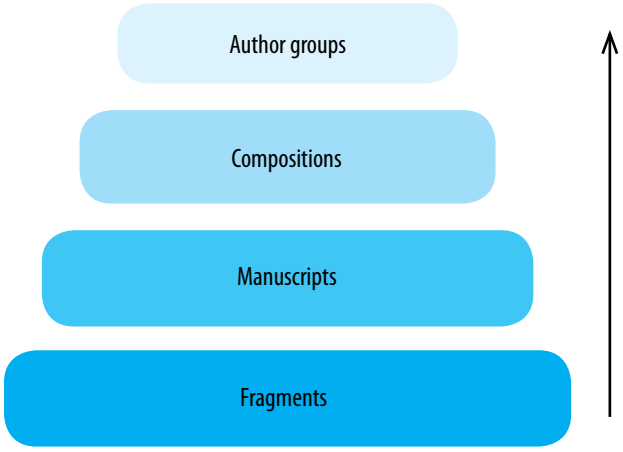


Fig. 1: Pyramid of increasing hypotheticality.

and came to a virtual halt for 30 years until the end of the eighties. Nobody apart from a very small team of about ten people had access to the fragments, photos, transcriptions or the concordance. Originally, this arrangement was probably a good idea for a quick and thorough publication, but it became increasingly problematic and eventually developed into what some people have called one of the biggest academic scandals of the twentieth century. Fanciful conspiracy theories about the Vatican trying to hide texts dangerous to Christian belief came into being. This is, of course, nonsense. Luckily, at this point, computers entered Qumran studies. Having laid their hands on one of the privately published and well-guarded concordances, of which only a handful existed at the time, and having entered the concordance into a computer, a small team around Ben Zion Wacholder and Martin Abegg simply reconstructed the complete fragments in a kind of reverse engineering and began to publish their work.[10] At about the same time, pictures had been made available in print in 'pirate editions'.[11] These were two of the factors in pushing the official editors forward with the publishing process that finally ended about five years ago with the final volume of the official edition. Everybody has access to many editions of all the texts today. In a limited number of cases, computer programs have assisted in the identification of some very small fragments containing only a few letters over several lines with large known texts.[12] However, some of these

[8] Eibert Tigchelaar's suggestion in an oral communication.

[9] Brown et al. 1988. For a history of discoveries and publications, see Fields 2009; Dimant, and Kottsieper 2012. See also Tov 2002.

[10] Wacholder, and Abegg 1991-1996.

[11] Eisenman, and Robinson 1991.

'identifications' or reconstructions can be shown to be highly hypothetical or even wrong.

Obviously, scholars differ – sometimes greatly – not only with regard to how they interpret a certain passage of writing, but also as to whether or not to ascribe a fragment to a particular scroll, or whether to differentiate a group of fragments into one or several scrolls. Designing a database that is capable of representing the different possibilities is quite a challenging task. The chain of philological problems is complex and resembles other complicated manuscript finds.

A) On a paleographical level, a chain of characters on a fragment can be interpreted by one scholar as *N* words with *X* characters, but by another one as *M* words with *Y* characters.

B) On a grammatical level, one scholar derives a word from root *A* linked to lemma *B* explained as form *C* with a syntactic function *D*. Other scholars will come up with other equally plausible explanations that are different.

C) On a level of textual criticism, one manuscript may represent a passage with 5 words in the order 1,2,3,4,5, while a clear parallel in another manuscript can consist of only 4 words in a different order and/or forms and/or lexemes and/or in a different language.

D) On a codicological level, one fragment can be linked by one scholar to a group of fragments *I* from place *J*, while other scholars will place it somewhere else and/or link it to a different group of fragments.

E) On an ideological level, one text may be interpreted as a copy of composition *Alpha*, while another scholar considers it a copy of composition *Beta*.

F) While some alternative suggestions can clearly be discarded as wrong, others may be judged as equally probable. One also needs a system that takes the hierarchization of proposed philological solutions into consideration.

Ingo Kottsieper from the Forschungsstelle Qumranwörterbuch (Qumran Dictionary Research Project) of the Academy of Sciences in Göttingen, a project directed by Annette Steudel and Reinhard Kratz, has succeeded in the complex task of constructing a database that can take all the problems mentioned above into account. Moreover, all parallels can be noted and texts interlinked. Different readings of a given passage can be aligned and displayed. Critical editions of any passage can be produced on the fly. Many scrolls bear several names, revealing various levels of confidence with regard to our knowledge of contents and/or genre. This state-of-the-art database can handle all these tasks.

While still rather exceptional in manuscript studies of other languages in the late forties and early fifties, infrared photography helped in distinguishing the background from the ink from the very beginning.[13] Today, these PAM photos are extremely important not only for reconstructions of scrolls (because many of the fragments have deteriorated over the last 50 years), but also for recapitulating the reconstruction process of the fifties.

Shortly after its invention, radiocarbon dating was tested on materials from Qumran in 1950 and confirmed their dating to the turn of the era in a very general way.[14] Before the invention of Accelerated Mass Spectrometry (AMS), a great amount of material was needed – and destroyed – for such tests, so testing was almost exclusively done on peripherals of the scrolls: first and foremost on the wrapper, and later on palm wood. With the invention of AMS, the amounts needed for C14 tests decreased. Several tests have been undertaken on about 30 scrolls since the early nineties.[15] In a general way, these tests have confirmed the paleographical scheme suggested by the late Frank Cross, which was based on comparison with very few other finds, including ones in other related scripts, and on general typological observations (see below).[16] In some cases, we have to assume that the sampling was biased, probably due to the application of modern castor oil by the early paleographers, which they sometimes used to enhance the contrast between the ink and the blackened parchment.[17]

In manuscript studies, the classical papyrologists are arguably the most advanced with regard to shared and openly accessible catalogues, downloadable high-quality images and crowd-sourcing platforms for manuscript descriptions

[12] See Pfann 2001, 213–225.

[13] Bearman, Pfann, and Spiro 1998; Zuckerman 2010.

[14] Libby 1951.

[15] The best explanation of the possibilities and limits of C14 dating is probably van Strydonck et al. 2000. For the first C14 analyses on Qumran scrolls, see Bonani et al. 1991 and 1992; Jull et al. 1995.

[16] Bibliography given above: fn 4.

[17] Doudna 1998. Rasmussen et al. 2006 and 2001; Atwill, and Braunheim 2004; van der Plicht 2007.

and transcriptions.[18] Qumran studies are not quite there yet, but the situation has considerably improved over the last decade. Today, all Qumran scholars work with one of two commercial databases on a daily basis: either *Accordance* or *DSSEL*, which include one transcription, one translation, grammatical analysis and one picture for most fragments, most of them as 300 or 900 dpi scans of the old infrared photos from the fifties.[19] Many related texts in different languages and from different periods, such as Bibles in various languages, Josephus, Philo, some pseudepigrapha and some Rabbinic literature complete the *Accordance* suite. Researchers using many other linguistic and cultural corpora, such as those available in Buddhist studies, can only dream of a similar infrastructure. Even so, a free online infrastructure usable by any group in any culture and to which different research institutes could provide plug-ins would still be a huge step forward.

Since September 2011, the Shrine of the Book website permits access to color photos of five of the almost complete scrolls, including some annotations for some 1QIsaa.[20] In December 2012, the Israel Antiquities Authority in collaboration with Google constructed the Leon Levy archive, a website that permits free internet access to scans of the historical infrared photos as well as some color photos[21] and, since February 2014, also to new high-resolution multispectral images of the scrolls. These photos have a 24-megapixel resolution in 1,215 ppi and were taken in 12 wavelengths from 445 nm to 924 nm. For some wavelengths, additional images with light coming from different angles were added. The images are freely viewable online, which is an immense step forward. They are not downloadable yet (a considerable deficit compared with the Greek papyrological world), but the IAA has kindly agreed to make them freely available to researchers upon request. Despite the fact that the state of preservation of many fragments has deteriorated since the old IR images were made, the multispectral images of the current state will lead to the identification of some new letters that were unreadable using the old techniques.

There is as yet no collaborative platform for transcribing, translating, commenting or annotating the Scrolls. The International Image Interoperability Framework (IIIF) developed at Stanford University would be very interesting in this respect.[22] IIIF provides, on the one hand, a standard for selected regions on an image (a rectangle that can be turned around) or, more precisely, on a canvas that can be linked to a stack of several pictures of the same object, which is ideal for the multispectral images of the IAA. The Shared Canvas and Mirador viewers developed in collaboration between Stanford and the Bibilissima project in Paris are very promising steps in the direction of an infrastructure.[23]

Other IT tools that would be extremely practical should facilitate the very time-consuming reconstruction process of the Stegemann method, which is based on destruction patterns and the regularly decreasing size of the revolutions of a scroll from the outside to the inside.

Patch pictures of recto and verso of fragmentary opisthographs (scrolls written on both sides) could also be done automatically.[24] The Tel Aviv team (see below) is pursuing work in this direction in a very promising way. Computer pattern analysis has great potential for reconstruction work on highly fragmentary papyrus by matching the papyrus fiber pattern of non-contiguous fragments from the same papyrus sheet using 'join distances' at the same height (for the recto pattern) or width (for the verso pattern). Up to now, such work has largely and very painstakingly been done by hand, for example by the Egyptologist Kim Ryholt for the Tebtunis papyri and by Barns and Pfann for 4Q249.[25] Together with the team around Lior Wolf and Nahum Dershowitz from Tel Aviv and Jonathan Ben Dov from Haifa, we have started a research project that should (if successful) considerably alleviate the reconstruction task through a process that includes the relative distribution of the fiber pattern coupled with the height and distribution of the inscribed lines similar

[18] One of the major projects is the shared platform www.papyri.info. The catalogue portal www.trismegistos.org is also useful, especially together with the *Leuven Database of Ancient Books* www.trismegistos.org/ldab, a catalogue of literary papyri; and finally there is the *Heidelberger Gesamtverzeichnis* for documentary papyri: http://www.rzuser.uni-heidelberg.de/~gv0/.

[19] Tov, and Parry 2005, *DSSEL*: *Dead Sea Scrolls Electronic Library*. Accordance: by Oaktree software: http://www.accordancebible.com/buzz/articles/dss_index.php.

[20] www.dss.collections.imj.org.il.

[21] www.deadseascrolls.org.il.

[22] www.iiif.io.

[23] Members of the Shared Canvas project at Stanford University and the Biblissima project at the Campus Condorcet in Paris are collaborating on creating an infrastructure.

[24] We are planning to construct such a macro at the EPHE.

[25] Barns 1977, 29 and Pfann 2000, 517–523; Ryholt 1999; 2006; 2013.

to the matching analysis of dendrochronological patterns. Of course, such an analysis would largely be independent of the type of script on the papyrus and could also be used by scholars working on completely different corpora. Promising preliminary results have already been obtained.

Perhaps computer pattern analysis can also be helpful in discerning the script of two fragments. However, as the fragments are often very small with only a few letters and are deformed three-dimensionally, a number of very complex issues will have to be solved before digital paleography can become a major factor in the reconstruction of tiny fragments. RTI (Reflectance Transformation Imagery)[26] coupled with 3D reconstruction could potentially help in overcoming these difficulties. It is particularly well suited for visualizing surface structures such as papyrus fibers and scratching.

In the long run, automatic paleography will arrive at OCR-like capabilities of a usable – albeit far from perfect – quality compared with human paleography. Since the Dead Sea Scrolls have been deciphered, this is important for other large corpora with unidentified and untranscribed texts, such as the cuneiform archives and libraries, Greek, Demotic and Coptic papyrology, Syriac and Geniza studies. Even before reaching an acceptable OCR level, these methods will enable computers to provide preliminary identifications of large collections of texts. Computer linguistics could also be of enormous help if joined to automatic paleography. The texts of the Qumran corpus have already been analyzed, so this is less relevant to them.

Material studies have enabled scholars to advance in ascribing fragments to specific manuscripts. Stephen Pfann has developed a hair-follicle pattern analysis, which examines the curve and distribution of hair follicles and compares them to the constant follicle pattern in the animal species that provided the skin to arrive at probable placements of a parchment fragment on a sheet.[27] I am not sure whether this method has been applied to or tested on manuscripts from other cultures. RTI could be helpful in hair follicle analysis. DNA could enable us to discern whether two fragments come from sheets that were cut from the hide of the same animal or not.[28] These hides of mainly goats, sheep and calves had an average size of about 60 cm x 90 cm, enough for two or more sheets, depending on the height of each sheet.[29] While this does not necessarily allow the ascription of two fragments to the same sheet,[30] knowing that two compositions were written on material from the same animal would obviously be most interesting regarding the hide trade, scroll production ateliers and the proximity of various scrolls or scribes during production.[31] This is a piece of information at least as important as the ascription of this or that fragment to this or that scroll.

Whether two fragments do *not* belong to a sheet can also be examined by applying X-ray fluorescence spectroscopy (XRF). A number of studies have been undertaken by the group around Oliver Hahn and Ira Rabin at the Bundesanstalt für Materialforschung (BAM Federal Institute for Materials Research and Testing) in Berlin.[32] It is important to note that XRF can only disprove certain claims: If the profile of chemical elements of two given fragments can be shown to be very different, then the two fragments most likely did not belong to the same skin, as we have seen in the case of 4Q413/4Q413a.[33] If XRF shows a similar profile for both fragments, they could also come from the same scribal atelier, but not necessarily from the same sheet. In antiquity, some skins waited longer than others before being prepared as writing support. These skins had to be treated with salt that was then mostly, but not entirely, washed off in the process of preparing them for writing. In the 2,000 years since their production, the remaining salt has crystallized on the surface of these fragments. With methods capable of discerning salt crystals, one can distinguish between salted and unsalted skins, thus providing an additional hint for sorting fragments, as was shown by the Berlin team.[34]

[26] See p. 25ff of the revised and updated version of Zuckerman's article quoted above, available online at: http://www.usc.edu/dept/LAS/wsrp/information/DynamicsDSS/.

[27] So far, only a brief summary has been published at http://orion.mscc.huji.ac.il/orion/programs/taskforce.shtml.

[28] On a new technique to arrive cheaply and quickly at a sample for DNA analysis with minimal involvement with the manuscript, see the contribution by XY in this volume. Also Woodward et al. 1996.

[29] According to Tov, most sheets had a length of about 30–40 cm (Tov 2011, 23); for 1QIsaa he gives a figure of 35–45 cm (2004, 80). For a scroll like 1QIsaa with a length of 28 cm, one could therefore produce about four to six sheets from one hide. For scrolls of greater length (e.g. 40–45 cm), the same hide size would only produce two long sheets.

[30] Tov 2011, 23–24.

[31] Tov 2011, 25.

[32] Rabin 2013.

[33] Hahn et al. 2007.

[34] Wolff et al. 2012.

3. Part Two

As other communications in this volume show as well, the physical remains of these texts obviously contain much more information about the owners than only the ideas expressed in the writing. Having pointed out some issues in which IT and material sciences and, especially, their combination could lead to progress in the classical tasks of editing fragmentary manuscript finds, let us now transit to three further complexes where material sciences have the potential to open alleys into largely untrodden fields. The fifth question of primary importance in the publication of the Scrolls or any other manuscript corpus is:

5. Scribal provenance: study of scribes and codicology to discern scroll production and networks or schools of scribes inside and outside of Qumran and the movement of the scrolls, scribes and/or owners of the library.

Of course, medievalists have addressed these questions for a long time. In Qumran's case, the main attempt has been Tov's controversial thesis of a special Qumran orthography.[35] Yet, provenance and clustering into schools is the first major question on which we can advance much further with the help of material studies and computer pattern analysis, especially if used in conjunction with each other, even though traditional paleography and codicology continue to be of fundamental importance.

DNA analysis could provide us with extremely valuable information regarding the kinship of the animals that provided the scribes with skins for making parchment. The more closely the DNA of the animal skins of two scrolls is related, the more probable it becomes that these skins came from the same production center or scribal school. Recent years have seen very few studies on DNA in the Qumran texts and none specifically for this purpose.

One of the biggest breakthroughs in the last decade has come from a different method in material sciences: In 2009, the team run by Oliver Hahn and Ira Rabin gave the definitive answer to the long-standing controversy as to whether the Scrolls came from a community that lived and worked locally at the site of Qumran or from a mixture of libraries in Jerusalem and elsewhere. The implications are huge. Local scrolls were certainly written (but not necessarily authored) by local Essenes, a splinter group of great intellectual but little numerical importance. If the scrolls are not local, however, they more likely bear witness to wider Judaism.[36] In antiquity, ink was prepared by grinding a solid block of ink on a stone, thereby creating a powder that was subsequently mixed with water. An XRF analysis of the composition of the ink of one fragment has shown an unusually high bromine to chlorine ratio in the ink, *but not* in the parchment. Such a high ratio of bromine to chlorine only appears in water from the Dead Sea area. Therefore, the Berlin team was able to show that at least this fragment had been written with local water. Establishing water, ink, parchment and papyrus profiles for other areas could undoubtedly help in provenancing scribal activity in various cultures and periods. It could also serve as a fundamental point of departure for studying the history of the science of writing and the commerce in scribal materials. Let me add that the XRF results by no means imply that all scrolls come from Qumran. In fact, the philological studies of the last two decades have shown that the majority of compositions were not authored by the local Essenes, even though they may have copied some. Furthermore, there is now a consensus that there were several other Essene sites in Judaea (and perhaps beyond). Fragments from at least a number of other scrolls have been investigated by the same team using XRF analysis and some other techniques such as Raman and FTIR.[37] The results point to a non-Qumranic origin for some of these scrolls. Additional studies of a similar nature are of the utmost importance to our understanding of the collection and the socio-intellectual network behind them. Which scrolls are local, which are not? These questions are, obviously, of the greatest interest to anybody working on material coming from a circumscribed collection.

Another task is establishing a palaeographical typology that would permit scrolls to be dated, almost all of which are in an undated book script. In the almost complete absence of fixed pegs, this dating has had to rely on the typological development of the Judean script and developments among its cousins, such as Nabatean.[38] After the fall of the Persian Empire, the once homogenous Aramaic chancellery script split up into local types in the various parts of the political successor entities until about the first century CE. Luckily, we are speaking of a 'hot' period of scribal development

[35] See the discussions in his *Scribal Practices and Approaches*. The most thorough critique has been that of Tigchelaar 2010.

[36] Rabin, Hahn, Wolff, and Masic 2009.

[37] Rabin, forthcoming.

[38] See above, fn. 4.

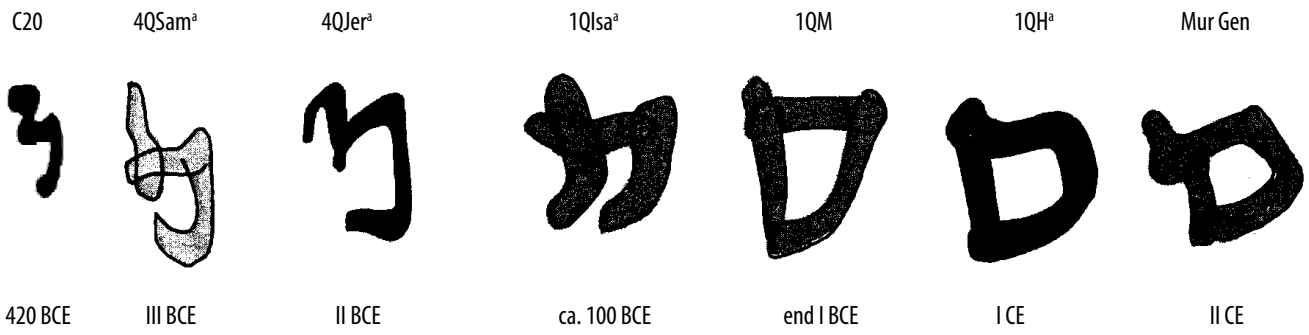


Fig. 2: Typological development of the *samekh* over time.

with a fast pace, when relative typologies of what is closer to Aramaic and what is closer to later square script can be established with more confidence compared with 'cold' periods such as the Middle Ages, when the different regional types used for Hebrew script changed much less.[39] This can be illustrated with the typological development of the *samekh*, which gradually lost its head on the left side and became a closed entity, as fig. 2 shows.

Still, the palaeography of early Judean scripts is an area that is needs to be revisited. I see the greatest potential for pattern analysis in providing 'big data' that is numerical and, therefore, can be analyzed and evaluated with statistics and lead to scribal clustering (hands, styles, schools, geographical and/or chronological distribution).

One idea would be an evaluation of the proposed paleographical evolution scheme for dating manuscripts. Can we actually derive a logical sequence from the proximity that puts the different stages of Judean script on a timeline between the Aramaic and the square script? When does the neat scheme become too noisy? Do we have to, or rather, can we ascribe certain features to differences between contemporaneous schools? Can we describe the transition between cursive and formal styles and their intermediaries mathematically? Can we arrive at the level of distinction between scribes for texts as fragmentary as the Qumran scrolls? Another result could be some kind of check system that tells us where the current attribution of fragments to hypothetical manuscripts makes sense or whether one should reconsider comparisons with another group of fragments.

Bernd Neumann, Rainer Herzog and Arved Solth at CSMC, University of Hamburg, have done some very promising work on stroke extraction, automatic recognition of lines and automatic recognition of inscribed areas of interest in manuscripts.[40] Their method seems to be particularly well adapted to the nature of Hebrew, which (similar to Chinese) is written in mostly rather well separated characters or sometimes short chains of characters connected by a nexus.

The Friedberg Genizah project team, whose highly impressive and innovative IT component is being developed by the group around Lior Wolf and Nahum Dershowitz, has been able to provide scholars with tools of stunning effectiveness for join matching by means of automatic codicological and palaeographical analysis and for automatic classification of scripts such as Ashkenazi, Italian and Oriental.[41] The Tel Aviv team has also done automatic alignment of existing transcriptions of Hebrew manuscripts with ink traces on images using algorithms developed by Tamar Lavee.[42] Of course, this has great potential not only for training the pattern recognition algorithms but also for updating classical digital editions for the purpose of quickly publishing previously transcribed texts directly linked to photos of these manuscripts.

Another objective that could be achievable with the help of a large database and computer paleography would be to study the development of one scribe's hand over a short time (from the first to the last column of a scroll) and over the long period of his lifetime. This is of the greatest interest to anybody working on scribal attribution, archives or biographies. Again, a triangle involving material studies and IT would vastly surpass studies that only employ one

[39] The drawings are taken from Yardeni 2003.

[40] Solth, Neumann, Stelldinger 2009; Herzog, Neumann, Solth 2010.

[41] www.genizah.org. Wolf 2010; Shweka et al. 2011.

[42] Lavee 2013 (available at: http://www.cs.tau.ac.il/thesis/thesis/Lavee.Tamar-MSc.Thesis.pdf).

methodology without the other. Computer-supported studies of the development of the script of one scribe would naturally benefit significantly from a network that assembles the greatest amount of data possible from precisely dated and/or provenanced sources in other scripts whose scribe is known and then extrapolates general variation algorithms for the less well known.

The GIWIS and *Monk* projects led by Lambert Schomaker in Groningen, Holland, have already assembled a vast amount of data on Latin scripts and are a hugely important step in this direction. According to Schomaker, GIWIS is able to *typologically* date medieval Latin manuscripts within periods of 25 years.[43] This does not mean that the manuscript actually comes from such a short time span, only that it is typologically in between two points that have been established with the help of a huge amount of paleographically analyzed manuscripts. It goes without saying that such a database should be openly accessible to everyone.

Another objective could be the question of whether a scribe wrote manuscripts in different scripts or even alphabets. This is a very challenging problem currently unsolved in the world of classical paleography. To establish a ground truth, it might be a good idea to assemble a large database of modern manuscripts written by the same scribes, but in different alphabets (e.g. Greek, Latin) or scripts (e.g., cursive, book script).

Beyond a doubt, philology, pattern analysis and material studies need to be applied to these questions in a combined approach to methodology. Coupling the results of material studies with traditional paleography and IT cluster analysis, one could establish a chronological profile that could provide insights into the history of activity of the site and possible waves of imports of outside manuscripts. This strategy is, of course, transferable to any manuscript collection whose production is linked to a particular place, no matter whether it is Ugarit, Elephantine, Herculaneum, Nag Hammadi, Athos, St. Gall, Timbuktu or the monastery of the Dharmaguptaka sect in Nagarāhāra (where the Gandhari manuscripts presumably came from). How many recipes for ink or parchment preparations are attested? How many scribes can we discern? When did local manuscript production start, peak or decline? Which ingroup manuscripts do *not* come from the center studied and point to the existence of further centers? Which outgroup manuscripts come from the center? What proportion of ingroup and outgroup compositions did the scribes at the center write and when? Did local ingroup scribes write in all scripts, languages, styles and genres?

For the analysis of local schools, a triangle that combines philology with a database and the analyses generated by both material studies and IT has great potential for reinforcing any argument, e.g., if the pattern analysis data matches the chemical element profile. If it does not match, this poses further questions and could result in some interesting conclusions with regard to scribal migration or networks of scribal schools. Similarly, for the transmission of techniques and the development of writing styles, a combination of IT and material sciences could open up new avenues of investigation. Are specific text groups, distinguished by ideology, genre, dialect, orthography etc. linked to specific ink or parchment recipes, scribal schools or scribes? Of course, such a working program would imply substantial financial support, but the general idea is transferable to any project of major complexity with great historical implications.[44]

Let me move on to a rather more complex version of the parable mentioned at the beginning of this paper for which material studies are indispensable. The situation of the Qumran data is, in fact, more complicated than pointed out at the beginning: Let's say your mother-in-law did not actually give you the eleven boxes, but only six of them (all minor ones). Instead, she transferred most of the boxes – and all of the important ones – to a group of smugglers (the Bedouin). Furthermore, she also gave the same group of smugglers additional boxes with other puzzles that she had prepared, not for you but for your brother-in-law (other sites from the Judaean desert). When you buy the pieces from the smugglers one by one, *they* are your sole source of information regarding which box they found a piece in. Sometimes the smugglers lie. At least, you are lucky enough to strike a deal with the smugglers so that 99% of the pieces

[43] Oral information by L. Schomaker. For some of his publications, see Schomaker et al. 2010; Schomaker et al. 2007.

[44] The École Pratique des Hautes Études in Paris has forged a cross-cultural and interdisciplinary working group of palaeographers skilled in the scripts of many major human cultures (cuneiform, Chinese, Japanese, Arabic, Hebrew, Greek, Latin, Tibetan, Coptic, Syriac, Central Asian scripts). Our first workshop was held on June 3, 2014 in Paris and concerned the theory and practice of dating methods. Throughout its history, the École Pratique has always had a tradition of quality and quantity with regard to scholars with this somewhat old-fashioned expertise in Europe and beyond. Often, giving chronological or geographical indications about the provenance of a manuscript is unjustifiably belittled as a *Hilfsdisziplin* (ancillary discipline). We are looking forward to suggestions for collaboration with experts from other universities in the coming years, be they experts in traditional palaeography, computer-supported pattern analysis, image treatment or material studies.

end up in your collection; other people (tourists and private collections) only get 1%. So, to the fragmentary state of the findings, we also have to add the difficulty of the black-market intermediary and unprovenanced data.

6. Transmission provenance: attribution of fragment to manuscript collection – origin in one or another specific cave – ancient, medieval and modern interventions.

In order to understand the collections in the caves and their relationships to the inhabitants of Qumran, proof of the authenticity of the provenance in this or that cave is crucial. We are certain that for some fragments the Bedouin have given us wrong information about the provenance, however, since some fragments purportedly from Qumran can physically be shown to belong to fragments from Murabbaat and Nahal Hever, sites of great importance further to the South. They contained Jewish archives from the Second Revolt and even personal letters by the leader of that uprising, Bar Kokhba, from about a century after the destruction of Qumran.

Scholars of the first generation (i.e., Strugnell) claimed to be able to recognize fragments from Cave 11, a major cave containing some very important scrolls. The fragments had a bitter taste (!) and a strong odor. Ira Rabin claims to have found the chemical compounds responsible for these features.[45] The large mass of bat guano accumulated over millennia marked the scrolls. Many scrolls were inevitably in contact with bat guano, which left microscopic but detectable traces on the fragments. The mass of guano was so large, in fact, that the resulting ammonia gas even penetrated into jars and under textiles and contaminated the outmost layers of otherwise protected scrolls such as the Temple Scroll. Was Cave 11 the only cave with bats in it? We need further chemical fingerprints to confirm the origin of fragments from other caves, especially for Cave 4. We may even eventually arrive at the conclusion that one of the so-called Qumran cave manuscript assemblies does *not* in fact belong together with the others. One challenge we face is that these methods will have to distinguish traces that are informative about ancient deposits from deposits that belong to either (1) modern conservation, (2) deterioration or (3) scroll production. Rather like policemen at the scene of a crime, archaeologists should always leave some of their material uncleaned and preserve the earth at the excavation site as it is so their finds remain in their original context.

If we transpose this possibility to other cultures, it would be trying to find traces of whether a manuscript was once in a specific city or library (e.g., St. Catherine's) or belonged to a specific scholar. I think it is possible in certain — but probably rare — cases to obtain fingerprints for the history of transmission, for example through ink analysis of marginal annotations, restored letters or bindings. Special places, their meteorological, biological (parasites!) or physical conditions and the techniques of their workers may leave special deposits behind.

7. Modern Conservation: Preservation.

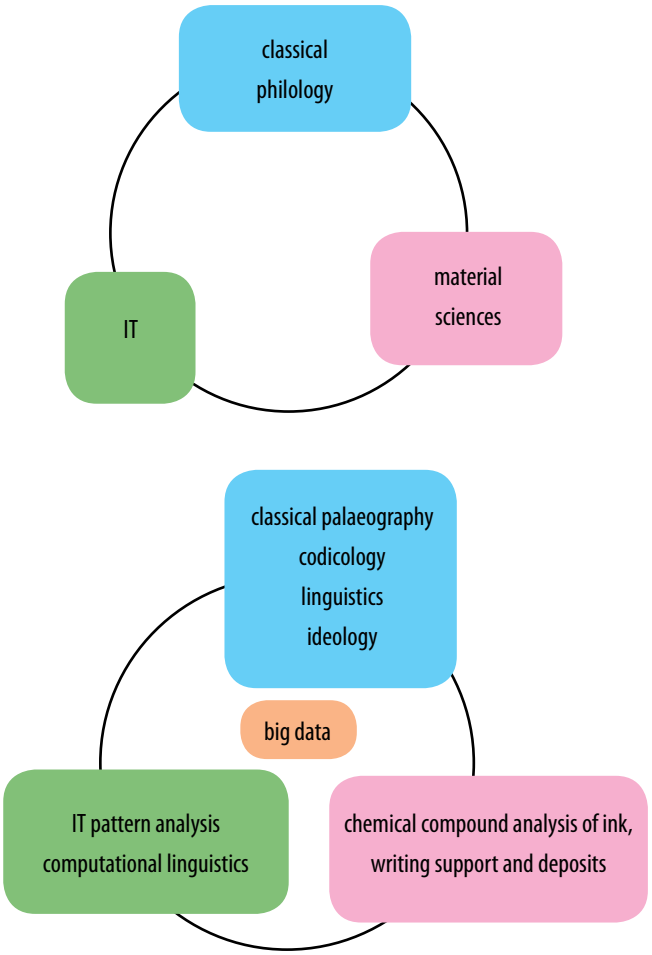
Knowing more about the effects of transmission will obviously also lead to progress in modern preservation techniques.

4. Conclusions

The readable text is only a chapter in this history that remains to be written. The material features, many of them invisible to the human eye, are a chapter of at least the same importance in disentangling the complex history of the Dead Sea Scrolls. Bringing these issues together will only be possible by employing a combination of techniques consisting of classical philological analysis, material studies and image pattern analysis. One could also add complex data management to this (as in the example of the Göttingen database) and, of course, a true willingness to collaborate and share data in an infrastructure open to everyone.[46] The Tel Aviv team's algorithms for finding joins are so successful because they combine the results of several automatic analyses with paleographical, codicological and content information. Analogously, combing the results of XRF analyses of ink and parchment with automated paleography and DNA testing based on new imaging methods will undoubtedly lead not only to greater accuracy, but, through geographical, chronological and stylistic clustering, will also enable us to ask completely new questions regarding production techniques, transmission and interdependence between scribal schools.

[45] Rabin et al. 2010.

[46] A database assembling all the results of examinations of inks, parchment, papyrus and paper would be of great importance here, especially if it also contained details about pottery gained with the help of material sciences (e.g., XRF, DNA, NAA), similar to what exists in dendrochronology, C14 dating and DNA sequencing.



REFERENCES

Atwill, J., and Braunheim, S. (2004), 'Redating the Radiocarbon Dating of the Dead Sea Scrolls', *Dead Sea Discoveries,* 11: 143–157.

Barns, J. (1977), 'Note on Papyrus Fibre Pattern', in R. de Vaux, J. T. Milik, *Qumran Grotte 4, vol. 2, I. Archéologie, II. Tefillin, Mezuzot et Targums (4Q128-4Q157)* (Oxford: Clarendon; Discoveries in the Judaean Desert, 6).

Bearman, G., Pfann, St. J., and Spiro, S. (1998), 'Imaging the Scrolls: photographic and direct digital acquisition', in P. Flint and J. Vanderkam (eds.), *The Dead Sea Scrolls after Fifty Years: A Comprehensive Assessment* (Leiden: Brill), 472–495.

Bonani, G., Ivy, S., Wolfli, W., Broshi, M., Carmi, I., and Strugnell, J. (1991), 'Radiocarbon Dating of the Dead Sea Scrolls', *Atiqot,* 20: 27–32.

——, Ivy, S., Wolfli, W., Broshi, M., Carmi, I., and Strugnell, J. (1992), 'Radiocarbon Dating of Fourteen Dead Sea Scrolls', *Radiocarbon,* 34: 843–849.

Broshi, M. (2004), 'The Dead Sea Scrolls, the Sciences and New Technologies', *Dead Sea Discoveries,* 11: 133–142.

Brown, R., et al. (1988), *A Preliminary Concordance to the Hebrew and Aramaic Fragments from Qumran Caves II–X: Including Especially the Unpublished Material from Cave IV. editorum in usum* (Göttingen: privately published).

Cross, F. (1961), 'The Development of the Jewish Script', in G. Wright (ed.), *The Bible and the Ancient Near East. Essay in Honor of William Foxwell Albright* (Garden City, NY), 133–202.
—— (1998), 'Palaeography and the Dead Sea Scrolls', in P. Flint and J. VanderKam (eds.), *The Dead Sea Scrolls After Fifty Years: A Comprehensive Assessment* (Leiden: Brill), vol. 1, 379–402.

Dimant, D., with the assistance of I. Kottsieper (eds.) (2012), *The Dead Sea Scrolls in Scholarly Perspective: A History of Research*, 2 vols. (Leiden: Brill; Studies on the Texts of the Desert of Judah, 99).

Doudna, G. (1998), 'Dating the Scrolls on the Basis of Radiocarbon Analysis', in P. W. Flint and J. C. VanderKam (eds.), *The Dead Sea Scrolls after Fifty Years: A Comprehensive Assessment*, (Leiden: Brill), vol. 1, 430–471.

Eisenman, R. H., Robinson, J. M. (1991), *A Facsimile Edition of the Dead Sea Scrolls*, 2 vols. (Washington, D.C.: Biblical Archaeological Society).

Fields, W. (2009), *The Dead Sea Scrolls: A Full History*, vol. 1 (Leiden: Brill).

Galor, K., Humbert, J.-B., and Zangenberg, J. (eds.) (2006), *Qumran, the Site of the Dead Sea Scrolls. Archaeological Interpretations and Debates* (Leiden: Brill; Studies on the Texts of the Desert of Judah, 57).

Gunneweg, J., Greenblatt, Ch., and Adriaens, A. (eds.) (2006), *Bio- and Material Cultures at Qumran* (Stuttgart: Fraunhofer IRB-Verlag).

Hahn, O., et al. (2007), 'Non-Destructive Investigation of the Scroll Material: 4QComposition Concerning Divine Providence (4Q413)', *Dead Sea Discoveries,* 15: 359–364.

Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens einschließlich der Ostraka usw., der lateinischen Texte, sowie der entsprechenden Urkunden aus benachbarten Regionen (HGV) (http://www.rzuser.uni-heidelberg.de/~gv0/).

Herzog, R., Neumann, B., Solth, A. (2010), 'Computer-based Stroke Extraction in Historical Manuscripts', *manuscript cultures*, 3: 14–24.

Humbert, J.-B., and Gunneweg, J., (eds.) (2003), *Khirbet Qumran et Ain Feshkha: Études d'anthropologie, de physique et de chimie* (Fribourg: Academic Press; NTOA.SA 3).

Jull, A. et al. (1995), 'Radiocarbon Dating of Scrolls and Linen Fragments from the Judean Desert', *Radiocarbon,* 37: 11–19.

Lavee, T., *Computer Analysis of the Dead Sea Scroll Manuscripts* (M.A. thesis, Tel Aviv University 2013; available at: http://www.cs.tau.ac.il/thesis/thesis/Lavee.Tamar-MSc.Thesis.pdf).

Leuven Database of Ancient Books (LDAB) (www.trismegistos.org/ldab, a catalogue of literary papyri).

Libby, W. (1951), 'Radiocarbon dates II', *Science,* 114: 291–296.

McLean, M. (1982), *The Use and Development of Palaeo-Hebrew in the Hellenistic and Roman Periods*, (unpubl. Ph.D. diss., Cambridge, Mass: Harvard University).

Pfann, St. J. (2000), *Qumran Cave 4: Cryptic Texts* (Oxford: Clarendon Press; Discoveries in the Judaean Desert 26).

—— (2001), *The Character of the Early Essene Movement in the Light of the Manuscripts Written in Esoteric Scripts from Qumran* (Ph.D. diss., Hebrew University of Jerusalem).

van der Plicht, J. (2007), 'Radiocarbon Dating and the Dead Sea Scrolls: A Comment on "Redating"', *Dead Sea Discoveries,* 14: 77–89.

Rabin, I., Hahn, O., Wolff, T., and Masic, A. (2009), 'On the Origin of the Ink of the Thanksgiving Scroll (1QHodayota)', *Dead Sea Discoveries,* 16: 97–106.

—— et al. (2010), 'Analysis and preservation of an Antique Alum-tawed Parchment', *ICOM-CC*, (Working Group 'Leather and related materials', Rome).

—— (2013), 'Archaeometry of the Dead Sea Scrolls', *Dead Sea Discoveries*, 20: 124–142.

—— (forthcoming), 'Material Analysis of the Fragments', in T. Elgvin (ed.), *Gleanings From the Caves. Dead Sea Scrolls and Artifacts of The Schoyen Collection*.

Rasmussen, K. L. et al. (2001), 'The Effects of Possible Contamination on the Radiocarbon Dating of the Dead Sea Scrolls I: Castor Oil', *Radiocarbon*, 43: 127–132.

—— et al. (2006), 'Cleaning and Radiocarbon Dating of Material from Khirbet Qumran', in Gunneweg, Greenblatt, and Adriaens, *Bio- and Material Cultures* (Stuttgart: Fraunhofer IRB-Verlag), 139–163.

Ryholt, K. (1999), *The Carlsberg Papyri. 4: The story of Petese, son of Petetum, and seventy other good and bad stories (P. Petese)* (Copenhagen: Museum Tusculanum Press; CNI publications 23).

—— (2006), *The Carlsberg Papyri. 6: The Petese stories II (P. Petese II)* (Copenhagen: Museum Tusculanum Press, CNI publications 29).

—— (2013), *The Carlsberg Papyri. 10: Narrative Literature from the Tebtunis Temple Library* (Copenhagen: Museum Tusculanum Press; CNI publications 35).

Schomaker, L., et al. (2007), "Text-independent writer identification and verification using textural and allographic features', *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 29: 701–717.

—— et al. (2010), 'Towards robust writer verification by correcting unnatural slant', *Pattern Recognition Letters*, 32: 449–457.

Shweka, A., et al. (2011), '"Bring them both close together": Handwriting Identification and Computer-assisted Reconstruction of Genizah Fragments', *Ginzei Qedem,* 7: 171–204.

Solth, A., Neumann, B., Stelldinger, P. (2009), 'Strichextraktion und -analyse handschriftlicher chinesischer Zeichen', in *Report FBI-HH-B-291/09* (Department of Informatics, University of Hamburg).

Stökl Ben Ezra, D. (2009), 'Archives and Libraries. II Greco-Roman World, New Testament and Early Christianity', in *Encyclopedia of the Bible and Its Reception* (Berlin – New York: De Gruyter), vol. 2, 683–687.

—— (2011), 'Wieviele Bibliotheken in Qumran?', in J. Frey et C. Claussen (eds.), *Qumran und die Archäologie* (Tübingen: Mohr-Siebeck; WUNT 278).

—— (forthcoming), *Lehrbuch Qumran* (Tübingen: Mohr Siebeck).

van Strydonck, M. et al. (2000), 'Rapport du Groupe de travail : Les limites de méthode du carbone 14 appliquée à l'archéologie', in *Actes du 3ème Congrès International 14C et Archéologie, 6–10 Avril 1998, Lyon, Mémoires de la Société Préhistorique Française, XXVI & Supplément 1999 de la Revue d'Archéométrie*, 433–448.

Tigchelaar, E. (2010), 'Assessing Emanuel Tov's 'Qumran Scribal Practice', in S. Metso, H. Najman and E. Schuller (eds.), *The Dead Sea Scrolls: Transmission of Traditions and Production of Texts* (Leiden: Brill; Studies on the Texts of the Desert of Judah,92), 173–207.

—— (2012), 'Classifications of the Collection of Dead Sea Scrolls and the Case of Apocryphon of Jeremiah C', J*ournal for the Study of Judaism*, 43: 519–550.

Tov, E. (2002), 'The Discoveries in the Judaean Desert Series: History and System of Presentation', in id. (ed.), *The Texts from the Judaean Desert: Indices and an Introduction to the Discoveries in the Judaean Desert Series* (Oxford: Clarendon; Discoveries in the Judaean Desert, 39).

—— (2004), *Scribal Practices and Approaches reflected in the Texts found in the Judean Desert* (Leiden: Brill; Studies on the Texts of the Desert of Judah, 54).

——, Parry, D. (eds.) (2005), *Dead Sea Scrolls Electronic Library (DSSEL)* (Leiden: Brill).

—— (2011), 'The Sciences and the Reconstruction of the Ancient Scrolls: Possibilities and Impossibilities', in A. Lange, E. Tov, and M. Weigold, with B. H. Reynolds III (eds.), *The Dead Sea Scrolls in Context: Integrating the Dead Sea Scrolls in the Study of Ancient Texts, Languages, and Cultures* (VTSup 140/1; Leiden – Boston: Brill), 3–25.

—— (2010), *Revised Lists of the Texts from the Judaean Desert* (Leiden – Boston: Brill).

Trismegistos. An interdisciplinary portal of papyrological and epigraphical resources dealing with Egypt and the Nile valley between roughly 800 BC and AD 800 (www.trismegistos.org).

Wacholder, B.-Z., Abegg, M. G. (1991–1996), *A Preliminary Edition of the Unpublished Dead Sea Scrolls: The Hebrew and Aramaic Texts from Cave Four*, 4 fascicles (Washington D.C.: Biblical Archaeological Society).

Wolf, L., et al. (2010), 'Automatically Identifying Join Candidates in the Cairo Genizah', International Journal of Computer Vision, 90: 1–18.

Wolff, T., et al. (2012), 'Provenance Studies on Dead Sea scrolls Parchment by Means of Quantitative Micro-XRF', *ABC,* 402: 1493–1503.

Woodward, S. R., et al. (1996), 'Analysis of Parchment Fragments from the Judean Desert Using DNA Techniques', in D. W. Parry and S. D. Ricks (eds.), *Current Research and Technological Developments in the Dead Sea Scrolls* (Leiden: Brill; Studies on the Texts of the Desert of Judah, 20), 215–238.

Yardeni, A. (2000), *Textbook of Aramaic, Hebrew and Nabatean Documentary Texts from the Judaean Desert and Related Material* (Jerusalem).

—— (2003), *The Book of Hebrew Script* (New Castle, DE: Oak Knoll Press).

Zuckerman, B., 'The Dynamics of Change in the Computer Imaging of the Dead Sea Scrolls and Other Ancient Inscriptions', in M. L. Grossman (ed.), *Rediscovering the Dead Sea Scrolls: An Assessment of Old and New Approaches and Methods* (Grand Rapids: Eerdmans, 2010), 69–88.

PICTURE CREDITS

Fig. 2: © Ada Yardeni, *The Book of Hebrew Script*, New Castle 2003.

s

Article

The Basics of Fast-scanning XRF Element Mapping for Iron-gall Ink Palimpsests

Leif Glaser and Daniel Deckers | Hamburg

Abstract

Synchrotron radiation X-ray fluorescence (srXRF) mapping of elements is a good tool for digitising iron-gall ink handwriting, even if the ink has been covered or erased, as was often the case in the Middle Ages in order to re-use pieces of parchment. In this paper, the influence of the excitation energy on the measuring process and resolution will be discussed, showing that 17 keV of excitation energy and a resolution of more than 100 dpi give the best results. Two typical systems of re-used parchment in book bindings were investigated with mock-up samples in a test, one with written parchment in close contact with wood, the other with leather. The results are discussed here with a special focus on the evaluation of what the minimum requirements of a dedicated set-up would have to be to make this method mobile, using an X-ray tube as the light source instead of a storage ring beam.

1. Introduction

In the Middle Ages, the scribal practice of re-using parchment produced numerous palimpsests, manuscripts that contained a newly written text on top of an erased older one. Iron-gall ink was the predominant choice for producing the historical manuscripts under consideration here, and in many cases, the original text was erased chemically. By removing the gallic acid from the organo-metallic compound responsible for the ink's bluish-black colour,[1] the remaining ink was rendered more or less transparent to visible light. This method left all the metallic compounds (mainly metal sulphates) of the old iron-gall ink in the parchment, making it appear more or less unused. Thus the parchment could be re-used to produce a second manuscript, although oxidation of the remaining iron content of the old ink would frequently lead to the eventual reappearance of the old text in a yellowish-brown tint, sometimes clearly readable, sometimes only as a faint trace. In the 19th century, chemicals were used to enhance the readability of the erased script on many of the remaining manuscripts, yielding some stupendous results in the short term, yet often resulting in damage to both parchment and texts (old and new alike) in the long term. Some less invasive, non-destructive approaches that also provide good results in recovering the older script are the use of UV light (since the early 20th century, both for examination and photography), multispectral imaging[2] and other optical imaging methods. In cases where the use of UV light or multispectral imaging will not provide adequate results or is rendered futile by solid layers of paint on top of the older text, for example, another approach that can be considered non-destructive[3] is the use of X-ray fluorescence spectroscopy employing a monochromatic hard X-ray light source of very high intensity (only available in storage rings today), which has proved to be the perfect tool in digitising and visualising hidden texts written in iron-gall ink.[4] Since the first successful experiments on the Archimedes Palimpsest,[5] erased text in several palimpsests has been deciphered as a result of using the synchrotron radiation XRF method, which always requires the documents to be transported to a storage ring facility. There are still a large range of objects that cannot be investigated, however, including manuscripts that are not available for transport to a storage ring facility, even for the short duration of the measurements, due to considerations relating to manuscript preservation, the manuscript's value or library and archive policies.

The same storage-ring-based XRF method was used for

[1] Krekel 1990.

[2] Easton et al. 2010.

[3] Young 2005.

[4] Bergmann 2011.

[5] Bergmann 2007.

the investigation of hidden paintings and led to similarly spectacular results.[6] As with manuscripts, in these cases the often considerable value of the objects investigated (the under-drawings of most interest to researchers tending to be underneath famous paintings) can be one of several reasons for wishing to avoid transport to a storage ring facility. Given the rather intense signals solid paint layers produce, the use of mobile X-ray-lab-source-based XRF set-ups has proved to be possible.[7] Changing the light source from a storage-ring-based system, which is highly monochromatic and most often linearly polarised, to a non-monochromatic unpolarised X-ray source, as most mobile systems are, the quality of the XRF spectrum recorded is decreased dramatically, especially for trace elements.[8] Hence storage-ring-based XRF images of paintings are still of significantly higher quality, even if the mobile equipment available as a prototype at the University of Antwerp today is perfectly sufficient for most paintings.

However, in the case of the slight traces of the erased inks and thus much weaker XRF signals emitted during the measurement of manuscripts as well as the higher spatial resolution required for the results, it is significantly harder to make use of a mobile set-up. As a first step pertaining to the eventual choice of the most suitable X-ray source, we have investigated the minimal resolution required for such a system as well as the dependence of readability contrast of the element maps produced with srXRF spectroscopy on the excitation energy used during the measurements.

The set-up used was not state-of-the-art in synchrotron fast XRF mapping as used in the case of the Archimedes Palimpsest, for example.[9] Today's srXRF upper limit is to measure with a resolution of 600 dpi and illumination times of around 3 ms per spectrum using optimised equipment and a highly brilliant X-ray source with suitable focusing. In principle, it is possible to enhance the readability of the measured data, separating different inks due to their non-iron metal impurities using methods such as principal component analysis or non-negative matrix factorisation.[10] Additionally using XRF detectors on both sides of the illuminated parchment, the nature of the fluorescent light emitted in essentially every direction allows one to separate the signal coming from the front and back of the parchment.[11] This has not been done in the measurements presented in this paper, nor has post-data processing been performed to enhance readability for the data presented in this paper, focusing on the possibilities of the XRF mapping technique itself and the necessary minimal operating parameters for a transportable set-up based on a laboratory X-ray source. The limitations of an X-ray tube to a state-of-the-art beamline at a storage ring facility is in our case mainly in focus, flux, monochromaticity and polarisation. The non-monochromatic flux of modern laboratory X-ray sources is comparable to the monochromatic flux of the bending magnet beamline used for the presented measurements, the focus of those sources is limited to a diameter of roughly 100 µm, hence the resolution achievable with such a beam size was to be analysed. The lack of linear polarisation of light from an X-ray tube would make the otherwise advantageous positioning of the XRF detector at right angles to the light in the polarisation plane useless. The scattered X-rays are minimal in that direction for linear polarised X-ray excitation and energy-dispersive XRF detectors used to perform XRF mapping experiments measure both, hence the background is lowest in this detector geometry. The results obtained with non-monochromatic light[12] show reduced thresholds for trace-element detection – a change from a storage-ring-based source to lab equipment may limit the number of non-iron impurities one may use for analysis to the metallic compounds with sufficient concentration. The latter two effects were not part of the investigation discussed in this paper.

A palimpsest manuscript which we had previously measured fully to identify the erased undertext[13] was available for energy-dependent test measurements. In addition, several mock-up iron-gall ink texts were produced in the lab on modern goatskin parchment, applying the ink with glass ink pens or goose-feather quills. No differences in the XRF maps could be seen as expected between the two ink application methods, with the glass pens being the more convenient to write with and producing thinner lines that were more even, so we stuck to them in the end. Freshly prepared iron-gall

[6] Dik et al. 2008.

[7] Ahlfeld et al. 2011.

[8] Chen et al. 2008.

[9] Bergmann et al. 2012.

[10] Ahlfeld et al. 2014.

[11] Bergmann et al. 2009.

[12] Chen et al. 2008.

[13] Deckers and Glaser 2011.

inks (table 1) were used for our mock-up samples following historical iron-gall ink recipes,[14] but our own sample inks were produced from modern ingredients. To simulate the high impurity of the historic vitriols (containing iron sulphate), we combined the chemically pure compounds, especially various metal sulphates. Iron-gall inks currently in commercial production are based on the essential compounds of the ink, in particular pure iron sulphate, and thus lack the impurities of the vitriol. One modern ink (labelled 'IR' in table 1) was used in comparison. When examining historical inks by methods involving XRF spectroscopy, mapping of these impurities (which sometimes make up more than 50% of the ink's metallic compounds) can often be more interesting than that of the iron signal alone. Using these traces, it is possible to distinguish different inks and/or scribes.[15] To investigate the different points in time at which a document was expanded or revised, for example, this is achieved by determining the minimum number of different inks evident in a specific document. There is always an iron signal to be measured when dealing with iron-gall ink, but the intensities – and sometimes the mere existence (or absence) – of a specific trace element can help one separate the upper from the lower ink, as was the case for calcium in one of our previous investigation, for instance.[16] The words 'upper' and 'lower' in this case refer to the later, fully visible layer of writing and the older, erased one respectively. The gap in production time between the upper and lower layers of text can range from a few centuries to as little as a few decades. As long as the second ink is based on a vitriol of different origin than the one previously used, mapping the non-iron metal impurities will help one to distinguish the upper from the lower text.

Table 1: The metallic content of the prepared inks as an atomic percentage of the total metallic content of the ink. The ink labelled IR was a modern commercial ink.

Ink	Fe	Cu	Mn	Zn	Sum of Mg, Al, K and Ca
1	95	1	1	1	2
2	90	2	2	2	4
3	85	3	3	3	6
4	80	4	4	4	8
5	75	5	5	5	10
6	50	10	10	10	20
7	40	20	5	25	10
8	30	30	5	25	10
9	20	50	5	15	10
10	60	15	5	5	15
11	50	40	2	2	6
12	40	10	10	10	30
13	60	20	0	10	10
14	60	20	10	0	10
IR	100	0	0	0	0

To further investigate specific configurations of historical manuscript materials that have been considered for future measurements, we also tested the effects that solid pieces of wood and leather have on the XRF signal, simulating the re-use situation of parchment leaves glued to a book's leather or wooden cover. For the tests with wood, we applied different inks on parchment and placed the parchment face down on different kinds of wood (oak, beech, pine and balsa), measuring from the back of the parchment. To test the effect of leather, we covered a text on modern parchment with a 2-mm-thick leather cloth and measured the ink through the leather. The XRF scanning technique works fine, even in the presence of strong matrix effects, as demonstrated with the gold paint cover on leaves of the Archimedes Palimpsest or on paleontological samples such as the Archaeopteryx fossil (Bergmann et al. 2010, Wogelius et al. 2011). In principle, it is possible to use a confocal XRF set-up to minimise the matrix effects of a supporting material, as has been done for single-point measurements on wood (Malzer et al. 2004). For scanning areas of parchment with a confocal setting, the layer of the ink within the fast-moving, uneven parchment sample would have to be kept in the focus of the set-up throughout the measurement. This highly challenging task could be achieved by 3D laser scanning the surface of the parchment in advance and then synchronising a 3D sample stage, compensating for the surface structure throughout the mapping. However, a set-up such as this is not available anywhere at present.

[14] Kolar and Stirlic 2007.

[15] Hahn et al. 2004.

[16] Deckers and Glaser 2010.

2. Data acquisition and analysis

All our measurements were performed at Beamline L of the DORIS III storage ring at DESY in Hamburg, Germany. To preserve the parchment, the experimental hutch was acclimatised to 20°C and a relative humidity of 50%. The beam size was collimated to 100 µm vertically and 70 µm horizontally, while the parchment to be measured was at a 45° horizontal angle to the incident beam, thus producing a 100 µm x 100 µm X-ray footprint on the parchment. The chosen beam size was in the order of the minimal step size planned for the measurements and at the lower end of what is today's limit of focal sizes of laboratory-based high-flux X-ray tubes. A VORTEX EM XRF detector was positioned in reflectance geometry in the plane of the polarisation of the light at angles of 45° to the parchment and 90° to the light (fig. 1) to minimise noise due to the detection of scattered X-rays. The parchment could be scanned continuously in the horizontal plane, while XRF spectra were taken at a photon flux of 10^9 photons/second at 7 Hz, resulting in effective illumination times of 0.13 seconds per point. Most measurements were carried out at a distance of 150 µm between two measured points in the plane of the parchment (~170 dpi), resulting in approximately 4,500 spectra taken within 15 minutes for one square centimetre of mapped parchment. The XRF data was processed using the AXIL code,[17] while the element maps were produced and processed using IDL and Photoshop. The elemental maps were all scaled to use the maximum contrast within the individual element map, but no additional processing such as principal component analysis was used since the actual quality of the measured data of interest in this particular experiment can best be interpreted prior to any further alteration of the data.

3. Results and discussion

To optimise scanning times and estimate the spot size for which a laboratory X-ray source should preferably be optimised, the same areas of prepared parchment were measured repeatedly with different step sizes between the measured spectra. The test parchment for this verification was inscribed using a glass ink pen and our home-made ink no. 10. Lines, waves, dots and circles were applied to simulate different abstract parts of writing characters, choosing a line distance in a range proportional to the thickness of the ink strokes. The resulting test object was created to represent manuscript handwriting with character sizes between 3 and 6 mm, a suitable average, even if some small, densely written manuscripts can sometimes exhibit characters with a height as small as 2 mm. The measurement was performed with 17.4 keV of photon energy to record the X-ray fluorescence of all the non-iron metal impurities in the ink. As shown in fig. 2, fair readability is achieved for step sizes of 200 µm and below, while the contrast is insufficient for step sizes of 400 µm and above. The region in between is reasonable for the main metallic compound in the ink (iron), but mostly insufficient where the secondary metallic compound (copper) is concerned. For writing with larger or very simple characters, 300 µm steps (equivalent to approx. 85 dpi) may be sufficient, while for smaller or ornamented characters, the use of 150 µm (or smaller) steps seems more suitable and was used for most of our other measurements.

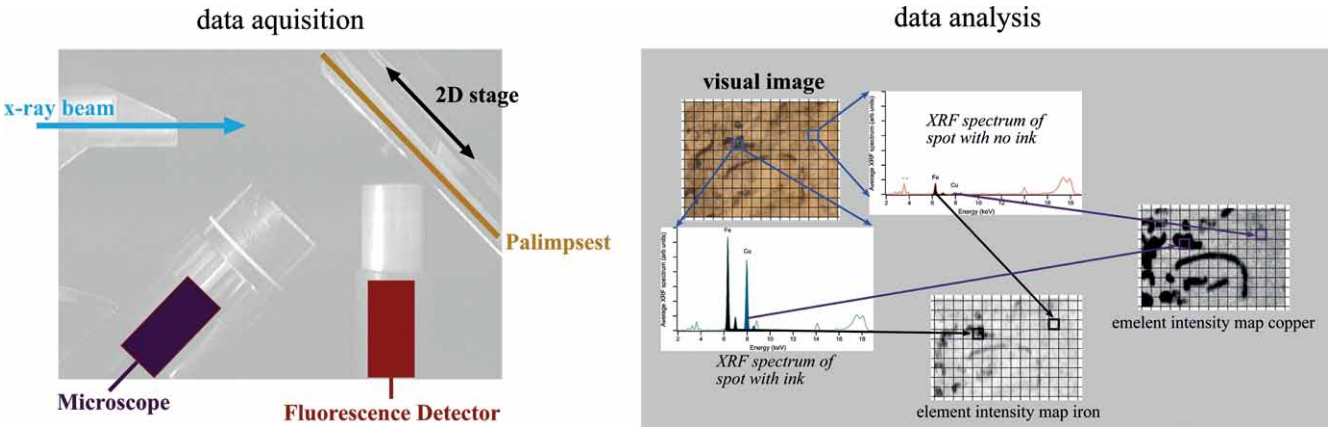


Fig. 1: Experimental set-up on the left-hand side and a schematic description of the analytical steps from image to elemental maps on the right.

[17] Vekemans et al. 1994.

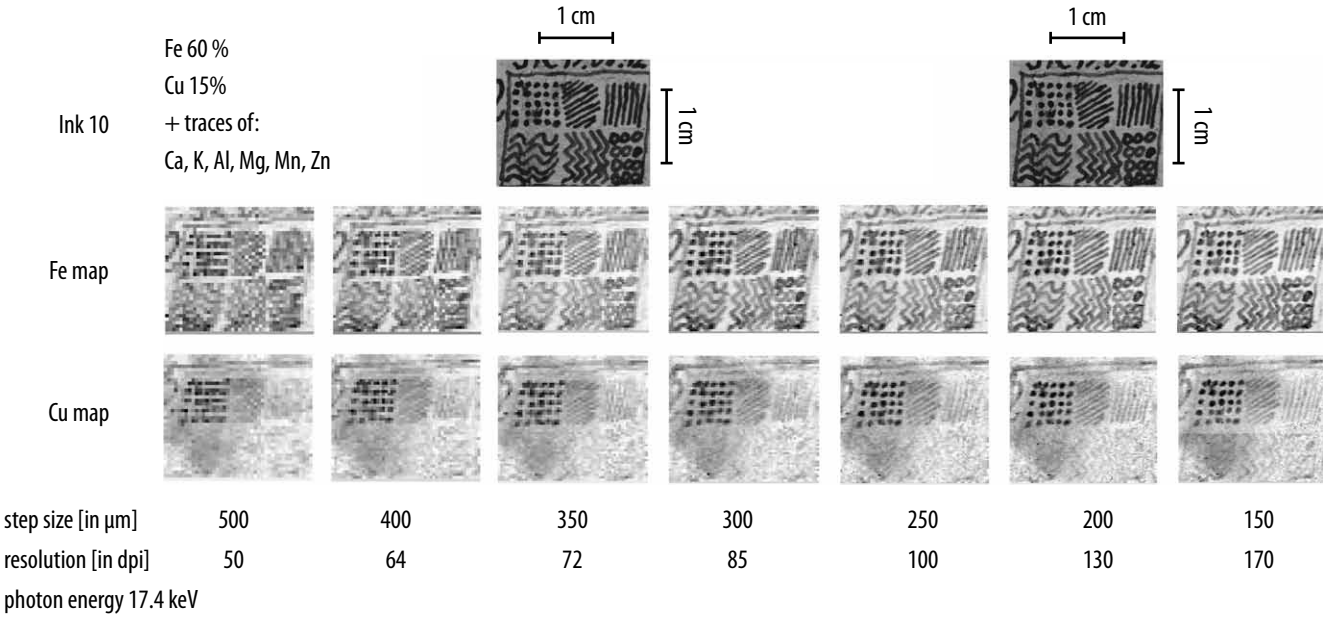


Fig. 2: Simulating a typical sample of handwriting with characters 3 to 6 mm in size, the test sample was prepared on modern thick parchment using circles, dots, lines and zigzags instead of actual letters. Good-quality mapping of the primary metallic ink component (iron) and the main secondary metallic component (copper) can be achieved with sampling steps of 150 to 200 µm, while insufficient contrast is achieved when using steps above 350 µm, even for the main component.

The usual expectation in X-ray absorption spectroscopy is that the signal of a specific element will be enhanced when exciting above but close to the corresponding absorption edges, hence we chose several energies above the iron K-edge, above the copper and zinc K-edges, above the lead M-edges and above the silver K-edge (7.15 / 10 / 17.4 / 31.6 keV). The energy above the silver K-edge was included even though there is no silver in iron-gall ink, as some historical drawings were done using silver-point pens. This was to investigate how much the quality of the iron-gall ink element maps would deteriorate when measuring at an energy level high enough to possibly excite silver-pen lines on the parchment. It was unclear whether the effects of the matrix from the parchment measured would change significantly from lower to higher excitation energies, but due to the chemical composition of parchment we expected to see decreasing effects (less noise) towards higher excitation energies.

As shown in fig. 3, the elemental contrast is quite good for iron when exciting with an energy slightly above the K-absorption edge, but with that energy being too low to

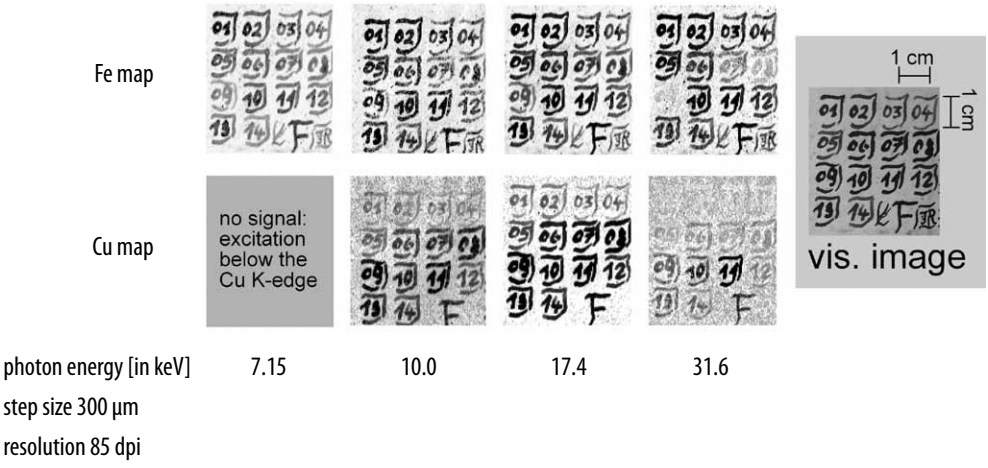


Fig. 3: The elemental contrast is increasing due to the reduced noise from the parchment towards higher photon energies, but when increasing the energy too much, the reduced absorption cross-section of the elements of interest lead to a decrease in contrast, especially for those of the ink's elements only present in low concentrations.

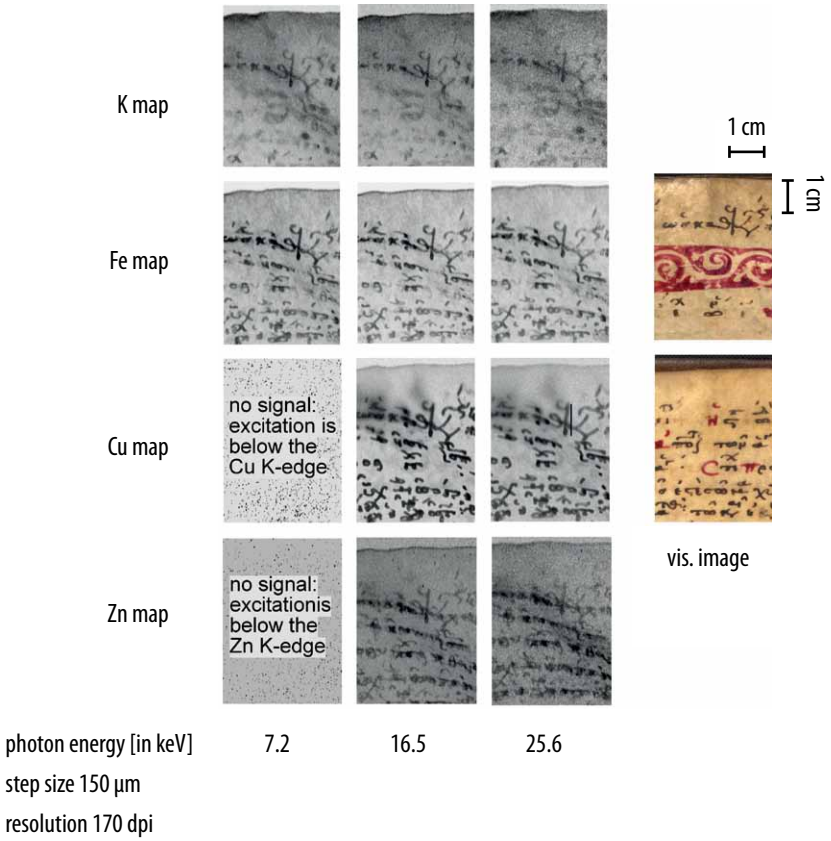


Fig. 4: The readability is quite good at 16.5 keV of excitation energy, even for light elements such as potassium, while contrast decreases for all elements if the excitation energy is too high. The red ink only shows up in the mercury elemental map, but it does not contain any potassium, iron, copper or zinc.

stimulate XRF of heavier elements (such as copper), it is not ideal for historic ink mapping experiments. The matrix effects of the parchment decrease with increasing photon energy, thus producing nicely readable elemental maps at 17.4 keV of photon energy. When using a rather high excitation energy such as 31.6 keV, it is obvious that the quality of the elemental maps decreases and inks with little metallic content (e.g. ink 9 in the Fe map or inks 1–5 in the Cu map) no longer show up in the results. If only one scan of the parchment can be performed (especially if a light source without or of limited tunability is to be used), it therefore seems a good default practice to use monochromatic X-ray photons whose energy is close to 17 keV. Note that for XRF measurements far above the excitation thresholds of the elements, as in the case of iron-gall ink, and excitation energy above 14 keV, slight changes in the excitation energy are not reflected very much in the resulting spectra, hence the excitation energy of 16.5 keV and 17.4 keV can be considered equal in this context. Using a real palimpsest from Leipzig University Library, the energy-dependent investigation was able to be reproduced. Some of the results are shown in fig. 4.

To investigate the effect of wood and leather upon the readability of the measured ink signals, two objects were produced using our 14 inks and the one commercial ink. The measurements with 17.4 keV of photon energy and a step size of 150 µm between two points (fig. 5) clearly show that the contrasts of the used inks are most visible in different elemental maps. The logo in the middle sketched with the commercial ink completely disappears in all but the iron elemental map, while ink no. 12 with a manganese content of only 10% exhibits the best contrast in the Mn element map, and even ink no. 10 with only 15% of copper (and containing four times more iron) is most readable in the Cu element map. It is quite obvious that the parchment used for our test objects contains a high amount of zinc, hence the Zn element map in these examples is not

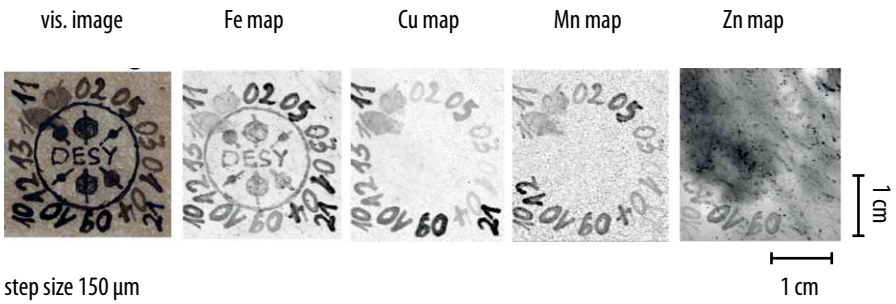


Fig. 5: A test with several differently prepared iron-gall inks (cf. table 1) shows that better contrast can be achieved for some inks from other metallic componentsthan iron. In this case, the modern parchment has a high zinc content, whereas historical texts (in those cases where the ink contains zinc) tend to yield a very good contrast between the writing and parchment in the zinc map.

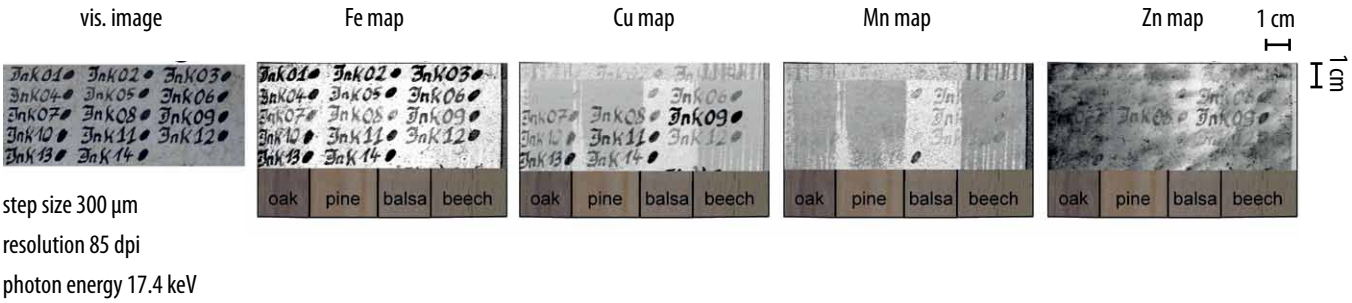


Fig. 6: Testing the method using iron-gall ink writing on modern thick parchment which was put face down onto different types of wood. The results show that the iron and copper maps are quite readable in all cases. The manganese and zinc maps from this test are rather illegible, with the zinc mostly showing the signal from the modern parchment rather than from the structure of the wood.

as useful as that in measurements of writing on historical parchment which typically contains less zinc.

For the test objects with parchment on wood, relatively large letters were used due to the use of a goose-feather quills instead of the glass ink pen employed for the other samples, hence the step size was increased to 300 µm, while all the other parameters were kept the same as before. The structure of typical European wood samples has a strong effect on the elemental maps of copper and manganese (fig. 6), while the balsa wood of tropical origin with little structure and no annual rings only adds a very small amount of noise. Due to the rather low iron content in wood, these effects are minimal in the Fe elemental map.

For the second special use-case examination, we prepared one more mock-up sample with writing and covered it with a 2-mm-thick leather cloth. We recorded the data scanning the ink signals through the leather, simulating the case of a text glued to a leather binding or otherwise obstructed by leather in way that would only allow an examination from the rear side if the structure of the object were to remain unaltered. The leather had a very strong obscuring effect on the contrast of the iron and manganese element maps, but this was not the case for the copper map. The contrast of the elemental distribution in the zinc remained surprisingly high as well, showing those inks with a zinc content of at least 10% quite clearly. It appears that the leather as a cover layer compensates (i.e. absorbs) some of the noise emitted by the modern parchment rich in zinc used for these experiments.

If the parchment is covered with a leather rich in iron and the ink only contains iron, as was the case for the ink used in our experiment for drawing the lines and circles (fig. 7), even a powerful technique such as srXRF has its limits. Fortunately, historical inks are never quite free of non-iron metal impurities, as has been discussed above, and thus far at least two elemental maps in our tests reproduced the inscribed text well.

4. Conclusions

In this paper, we have shown that when measuring iron-gall ink writing with the XRF scanning technique using a highly intense monochromatic source, the best results for element mapping of iron and non-iron metal impurities in examining characters just 3 mm in size can be achieved with

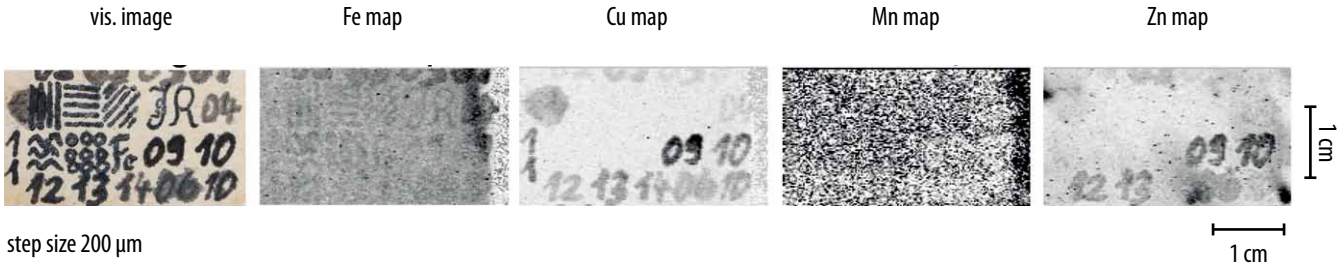


Fig. 7: Testing the method on iron-gall ink writing on modern parchment covered by a 2-mm leather cloth. The high iron and manganese content of the leather produces noise in the corresponding elemental distribution maps, while the copper and zinc channels are mostly undisturbed.

a grid below 200 µm. The best photon energy to use for the experiments is close to 17 keV. If a tunable light source is available, a single map can be improved by re-scanning the entire area for the element in question using a specific photon energy directly above the element's K-shell ionisation threshold (or M-shell threshold for heavier elements), which of course results in doubling the total amount of time for the mapping experiment. It should therefore be avoided if possible. The usefulness of recording further non-iron metal impurities in addition to the iron always present in these inks was demonstrated, showing that the higher-contrast copper and zinc elemental maps are especially valuable in restoring hidden or erased writing. The tests that involved using the scanning methods on a written surface covered by wood or leather proved successful, suggesting that original objects in this condition are suitable for this kind of examination. With respect to the option of using a mobile X-ray source, we have come to the conclusion that the requirements for achieving the best resolution can be met by a mobile source with a focal spot diameter in the order of 100 µm. The most suitable single excitation energy of close to 17 keV for the photons could be achieved using a molybdenum target material. The photon flux needed to scan pages within several days can be produced using at least non-monochromatic laboratory sources. The use of a non-monochromatic photon source that is not linearly polarised in conjunction with the fast-scanning XRF method still remains to be tested; this would enable us to estimate whether the quality of the elemental maps is still good enough to reproduce hidden or erased text. If we tried to keep the measuring time as short as possible, a drop in the quality of the elemental maps would be unavoidable due to changing from a storage-ring-based light source to a mobile X-ray source, but as long as measuring with a non-monochromatised beam of a portable X-ray source is possible, the examination of iron-gall ink writing should yield fairly good results in a dedicated mobile set-up.

ACKNOWLEDGEMENTS

We would like to thank the beamline scientists of Beamline L (Karen Appel and Manuela Borchert) for their great support during the measurements, Leipzig University Library (Christoph Mackert) for both facilitating access to the original manuscripts and assisting in determining safe handling procedures with regard to their preservation, and Ira Rabin for helpful discussions and providing the commercially produced iron-gall ink along with a glass ink pen.

REFERENCES

Alfeld, M., Janssens, K., Dik, J., de Nolf, W., and van der Snickt, G. (2011), 'Optimization of mobile scanning macro-XRF systems for the in situ investigation of historical paintings', *Journal of Analytical Atomic Spectrometry*, 26: 899–909.

——, Wahabzada, M., Bauckhage, C., Kersting, K., Wellenreuther, G., and Falkenberg, G. (2014), 'Non-negative factor analysis supporting the interpretation of elemental distribution images acquired by XRF', *JoP: Conference Series* 499: 012013.

Bergmann, U. (2007), 'Archimedes brought to Light', *Physics World*, *Physics World Archive*, November 2007.

——, Knox, K. (2009), 'Pseudo-color enhanced X-ray fluorescence imaging of the Archimedes Palimpsest', in *Document Recognition and Retrieval XVI*, edited by K. Berkner, L. Likforman-Sulem, *Proc. of SPIE-IS&T Electronic Imaging, SPIE* vol. 7247, 724702-1-13.

——, Morton, R. W., Manning, P. L., Sellers, W. I., Farrar, S., et al. (2010), 'Archaeopteryx feathers and bone chemistry fully revealed via synchrotron imaging', *Proc. Natl. Acad. Sci. USA*, 107: 9060-65.

—— (2011), 'Imaging with X-ray Fluorescence', in Netz, R., Noel, W., Wilson, N., Tchernetska, N. (eds.), *The Archimedes Palimpsest*, vol. 1 (Cambridge University Press).

——, Manning, P. L., Wogelius, R. A. (2012), 'Chemical Mapping of Paleontological and Archeological Artifacts with Synchrotron X-rays', *Annual Review of Analytical Chemistry*, 5: 361–389.

Chen, Z. W., Gibson ,W. M., and Huang, H. (2008), 'High-definition X-ray Fluorescence: Principles and Techniques', *X-ray Optics and Instrumentation*, 2008: Article ID 318171, (doi:10.1155/2008/318171).

Deckers, D., and Glaser, L. (2010), 'Zum Einsatz von Synchrotron-strahlung bei der Wiedergewinnung gelöschter Texte in Palimpsesten mittels Röntgenfluoreszenz', in F. Fischer, B. Assmann (eds.), *Kodikologie und Paläographie im Digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2* (Norderstedt; Schriften des Instituts für Dokumentologie und Editorik, 3), 181–190.

——, and —— (2011), in M. Holappa (ed.), *Eikonopoiia. Symposium on Digital Imaging of Ancient Textual Heritage: Technological Challenges and Solutions*, *Helsinki, Finland, 2010-10-28 – 2010-10-29* (Helsinki: University of Helsinki; Commentationes Humanarum Litterarum 129), 161–171.

Dik, J., Janssens, K., van der Snickt, G., van der Loeff, L., Rickers, K., and Cotte, M. (2008), 'Visualization of a Lost Painting by Vincent van Gogh Using Synchrotron Radiation Based X-ray Fluorescence Elemental Mapping', *Anal. Chem*., 80: 6436–6442.

Easton, R. L., Knox, K. T., Christens-Barry, W. A., Boydston, K., Toth, M. B., Emery, D., and Noel, W. (2010), 'Standardized system for multispectral imaging of palimpsests', *Proc. SPIE7531, Computer Vision and Image Analysis of Art*, 75310D (doi: 10.1117/12.839116).

Hahn, O., Malzer, W., Kannegiesser, B., and Beckhoff, B. (2004), 'Characterization of iron-gall inks in historical manuscripts and music compositions using X-ray fluorescence spectrometry', *X-Ray Spectrometry*, 33: 234–239.

Kolar, J., and Stirlic, M. (eds.) (2006), *Iron Gall Inks: On Manufacture Characterization Degradation and Stabilisation* (Ljubljana: National and University Library).

Krekel, C. (1990), Master's thesis (Institut für Anorganische Chemie der Georg-August-Universität zu Göttingen).

Malzer, W., Hahn, O., and Kanngiesser, B., (2004), 'A fingerprint model for inhomogeneous ink–paper layer systems measured with micro-X-ray fluorescence analysis', *X-Ray Spectrometry*, 33: 229–233 (DOI: 10.1002/xrs.676).

Vekemans, B., Jannsens, K., Vincze, L., Adams, F., and Van Espen, P. (1994), 'Analysis of X-ray Spectra by Iterative Least Square (AXIL): New Developments', *X-Ray Spectrometry*, 23: 278–285.

Wogelius, R. A., Manning, P. L., Barden, H. E., Edwards, N. P., Webb, S. M., Sellers, W. I., Taylor, K. G.; Larson, P. L., Dodson, P., You, H., Da-qing, L., Bergmann, U. (2011), 'Trace metals as biomarkers for Eumelanin Pigment in the Fossil Record', *Science*, 333: 1622–1626.

Young, G. (2005), *Effect of High Flux X-radiation on Parchment*. *Report No. Proteus 92195* (Canadian Conservation Institute) (http://www.archimedespalimpsest.org/pdf/archimedes_f.pdf>, last verified 17-5-2014).

Article

Multispectral Imaging of the San Lorenzo Palimpsest (Florence, Archivio del Capitolo di San Lorenzo, Ms. 2211)*

Andreas Janke and Claire MacDonald | Hamburg

Abstract

This paper details the findings presented at the International Conference on Natural Sciences and Technology in Manuscript Analysis held in Hamburg, Germany in December 2013 on the San Lorenzo Palimpsest (Florence, Archivio del Capitolo di San Lorenzo, Ms. 2211). The San Lorenzo Palimpsest contains over 200 secular compositions from the 14th and the beginning of the 15th centuries of invaluable importance to musicologists. However, most of these pieces have not been studied in detail due to the damage sustained in creation of the palimpsest. Recently, scholars and scientists from the University of Hamburg were granted the opportunity to image the palimpsest using an advanced multispectral system from the SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa' / Centre for the Study of Manuscript Cultures (CSMC) in Hamburg. While the processing and evaluation of each folio is still ongoing, preliminary results are presented in the following sections.

1. Introduction

Today, multispectral imaging has become one of the standard procedures used to cope with damaged manuscripts, especially with palimpsests.[1] 'Standard' means the general way of proceeding here. However, the development of powerful new cameras and scanner systems is still continuing. Besides the fact that new technologies continue to be developed, this field is driven by the manuscripts themselves, which require the development of specific methods as each manuscript is unique and has particular characteristics as a result of its history.

This article seeks to present a case study on imaging the palimpsest Florence, Archivio del Capitolo di San Lorenzo, Ms. 2211. This manuscript is entitled *Campione dei Beni, 1504* and was used to record church properties well into the 17th century. By the end of the 15th century it was made of 111 parchment folios that originally belonged to a music manuscript compiled around 1420 in Florence. Having such a big collection of music from the first few decades of the 15th century in the form of a palimpsest is extremely rare since musical palimpsests from the 15th century usually only survive as fragments or are simply the result of corrections or new scribal initiatives within music manuscripts.[2] Fig. 1 shows an example of how the manuscript appears today with only faint remains of the original musical notation.[3] Nearly all the surviving folios from the original music manuscript contain overwriting, with the exception of fols. 8r+v, 24v, 33v, 79r–80v, 83r, 86r+v, and 90r–109v.

* This article introduces work in progress on the San Lorenzo Palimpsest that will result in the forthcoming publication *The San Lorenzo Palimpsest: Florence, Archivio del Capitolo di San Lorenzo, Ms. 2211. Introductory Study, Commentary and Images*, edited by Andreas Janke and John L. Nádas (Lucca: Libreria Musicale Italiana). We would like to thank Monsignor Angelo Livi, Prior of San Lorenzo in Florence for kindly allowing us to image the San Lorenzo Palimpsest. We would also like to give special thanks to John L. Nádas. Further thanks go to Boryana Pouvkova (who took part in the imaging in Florence) and to Christian Brockmann, Roger L. Easton Jr., Ugo Giani, Giuseppe de Gregorio, Oliver Huck, Vito Lorusso, Sonja Puccetti, Miriam M. Wendling, and Francesco Zimei. The research for this article was carried out within the scope of the work conducted by the SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa' / Centre for the Study of Manuscript Cultures (CSMC) funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

[1] The advantages of this technique regarding palimpsests have been discussed intensively in the past in connection with a variety of projects (Archimedes Palimpsest, Rinascimento Virtuale, Sinai Palimpsest Project).

[2] See for example Nádas and Ziino 1990, Memelsdorff 2004, and Meconi 2011. For a list of music palimpsests from a different context, see Moran 1985.

[3] See for example the remains of the underwriting between the last two paragraphs of the overwriting.

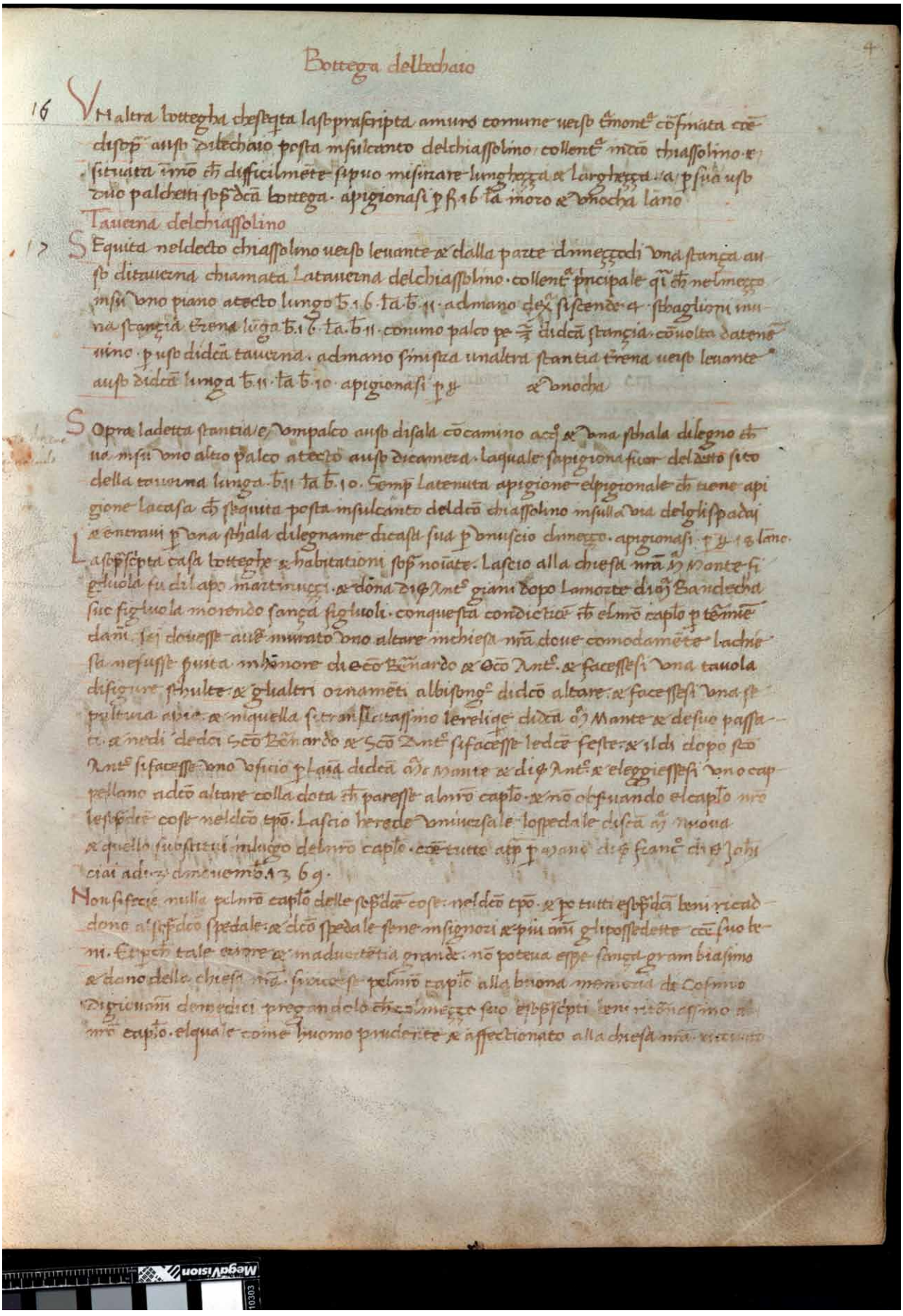


Fig. 1: ASL 2211, fol. 4r; the manuscript as it appears today with only faint remnants of the original musical notation.

The discovery of the San Lorenzo Palimpsest was reported in 1982 by the musicologist Frank A. D'Accone,[4] who emphasised the importance of the manuscript as it contains the remains of what had once been a vast collection of mainly Italian secular polyphonic music. The pieces were composed between the 14th and the beginning of the 15th century and therefore belong to the Trecento repertoire, or the so-called *Ars Nova Italiana*.[5]

Since this discovery, scholars have tried to identify the compositions – a difficult task not only because of the scraped content, but also because of the fact that by the end of the 15th century the manuscript had been completely disassembled in order to be scraped and then put back together in a different order for the *Campione dei Beni*. Therefore the compositions appear in the wrong order today. Furthermore, in many cases, the two or three voice parts of one composition which were originally arranged on one opening[6] are now separated and have to be located. In 1984, John Nádas undertook the first reconstruction of the original gathering structure.[7] The only aids available then were ultraviolet (UV) lamps and later, in 1989, photographs taken under UV light, which led to a revised version of the gathering structure that year.[8]

There is no question about the importance of the collection, which not only includes new readings for compositions known from other contemporary manuscripts, but which, more importantly, contains completely unknown compositions by Italian composers from the beginning of the 15th century, most of whom were connected to the Florentine cathedral *Santa Maria del Fiore* as organists or singers, including Giovanni Mazzuoli († 1426), his son Piero († 1430) and the music theorist Ugolino da Orvieto († 1452).

In 2000, the *Digital Image Archive of Medieval Music* (DIAMM)[9] started the first imaging campaign to provide scholars with high-resolution digital images (natural light and ultraviolet) from the palimpsest, permitting them to apply several digital restoration techniques using software like Adobe Photoshop.[10] However, despite the availability of the DIAMM images and the new techniques for enhancing the musical notation, the contents of the San Lorenzo Palimpsest were usually still described as 'unreadable' and therefore research focusing on the music and the context of this collection was limited.[11]

This article will report on a new imaging campaign – this time using multispectral imaging – that took place in the summer of 2013 in the Archivio del Capitolo di San Lorenzo in Florence. It describes the first steps in developing methods to finally create a publishable set of images that enhance the original musical notation. One of the advantages of using multispectral imaging is that no decision on what information belongs to the underwriting is needed from the scholar in the early stages of the work. However, every change in the image's appearance has to be made transparent to scholars[12] since the final processed images are by no means perfect representations of the original manuscript. Therefore these images will contain unambiguously falsified colours, which not only result in better contrast between over- and underwriting for the reader, but the images cannot be mistaken for the original state of the manuscript.

2. Multispectral imaging

In the past few decades, multispectral imaging has emerged as a vital tool in the recovery of lost writing in manuscripts. The ability to record information beyond what the human visual system can see and interpret is indispensable for cases like palimpsests, where the person who scraped the *scriptio inferior* deliberately tried to remove all visible information. In many ways, the imaging and processing done in this project were inspired by the Archimedes Palimpsest project.[13]

[4] D'Accone 1984.

[5] An overview of the music of the Trecento can be found in Gozzi 2011.

[6] For the layout of music manuscripts, see Schmidt and Vorholt 2009.

[7] Nádas 1992.

[8] John L. Nádas, 'The Lucca Codex and MS San Lorenzo 2211: Native and Foreign Songs in Early Quattrocento Florence', paper presented at the 55th Annual Meeting of the American Musicological Society, Austin, Texas, 27 October 1989.

[9] http://www.diamm.ac.uk

[10] DIAMM has organised many workshops on how to perform digital restoration and also provided a very useful workbook (Craig-McFeely and Lock 2006) which is available online (http://www.diamm.ac.uk/publications/digital-restoration-workbook).

[11] New research for example is found in Huck and Dieckmann 2007, Gehring 2012, 124–134, and Janke 2013 and 2014. A PhD dissertation focusing on the unique compositions in the San Lorenzo Palimpsest is currently being prepared by A. Janke.

[12] For a discussion on ethical considerations in image enhancing, see McFeely 2012.

[13] Christens-Barry, Easton, and Knox 2011.

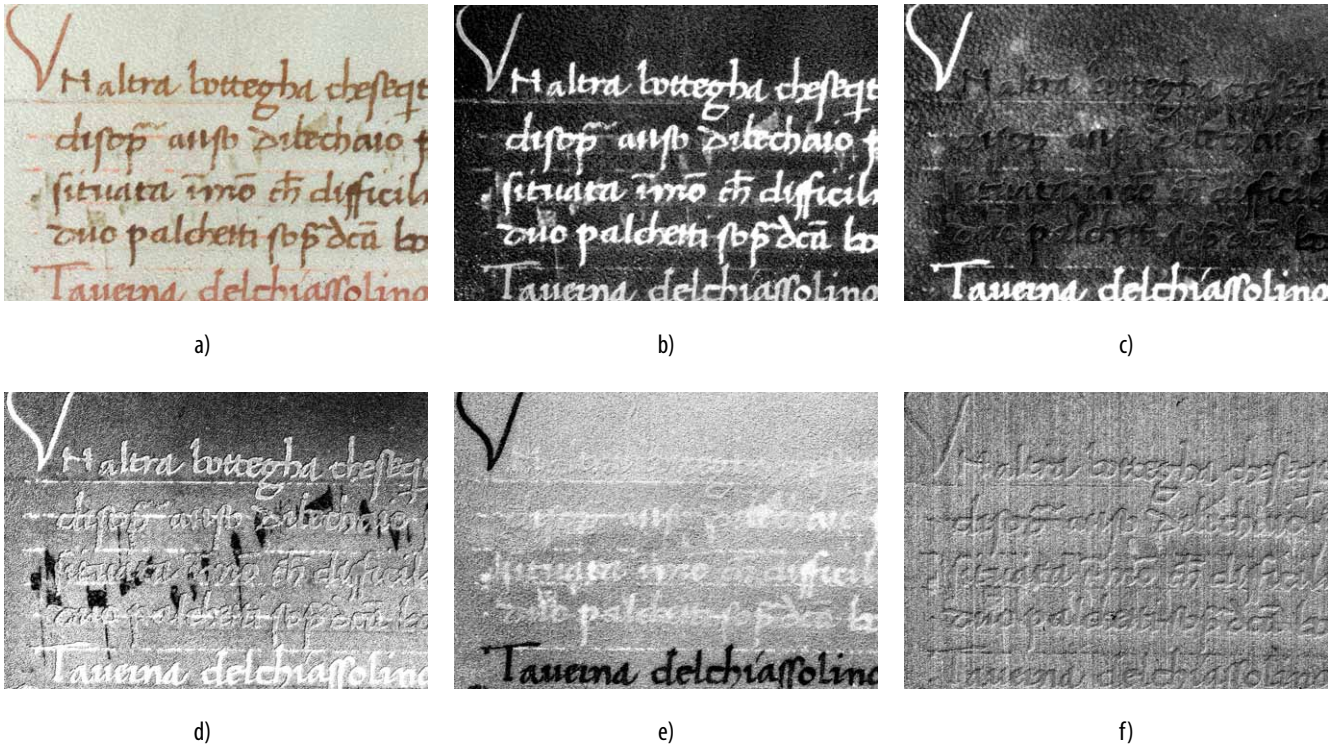


Fig. 2: ASL 2211, comparison of detailed area of fol. 4r: a) appearance under natural light; b) PCA band 1; c) PCA band 3; d) PCA band 6; e) PCA band 10; f) PCA band 20.

The camera and light system[14] used for imaging in this study employs a 50-megapixel monochromatic camera, 13 different wavelengths between 365 and 1,050 nm, five filters and two raking lights in blue and infrared.[15] The system is portable, so it was possible to transport it to the Archivio del Capitolo di San Lorenzo to image on site in one of the archive's rooms prepared to accommodate all the equipment. This included a cradle specifically designed to hold delicate manuscripts.[16] Since it was not possible to remove folios from the binding and produce images of each page individually, an acrylic plate[17] was placed over each folio to keep it standing upright and to keep the book from closing. The entire manuscript was imaged in two weeks, including the binding and notes folded into the cover accompanying the text.

The five filters allow both reflectance and fluorescence images to be captured. Fluorescence describes the phenomenon of light being absorbed and then re-emitted at a longer wavelength (lower energy). Filters are used to isolate a specific band of fluorescence; illuminating the manuscript in ultraviolet light and using a green filter would cause the camera to only capture green fluorescence, for example. Often, parchment fluoresces under UV light, creating an image where the background support appears to glow and anywhere covered by ink remains dark. When viewing the manuscript, distinguishing between light-brown parchment and the faded brown ink of the underwriting can be quite difficult, so the fluorescence images play a critical role in the imaging of palimpsests.

Another problem deals with the issue of 'show through', where it is difficult to distinguish whether a particular note or a group of notes is from the back of the parchment or the side the reader is viewing. In order to mitigate this problem, light must be prevented from reflecting from the page underneath and then being transmitted back through the target page. Since the manuscript could not be unbound to image each folio separately, a sheet of black, acid-free paper was placed between the target folio and the one behind it to absorb any transmitted light and prevent it from being reflected back and travelling through the target folio on its way to the camera.

[14] The system was developed by Megavision (see http://www.mega-vision.com/cultural_heritage.html).

[15] The wavelengths used were 365 nm, 455 nm, 470 nm, 505 nm, 535 nm, 570 nm, 625 nm, 700 nm, 735 nm, 780 nm, 870 nm, 940 nm, and 1,050 nm. The filter set included UV pass, UV block, red, green and blue.

[16] The Traveller's Conservation Copy Stand was developed by Manfred Mayer on behalf of VESTIGIA, the Manuscript Research Center at Graz University, Austria.

[17] Theoretically, the acrylic plate would increase scattering under UV light, but the impact is negligible in practice.

A total of 24 different reflectance and fluorescence images were taken of each folio, which were flattened (a form of pre-processing and calibration) on capture with images taken of a blank, white target. All of these captured images, except four which were raking-light images,[18] were included in the current processing work using the statistical methods of principal component analysis (PCA) and occasionally independent component analysis (ICA)[19] implemented in the Excelis ENVI software package.

3. Processing

The task of rendering captured multispectral images into a legible, publishable set of images of a palimpsest manuscript involves several steps. First, statistical techniques are applied to differentiate different kinds of information in the captured images. Second, perception-based decisions are made to render the results from the statistical techniques into a series of legible images for musicologists or other scholars.

Both PCA and ICA are eigenvector rotational transformations that produce the same number of output images, or bands, as input images used. It can be useful to think of these output images as analogous to displaying the differences between the captured images, where ideally at least one output image would show the difference between the parchment and the traces of underwriting. However, the more subtle the differences, the more likely random noise is to overwhelm the results, making many of the output images of no consequence.

Fig. 2 shows a selection of PCA output bands (b–f) along with the natural light image (a). While there is little difference visually between the red and brown ink in the natural light image, output bands (c) and (e) enhance this small difference, as does (d) with respect to the musical notation.

The generated output bands can now be used to create pseudo-colour images, a technique that allows us to combine up to three greyscale images into one colour image. Using this method, it would be possible to display both the underwritten music and the overwriting in the same image and distinguish them from one another, a necessity since it is important for the reader to know where overwriting might be covering underwriting. If the overwriting is removed completely or looks like the parchment background, the reader might not be able to judge the visible outcome properly.[20] In the case of a music manuscript, for example, the note head of a *minima* (♦) might be covered completely by the overwriting, so what would be left would only be a small vertical line. If the overwriting is not recognisable as such, the remains of the *minima* might be mistaken for a rest, falsifying or at least confusing the process of transcribing the respective composition.

Fig. 3: ASL 2211, fol. 30v; initial attempts at creating pseudo-colour images.

[18] The use of raking lights allows the capture of topographical and texture information relating to the folio. However, any buckling and warping can lead to dramatic shadows which overwhelm the subtler signal from the faded ink.

[19] Christens-Barry, Easton, and Knox 2011.

[20] As has been discussed intensively in other projects (Christens-Barry, Easton, and Knox 2011).

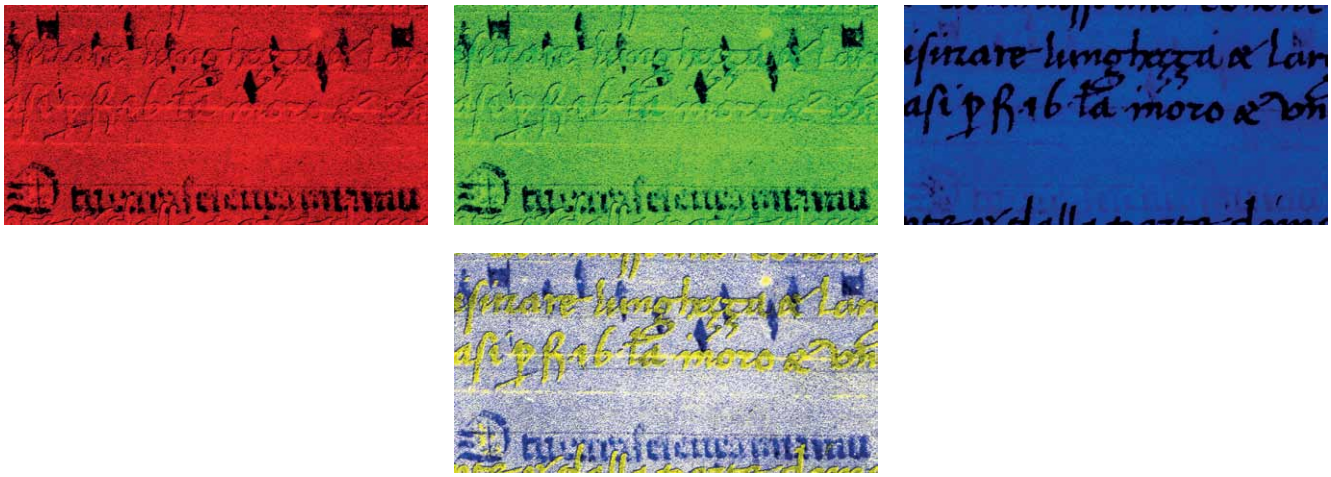


Fig. 4: ASL 2211, fol. 4r, detail; combining red, green and blue channels to create a pseudo-colour image.

Creating pseudo-colour images provides the researcher with the opportunity to combine up to three images (captured or processed) into the red, green and blue channels of a colour image instead of having to choose a single image to represent all the information and details on a page. By trying out several colour combinations, we found this could create results with chaotic and distracting colours that might overwhelm the data being displayed (fig. 3). This led to the search for a simplified and consistent colour scheme that could be applied throughout the manuscript as it was important to have a standardised colour scheme that both enhances the readability of the set of publishable images and preserves some continuity throughout the work. In order to achieve this, a set of criteria was needed to define a suitable image combination.

Since black text on a white background is the most familiar colour combination and the easiest one to read, creating images with the musical underwriting in black (or at least as dark as possible) on a light background became a starting point for developing the colour scheme. The challenge here was that some folios produced many processed output bands suitable for creating pseudo-colours, while others required the use of substandard images for the second and third colour channels. What was required was a method that would work with only a single useful processed image, would be consistent from one folio to the next and would display the

a) b)

Fig. 5: ASL 2211, fol. 52r; a) PCA band 4 and b) PCA band 7.

Fig. 6: ASL 2211, fol. 52r; mean of PCA band 4, PCA band 5 and PCA band 7.

two sets of writing and background in a way that made them clearly distinguishable from one another.

4. Our pseudo-colour method

We found that instead of trying to include as much information (and therefore as many colours) as possible, selecting the best processed output band of a folio and using it twice for both the red and green channels was a better approach. This had the advantage of keeping the colours consistent with dark underwriting, and only one acceptable output image was required. If necessary, the image that displayed the underwriting most clearly would be inverted to keep the musical notation as close to black as possible. The 455 nm blue reflectance image was used in the blue channel of the pseudo-colour. The blue image shows the difference between the parchment and overwriting well, but not the underwriting. It helps to separate the two sets of writing when combined with the red and green channels, as in fig. 4.

Using an image with distinct underwriting twice also means that two of the three channels display the underwriting as dark script, resulting in the musical notation being displayed in dark blue or black. While only using a single image to represent the *scriptio inferior* instead of three is a good solution for well-scraped pages, there is also a chance that by using only one of many good results, a great deal of useful information for other folios will be discarded. In some instances, there are many options from which to choose, each with their own pros and cons. Fig. 5 shows two pseudo-colour images (a and b) based on different processed output bands (PCA band 4 and PCA band 7), for example. In (a), it is especially difficult to distinguish the music on the left-hand side of the first image, which hides information, while image (b) does not provide enough contrast between the over- and underwriting to distinguish them both clearly.

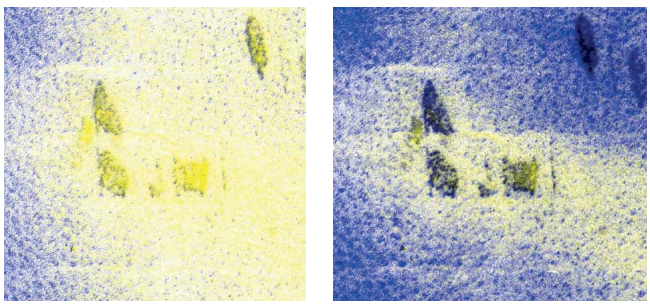


Fig. 7: ASL 2211, fol. 52r; details of figs. 5a and 6.

While neither image is ideal on its own, the mean of these two (or more) images performed in ENVI can preserve the contrast of the first (a) and the left-hand margin details of the second (b), as in fig. 6. This image preserves the contrast between the overwriting and underwriting notation from fig. 5a, but is not as susceptible to the problems of the left-hand margin in fig. 5b, mentioned above. Combining two or more images helps to mitigate the possible disadvantages of only being able to select one image and reduces the influence of very light or very dark areas on the readability of the underwriting.

A rest which appears as a small vertical line, can be used to demonstrate how details may be lost or recovered based on the band – or bands in the case of the mean used in fig. 6 – chosen (see details in fig. 7).

5. Results

It is important to keep in mind that there was a deliberate attempt made to erase the music. Therefore it must be emphasised that while multispectral imaging has produced an improvement on every examined folio compared to the actual state, it is impossible to recover anything that was completely removed due to the palimpsest creation process. In sum, the use of the method described above has been a success for the majority of folios processed to date. Fig. 8 shows the capabilities this method possesses (compared to the present state as shown in fig. 1). However, there are still a number of challenges to overcome such as folio 30v, shown in fig. 9: this folio suffered more from scraping than fol. 4r, for example (figs. 1, 2). Other techniques that are more powerful will therefore be necessary to further enhance and recover the music on the folio. While PCA and ICA are the default processing approaches, this project is by no means limited to those methods, and more techniques will be explored as needed.

The method presented here is designed to display the captured data in such a way that it is easy for scholars to read and study the musical notation. However, this requires an acceptable, legible processed image, which is sometimes a challenge in the case of particularly well-scraped or damaged folios. Uneven scraping, patching and other factors that vary spatially across a single page can contribute to output images that also vary from region to region. For example, it is not unusual for folios to have areas that are especially well scraped. These have different properties than the rest of the page.

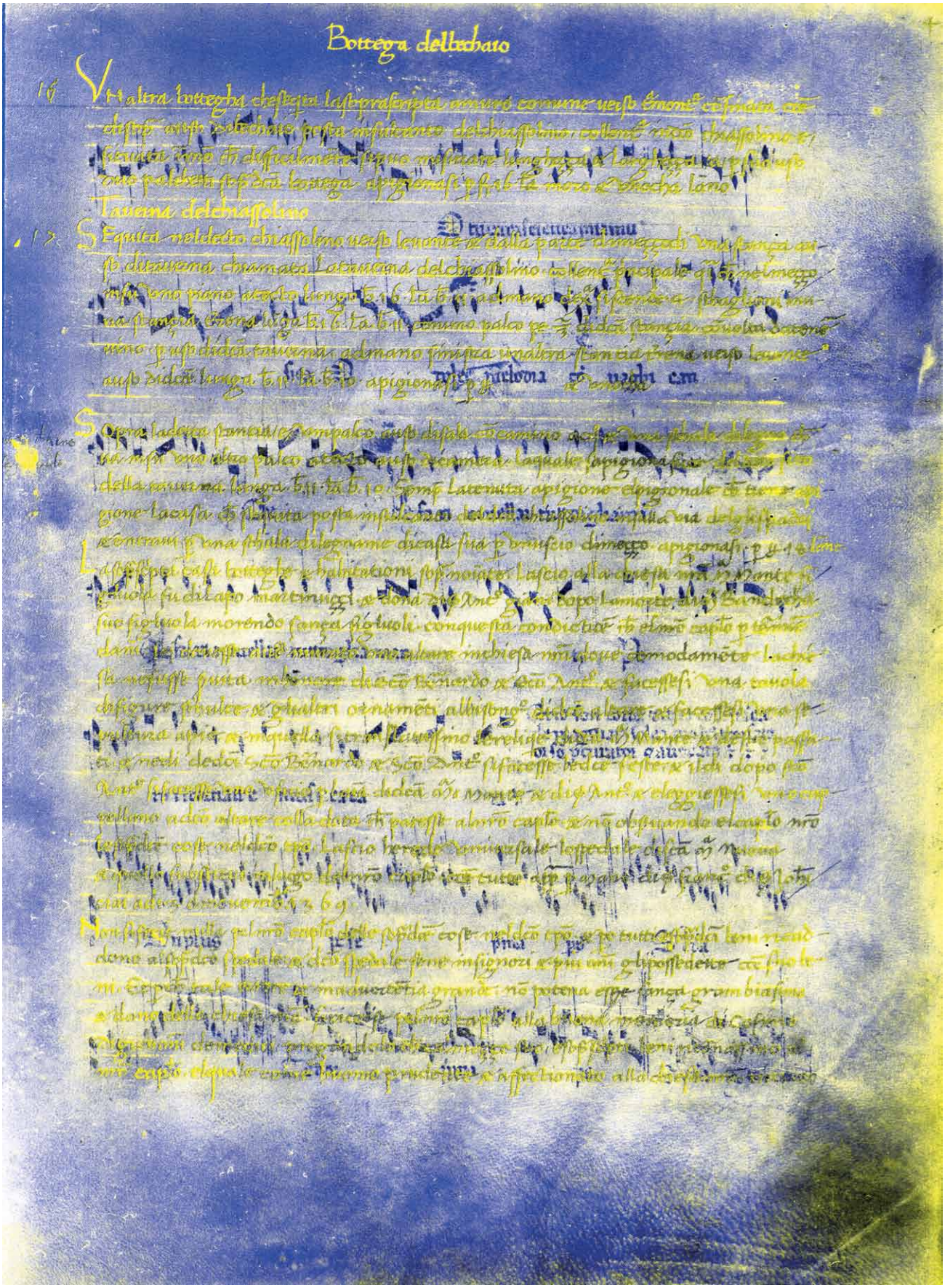


Fig. 8: ASL 2211, fol. 4r; pseudo-colour processed image showing the tenor voice of Giovanni da Firenze's madrigal *O tu cara scienza* and another French piece, *Con plus je se*, (added as a 'space filler') as *scriptio inferior*.

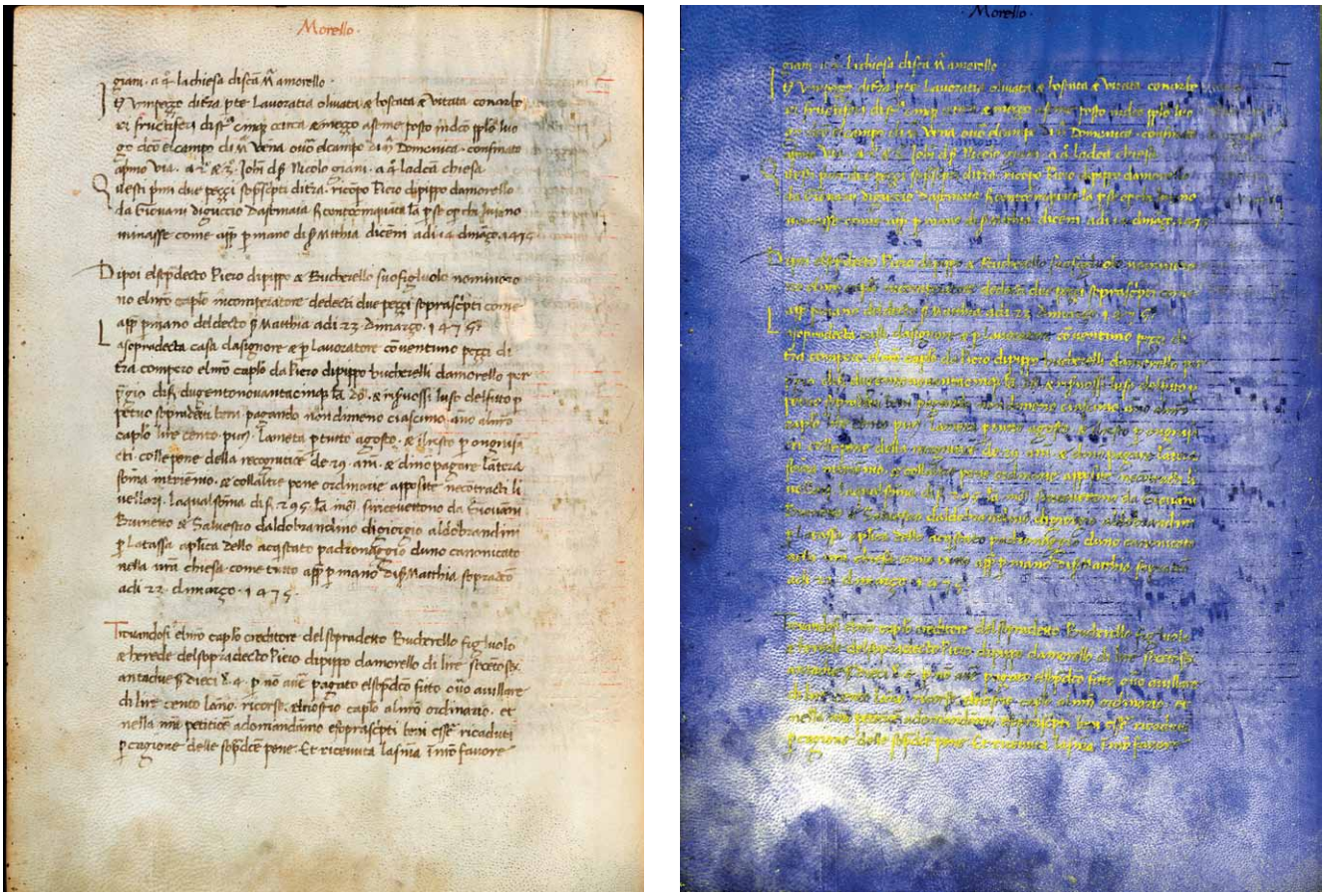


Fig. 9: ASL 2211, fol. 30v; present state of the folio compared to a pseudo-colour image created by the method described above.

The several different colours of ink used in the overwriting affect the colour distribution in each image, expanding this from a two-class to a multi-class problem. There may be four or even five different inks[21] with different spectral properties, all contributing signals that influence the processing.

The criteria required for an acceptable image are focused on the underwriting as the 'interesting part' or signal, and other parts of the original music manuscript will not necessarily be enhanced as well. This is particularly true for red ink that was used for the staves, composer attributions, roman foliation numbers and the groups of red notes found in one composition. This suggests that by using the current method, several images may be needed to fully recover different layers of information within a single folio. Fol. 11v represents a special case since it contains many types of inks in the overwriting, which means that each ink region has different statistical relationships between the overwriting, underwriting and the parchment. Processing the entire page as a single entity can cause the subtle differences the method aims to enhance to be overshadowed by stronger signals. Processing each area as a separate region can consequently lead to better results, as in fig. 10, which shows the beginning of a verse. The group of notes and the first syllable of the text underlay 'Si-' comes out much better when processing a small area compared to processing the whole folio.

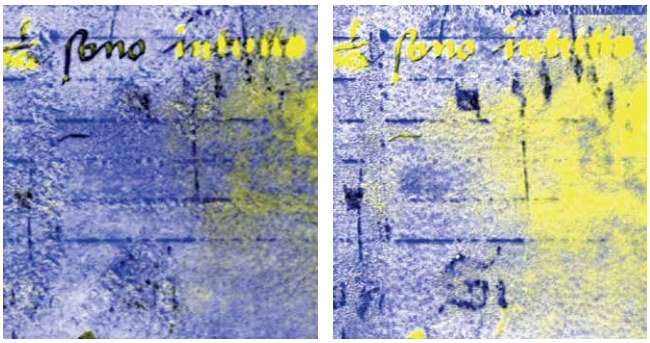


Fig. 10: ASL 2211, fol. 11v; detail, a) processing the whole page, b) processing a particular area.

[21] E. g. red overwriting, which is a little darker than the red underwriting, one or more sets of dark brown or even black overwriting due to the different scribes over the centuries, red underwriting, which is used for composer attributions and roman foliation (only on recto pages) both found in the upper margin, the staves and small groups of red notes in one piece of music, and finally the brown underwriting of the notation and text underlay.

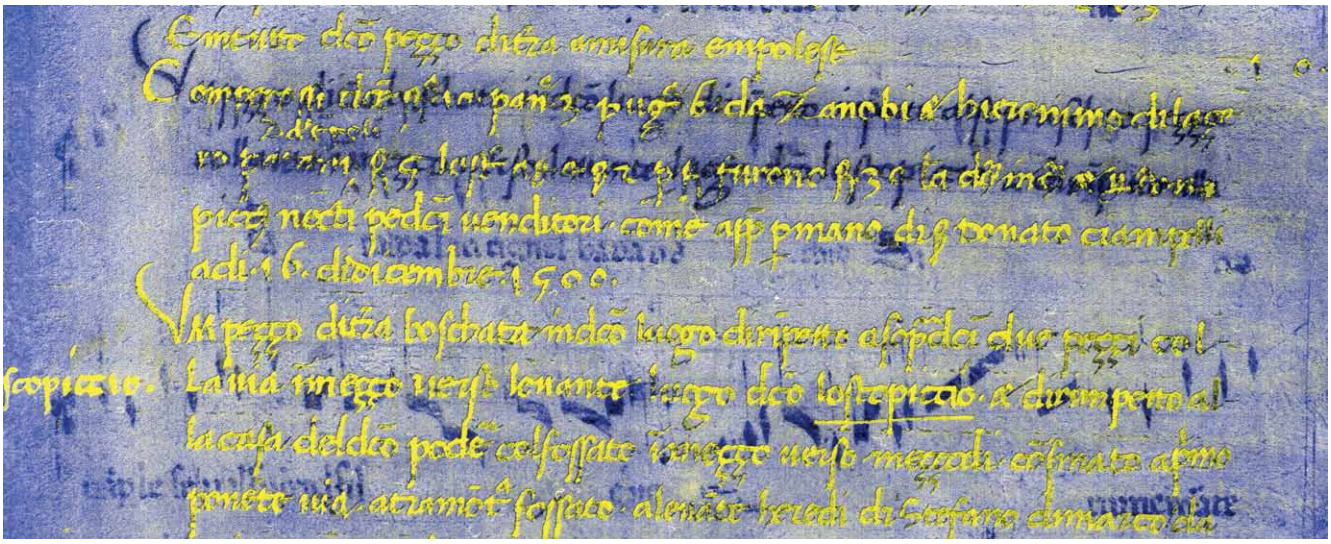


Fig. 11: ASL 2211, fol. 58r, detail.

Despite the challenges mentioned above, the results to date show improved readability of the original music and the text underlay in the *scriptio inferior*. In the case of fol. 4r (fig. 8), it is now possible to decipher the incipit of the added French composition *Con plus je se*.

Due to the fact that it is not necessary to decide which script belongs to the underwriting and to the overwriting in the first steps, unexpected outcomes can arise, as in the case of fol. 58r. In general, the separation between the music in the *scriptio inferior* and the overwriting is clear (fig. 11).

Surprisingly, another layer of erased text showed up in dark blue from part of the *Campione dei Beni* (in fig. 11). This provides an insight into the making of the manuscript. Apparently, after the scribe had copied the two lines beginning 'Un pezzo di terra boschata in dicto luogo […]', he realised that he had left out a whole paragraph from his exemplar, so he erased the two lines and carried on with his work.

One significant further step regarding the compositions transmitted as *unica* in this source features a specific piece by Giovanni Mazzuoli. The organist and composer has been described as an 'enigmatic figure'[22] due to the fact that a section was laid out for his music in the lavishly decorated Squarcialupi codex[23] made in Florence around 1415. However, his compositions were never written down in the codex (fig. 12). This is also the case for Paolo da Firenze, another Florentine Trecento composer whose compositions are known from other sources. Thanks to the specific decoration system of the Squarcialupi codex, it was possible to identify the madrigal *Girand' un bel falcon* as the opening piece of Paolo's section.[24] What proved helpful in the identification process was the historiated initial and the *bas-de-page* miniature, which always refers to the text of the composition to be written above it.

With the help of the San Lorenzo Palimpsest, it is now possible to identify the opening piece originally intended for the blank section of Giovanni Mazzuoli in the Squarcialupi codex. The decoration system on fol. 195v in the Squarcialupi codex includes a historiated initial – the letter 'C' – and a *bas-de-page* miniature with a dancing scene (fig. 12).

On the basis of previously deciphered text fragments, Janke[25] has formulated a hypothesis that this decoration might have been intended for the madrigal *Chome servi a Signor*. Thanks to multispectral imaging, this hypothesis has been verified since it is now possible to transcribe the entire text of the madrigal (see below). An important link between the decoration and the text of the madrigal is the description of women dancing a round dance with one dressed in white in the centre (for the text residuum, see fig. 13).

[22] D'Accone 1968, 23.

[23] Florence, Biblioteca Medicea Laurenziana, Mediceo Palatino 87. The images can be found in the facsimile in Gallo 1992.

[24] Günther, Nádas, and Stinson 1987, 204; Nádas 1989.

[25] Janke 2014, 248.

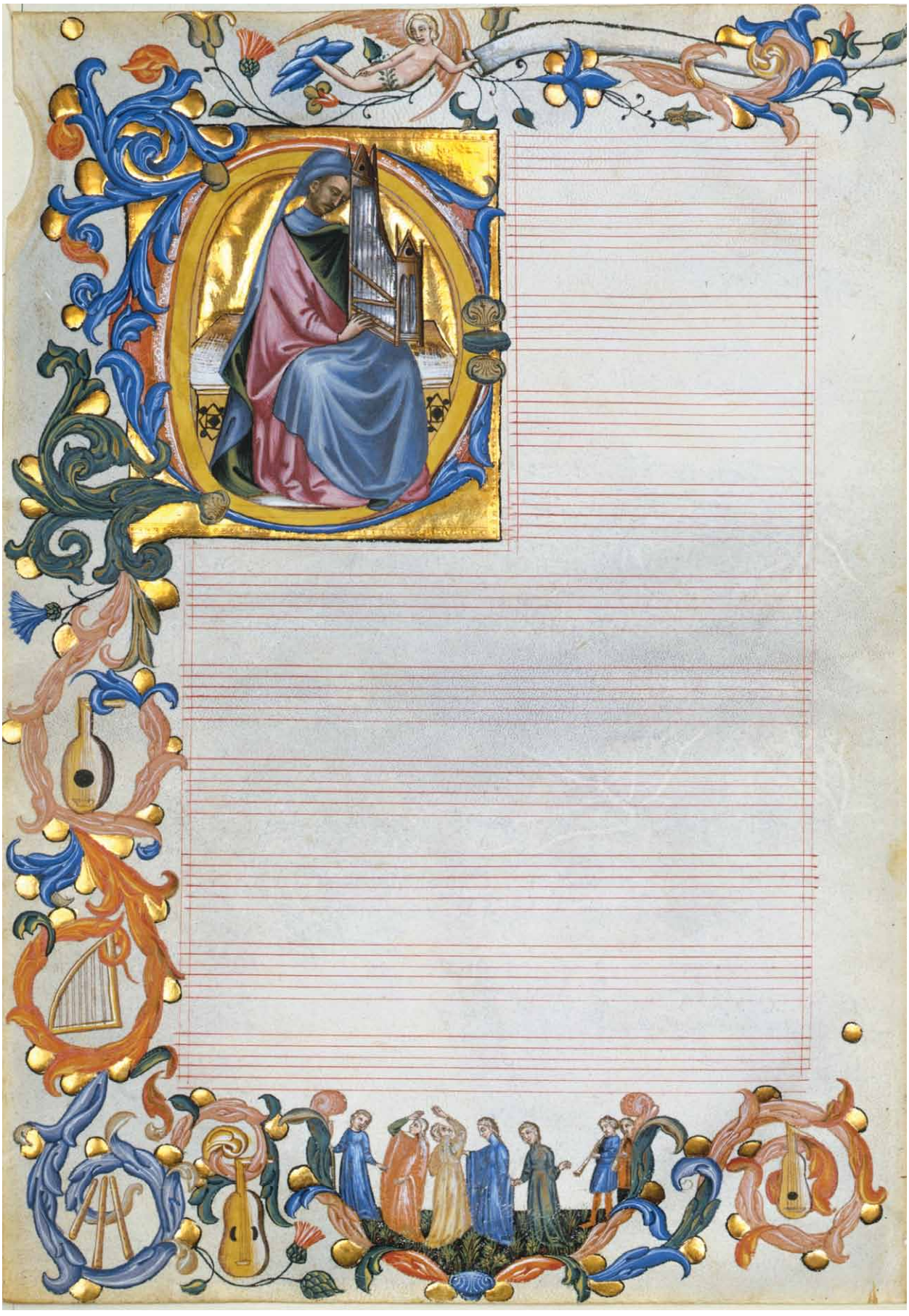


Fig. 12: Florence, Biblioteca Medicea Laurenziana, Mediceo Palatino 87, fol. 195v.

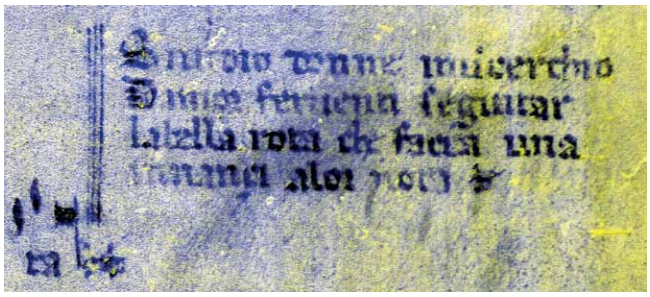


Fig. 13: ASL 2211, fol. 24v, text residuum.

Chome s*er*vi a signo*r* giust'e umile
attenti a 'ntende*r* lo suo disidero
e cho*mpre*sa suo voglia e ognu*n* legero

Sì vid'io donne in u*n* cerchio
D'Amor ferventi seguitar la bella rota
che facea una innançi a lor nota

In bia*n*cha vesta do*n*çella vestita
ch'è dat'a Lore*n*ça guido mie vita.[26]

6. Outlook

As the project progresses, it is hoped that more of the illegible pieces will be identified, and the possibility of finding more *unica* means that the San Lorenzo Palimpsest will prove to be a treasure trove to musicologists and other scholars. The method for recovering the music is still under development, and more techniques will be tried in order to recover the more difficult pages.

By publishing the images in a printed format[27] (not discounting the possibility of a digital database), intensive research can finally start and can begin a new chapter in understanding the music and the manuscript's history and significance.

[26] Authors' translation: 'Just as servants pay heed to the just and the humble lord to understand his desire and having understood his will, everyone is light-hearted, / Thus I saw women in a circle, glowing for love, they follow the lovely round dance that highlights one of them / The maiden dressed in a white dress, which was given to Lorença, guides my life.'

[27] See fn. *.

REFERENCES

Christens-Barry, Williams A., Easton Jr., Roger L., and Knox, Keith T. (2011), 'Imaging and Image-Processing Techniques', in Reviel Netz, William Noel, Natalie Tchernetska, and Nigel Wilson (eds.), *The Archimedes Palimpsest. Catalogue and Commentary* (Cambridge University Press), 175–197.

Craig-McFeely, Julia, and Lock, Alan (2006), *Digital Restoration Workbook* (Oxford: Alden Press), (http://www.diamm.ac.uk/publications.html).

—— (2012), 'From Perfect to Preposterous: How Digital Restoration Can Both Help and Hinder Our Reading of Damaged Sources', in Lynn Ransom, and Emma Dillon (eds.), *Cantus Scriptus. Technologies of Medieval Songs* (Piscataway, NJ: Gorgias Press), 125–141.

D'Accone, Frank (1968), 'Giovanni Mazzuoli, a late representative of the Italian Ars nova', in Alberto Gallo (ed.), *L'Ars nova italiana del Trecento*, vol. 2, *Convegni di studio 1961–1967*, 23–38.

—— (1984), 'Una nuova fonte dell'ars nova italiana: il codice San Lorenzo, 2211', *Studi Musicali*, 13: 3–32.

Gallo, Alberto (ed.) (1992), Il codice Squarcialupi. Ms. Mediceo Palatino 87 Biblioteca Medicea Laurenziana (Florence: Giunti Barbera, and Libreria Musicale Italiana).

Gehring, Julia (2012), *Die Überlieferung der Kompositionen Francesco Landinis in Musikhandschriften des späten 14. und frühen 15. Jahrhunderts* (Hildesheim et al.: Olms; Musica Mensurabilis, 5).

Gozzi, Marco (2011), 'Trecento', in Mark Everist (ed.), *The Cambridge Companion to Medieval Music* (Cambridge University Press), 136–160.

Günther, Ursula, and Nádas, John, and Stinson, John A. (1987), 'Magister Dominus Paulus Abbas de Florentia. New Documentary Evidence', in *Musica Disciplina*, 41: 203–246.

Huck, Oliver, and Dieckmann, Sandra (eds.) (2007), *Die mehrfachen überlieferten Kompositionen des frühen Trecento. Übertragungen, Texte, Kommentare* (Hildesheim et al.: Olms; Musica Mensurabilis, 2).

Janke, Andreas (2013), '"Hoc enim in plano cantu raro videtur contingere." Modus und Mehrstimmigkeit im späten Trecento', in Jochen Brieger (ed.), *Das modale System im Spannungsfeld zwischen Theorie und Praxis* (Frankfurt am Main: Peter Lang; Hamburger Jahrbuch für Musikwissenschaft, 29), 55–67.

—— (2014), 'Giovanni e Piero Mazzuoli. Due compositori del tardo Trecento', in Marco Gozzi, Agostino Zino, and Francesco Zimei (eds.), *Beyond 50 Years of Ars Nova Studies at Certaldo 1959–2009, Atti del Convegno internazionale di Studi (Certaldo, Palazzo Pretorio, 12–14 giugno 2009)* (Lucca: Libreria Musicale Italiana; L'Ars Nova italiana del Trecento, 8), 241–253.

Meconi, Honey (2011), 'Shedding New Light (Literally) on the Rochester Fascicle: A Preliminary Report', in Fabrice Fitch, and Jacobijn Kiel (eds.), *Essays on Renaissance Music in Honour of David Fallows: Bon jour, bon mois, et bonne estrenne* (Woodbridge: Boydell and Brewer; Studies in Medieval and Renaissance Music, 11), 52–59.

Memelsdorff, Pedro (2004), 'New Music in the Codex Faenza 117', *Plainsong & Medieval Music*, 13: 141–161.

Moran, Neil K. (1985), 'A List of Greek Music Palimpsests', *Acta Musicologica,* 57: 50–72.

Nádas, John L. (1989), 'The songs of Paolo Tenorista. The Manuscript Tradition', in Fabrizio Della Seta, and Franco Piperno (eds.), *In cantu et in sermone. For Nino Pirrotta on his 80th Birthday* (Florence: Olschki and University of W. Australia Press), 41–64.

——, and Ziino, Agostino (eds.) (1990), *The Lucca Codex: Codice Mancini; Lucca, Archivio di Stato, MS 184; Perugia, Biblioteca Comunale 'Augusta', MS 3065. Introductory study and facsimile* (Lucca: Libreria Musicale Italiana).

—— (1992), 'Manuscript San Lorenzo 2211: Some further Observations', in Giulio Cattin, and Patrizia Dalla Vecchia (eds.), *L'ars nova italiana del trecento, 6, Atti del congresso internazionale Certaldo 1984* (Certaldo: Ed. Polis), 145–168.

Schmidt, Thomas, and Vorholt, Hanna (2009), 'Mise-en-page in Choirbooks, ca. 1450–1550', *Gazette du livre médiéval*, 55: 31–42.

Wathey, Andrew, and Bent, Margaret, and Craig-McFeely, Julia (2001), 'The Art of Virtual Restoration: Creating the Digital Image Archive of Medieval Music (DIAMM)', in Hewlett, Walter B., and Selfridge-Field, Eleanor (eds.), *The Virtual Score. Representation, Retrieval, Restoration* (Cambridge, Mass.: The MIT Press, and Stanford, University of California: Center for Computer Assisted Research in the Humanities – CCARH; Computing in Musicology, 12), 227–240.

PICTURE CREDITS

Article

Combining Codicology and X-Ray Spectrometry to Unveil the History of Production of Codex germanicus 6 (Staats- und Universitätsbibliothek Hamburg)

Ira Rabin, Oliver Hahn, and Mirjam Geissbühler | Berlin – Hamburg – Bern

1. Introduction

The investigation of physical properties and chemical composition generates important data for answering cultural-history questions that cannot be solved by historical and philological methods alone. In its individual materiality, each manuscript is the result of a wide variety of influences. The 'life' of a manuscript starts with its production, followed by the use and storage of the manuscript, and is finally characterized by its treatment during restoration. Some of these characteristics are still in existence and may provide insights into the production process and history of a manuscript.

Codex germanicus 6, which consists of a compilation of twelve different texts, is an excellent example of a manuscript with a complex history. It is a plain, 614-page manuscript without illuminations and was created around 1450. Most of the twelve different texts are composed in Middle High German. The entire manuscript was written and rubricated by a scribe who called himself Jordan – of whom little else is known – and was intended for his personal use, as he conveyed in two colophons on pages 365 and 560.

The combination of classic codicology and scientific analysis, i.e. advanced codicology, should assist in clarifying the chronology of the production process.

The most important prerequisite for investigating historical objects is the use of techniques that are non-destructive or only require minimal sampling. The unchanged sample should preferably still be available for further study after it has been analysed. X-ray fluorescence analysis is one of the most suitable methods for obtaining qualitative and semi-quantitative information on a great diversity of materials and is a convenient technique for the investigation of inorganic compounds. In this article, we present preliminary results from XRF examinations of the red inks at relevant points in the manuscript, especially at points where one text ends and another begins.

2. Codicological analysis

According to codicological research, the sequence of the texts in Cod. germ. 6 does not correspond to the order in which they were penned. Table 1 shows where the individual texts are positioned within the codex (under 'Text index') and the order in which they were transcribed (under 'Evidence about the order of transcription of the texts'). The last column summarizes the questions we have attempted to answer in this paper.

The codex begins with two *Meisterlieder*: *König Artus' Horn* and *Luneten Mantel.* They were obviously added after the codex was bound since a sheet has been inserted (page 5/6) so that they could be placed before the beginning of the *Parzival* text on page 8. The index, written by the same scribe, has also been placed before the texts. It only mentions texts 3, 5, 6, 8, 9, 10 and 11. However, text 7 appears seamlessly between texts 6 and 8 in the middle of quire 24, indicating that Jordan most likely forgot to note it in the index. This is plausible since texts 6 and 7 have the same title in the codex (*Von dem Soldane)*. It is noteworthy that texts 4 and 12 are very short and also could have been forgotten by Jordan. Alternatively, they could have been penned after the index was composed.

Text 3, entitled *Parzival,* is dated to 2 February 1451 in a colophon which concludes the text of the romance. *Artusnotiz* – text 4 – appears on the same page and could have been added later since it is extremely short. The last page of the quire is blank. The next quire begins with text 5, the Arthurian romance *Wigalois*, which also has a colophon dating completion of the text to 11 November 1451. It is

Table 1: Structure of Codex germanicus 6

Quire	Pages*	Title	Text index	Evidence about the order of transcription of the texts	To clarify
1	2a–4a	König Artus' Horn[1]	1	Before: 2 , After: 3, 5, 6, 7, 8, 9, 10, 11	Before or after 4/12?
1	4a–6b	Luneten Mantel [2]	2	Before: ?, After: 1, 3, 5, 6, 7, 8, 9, 10, 11	Before or after 4/12?
1-15	8a–365a	Parzival[3]	3	Before: 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12	
15	365a	Artusnotiz[4]	4	Before: ?, After: 3	Added later?
16-23	367a–560a	Wigalois[5]	5	Before: 1, 2, 6, 7, 8, 9, 10, 11, 12 , After: 3	Before or after 4?
23-24	560a–567a	Sultansbrief Abul Nasr[6]	6	Before: 1, 2, 7, 8, 9 , After: 3, 5, 10	Before or after 4/11/12?
24	567a–569a	Sultansbrief Almansor[7]	7	Before: 1, 2, 8, 9, After: 3, 5, 6, 10	Before or after 4/11/12?
24	569a–575b	Der König im Bad[8]	8	Before: 1, 2, 9, After: 3, 5, 6, 7, 10	Before or after 4/11/12?
24	576–587a	Friedrich[9]	9	Before: 1, 2, After: 3, 5, 6, 7, 8, 10	Before or after 4/11/12?
25	589a–610b	Jeanne d'Arc[10]	10	Before: 1, 2, 6, 7, 8, 9, 11, 12, After: 3, 5	Before or after 4?
25	611a–612b	Lüttich[11]	11	Before: 1, 2, 12, After: 3, 5, 10	Added later?
25	612b	Notabile[12]	12	Before: ?, After: 3, 5, 10, 11	Added later?

Quire formula: $(VI+1)^{13} + (VI)^{169} + (VII)^{183} + (VI)^{267} + (VII)^{281} + (VI)^{293} + (VII)^{307}$.

* Letters 'a' and 'b' correspond to the left and right columns.

[1] This *Meisterlied* is about a test of fidelity at the court of King Arthur and originates from the end of the 14th century or the first half of the fifteenth century; cf. Schanze 1985, 69–70.

[2] The second *Meisterlied* deals with another test of fidelity at the court of King Arthur and was presumably written in the first half of the 15th century; cf. Schanze 1985, 1068–1069.

[3] An Arthurian romance by Wolfram von Eschenbach written between 1200 and 1210; cf. Bumke 2004, 19–21.

[4] A short note containing biographical information about King Arthur.

[5] An Arthurian romance by Wirnt von Grafenberg written between 1210 and 1220; cf. Wennerhold 2005, 80.

[6] Letter from the Egyptian sultan Abul Nasr to Antonio Fluvian de Rivière from 1426; cf. Putzo 2002, 64.

[7] A fictitious letter from the Babylonian sultan Almansor to the Roman Pope, the emperor and all kings; cf. Putzo 2002, 64.

[8] Narration in couplets about a king who loses everything, but returns to the throne after experiencing catharsis. It originates from the second half of the 13th century.

[9] This text lists the order of entry at the coronation of Emperor Friedrich III in Rome in 1452; cf. Putzo 2002, 64. This is the most recent text in Cod. germ. 6.

[10] These 22 pages contain diverse documents (such as letters) on the life of Jeanne d'Arc, who lived from 1412–1431.

[11] This text presents the articles of the peace treaty between the Bishop of Lüttich (Johannes VIII von Heinsberg) and the town of Lüttich in 1408.

[12] This note tells of a woman who gave birth to an animal form in Strasbourg in 1412.

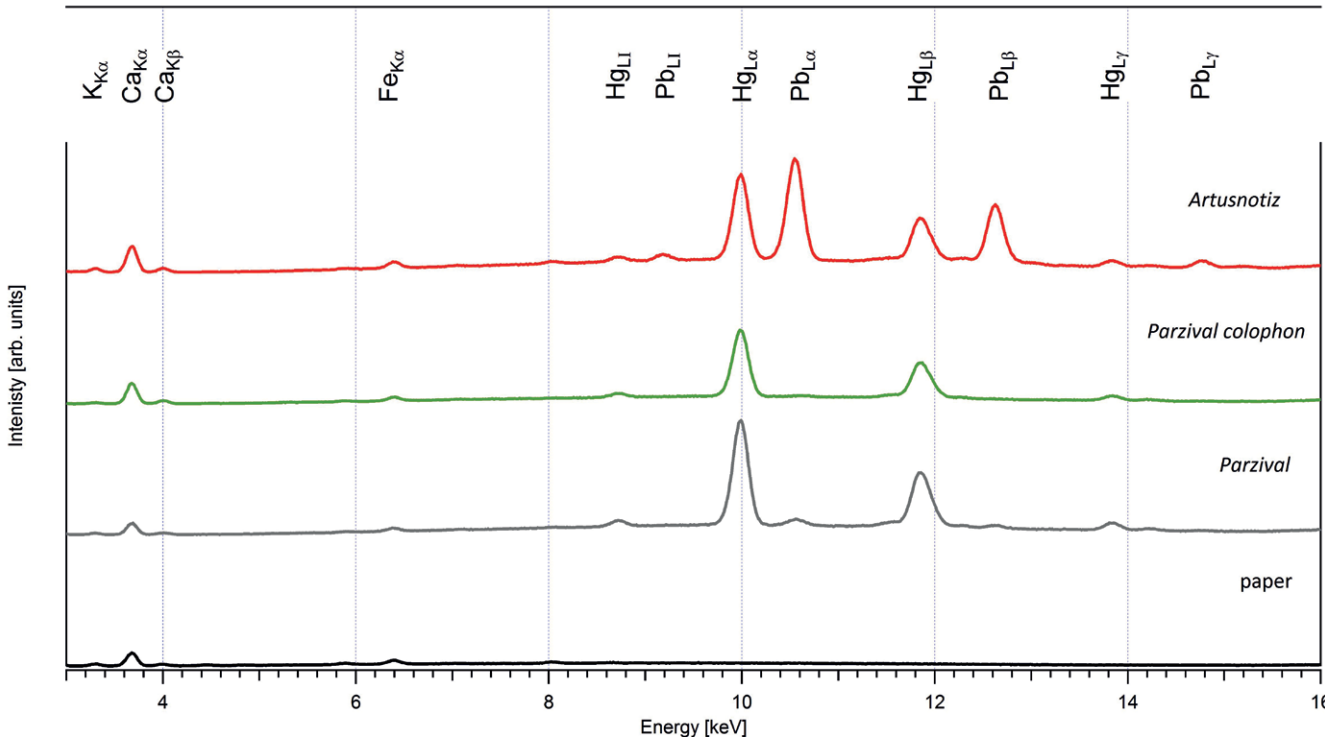


Fig. 1: Excerpt from the XRF spectra collected from the paper and selected red inks.

thus certain that *Parzival* was transcribed before *Wigalois*. There is no indication that the remaining texts in the two last quires of the codex were penned before the two Arthurian romances, *Parzival* and *Wigalois*. It is therefore most likely that the two longer texts in the codex were the first to be transcribed.

There are, however, two factors indicating that quire 25 was transcribed before quire 24. Firstly, the 25th quire is very likely to have initially been a senion, but appears in the bound manuscript as a 7-bifolium quire: it is enclosed by a double sheet (p. 587/588 and 613/614) so that the last text of quire 24 could end on page 587. Page 588 is blank, and the first text of the 25th quire begins on page 589. Secondly, the 24th quire and the double sheet added around quire 25 show watermarks that differ from those in the other quires in Cod. germ. 6.

This suggests that the first text of the 25th quire, text 10, was initially placed after *Wigalois,* which is text 5. This text nearly completes quire 25, leaving only two pages. On these pages we find two short texts (11 and 12) that could easily have been added later. Furthermore, we can conclude that text 11 was written before texts 1 and 2 were added since it is listed in the index of the codex, whereas texts 1 and 2 are not.

Text 6, the first letter from the sultan, comes after the *Wigalois* text, starting in quire 23 and ending in quire 24. If quire 24 is indeed the last quire of the codex, it is likely that all the texts in this quire were penned in consecutive order.

Codicological analysis employed in this research led to an almost complete reconstruction of the transcription of Cod. germ. 6. The remaining questions that could not be resolved by classical codicology alone concern short texts (4, 11 and 12), which could have been added at a later stage. We recently demonstrated that expanding codicology to include physical and chemical analysis of writing materials offers new possibilities for studying the history of manuscripts. Expanded codicology includes information ranging from simple recognition of the ink typology by visible and infrared reflectography to determination of the chemical composition using complex analysis methods.[13] Ink composition in particular is used to help reconstruct the chronology of the texts. The method is based on the observation that a change in the scribe's hand is often accompanied by a change in the ink composition. Moreover, handmade inks in the Middle Ages were never completely reproducible, with the result that different batches of ink prepared with the same recipe display sufficient differences to be detected by chemical analysis.

[13] Cf. Rabin et al. 2012.

3. X-ray fluorescence analysis

To take X-ray fluorescence measurements, we used a commercial, transportable (though not portable) ARTAX micro-XRF spectrometer (made by Bruker Nano GmbH) specially designed for archeometric studies in situ.[14] It consists of a low-power, air-cooled X-ray tube, polycapillary optics resulting in a beam spot of 70 µm in diameter, an electro-thermally cooled Xflash detector and a CCD camera for sample positioning. All measurements were taken using a 30 W low-power Mo tube operated at 50 kV and 600 µA and with an acquisition time of 20 s (live time). Peak fitting and semi-quantitative data evaluation were conducted using Bruker's SPEKTRA software.

Fig. 1 presents an example of the spectra collected in this work. The elements K, Ca and Fe are present in the paper and constitute a constant background. Hg and Pb correspond to the red inks. We observe practically only Hg in the red inks used for rubrication of *Parzival* and its colophon, whereas the inks used in *Artusnotiz* have a large amount of Pb.

4. Results

With the aid of XRF, six different kinds of red ink have been distinguished to date in Cod. germ. 6. The majority of the red inks analysed show cinnabar with lead contamination below 10% – the red ink in one text had more lead (Pb) than mercury (Hg), however. In other words, this ink has a different chemical composition than the rest of the red inks and is most probably a mixture of minium and cinnabar. This particular red ink has been found in text 4, *Artusnotiz* – one of the three texts in Cod. germ. 6 that could conceivably have been added at a later date. Ink analysis suggests that text 4 was not written at the same time as the colophon of *Parzival*. On the basis of the codicological study, we believe that this was, in fact, the last text to be penned. To validate our thesis, it is necessary to consider the positioning of *Artusnotiz* in the manuscript.

Table 2: Summary of the red inks tested in our study.

Text	Title	Use in the text	Pb/Hg	Ink no.
2	Luneten Mantel	Rubrication	0.07±0.007	1
3	Parzival	Rubrication	0.06±0.006	1
3	Parzival	Colophon	0.01±0.001	2
4	Artusnotiz	Rubrication	1.24±0.12	3
5	Wigalois	Rubrication	0.3±0.03	4
5	Wigalois	Colophon	0.29±0.03	4
6	Sultansbrief Abul Nasr	Heading	0.09±0.01	5
10	Jeanne d'Arc	Last passage	0.07±0.007	1
11	Lüttich	Rubrum	0.032±0.003	4
11	Lüttich	Rubrication	0.1±0.01	5
12	Notabile	Heading	0.045±0.005	6
12	Notabile	Rubrication	0.07±0.007	1

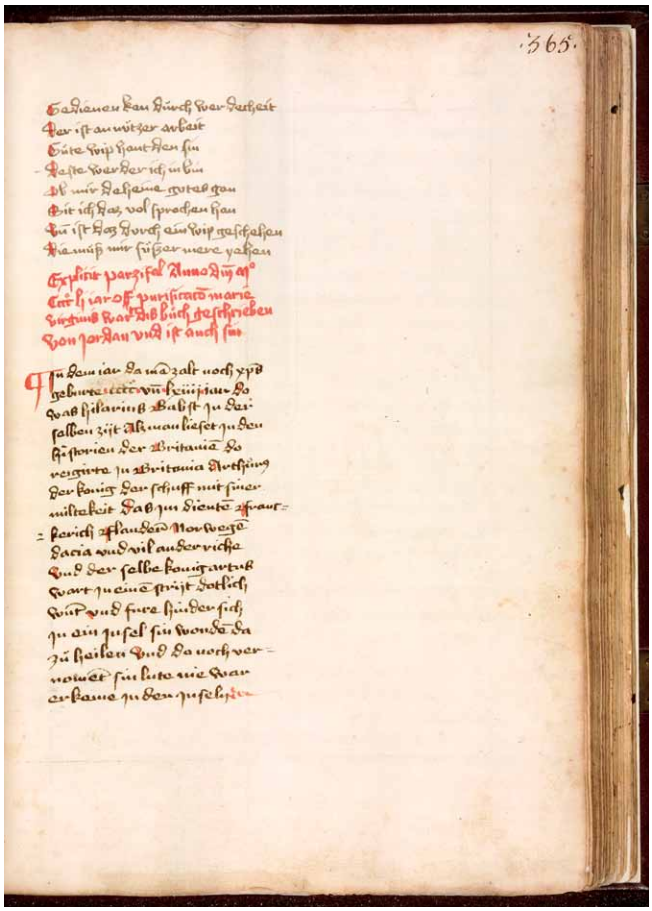


Fig: 2: Page 365 of Cod. germ. 6.

[14] Cf. Bronk et al. 2011.

Fig. 2 shows page 365 of Cod. germ. 6 featuring the text of *Artusnotiz*.[15] We are concerned here with the 17 lines at the bottom of the page, which are separated by an alinea. The first eight lines are the last verses of *Parzival*, and the following four lines constitute the colophon written in red ink. Since *Artusnotiz* is positioned directly after the colophon, we can be certain that it was written after *Parzival*. The measurements for the black ink also indicate that the verses of *Parzival* on this page and the text of *Artusnotiz* were written in different inks. Furthermore, the red ink used for respective rubrication of the two texts is not identical. This makes it extremely probable that *Artusnotiz* was transcribed after *Parzival* had been written and rubricated, otherwise both texts would most likely have been rubricated with the same red ink.

Based on the codicological examination, there are two other texts besides *Artusnotiz* which could be the last ones to have been transcribed, namely text 2 and text 12 – the *Meisterlied Luneten Mantel* and *Notabile.* Interestingly, the measurements show that these two texts were rubricated with the same red ink. Exactly this sort of ink was found in two other passages in the codex as well: in the last paragraph of text 9 and in the rubrication of *Parzival.* As mentioned above, *Artusnotiz* was transcribed after *Parzival* had been written and rubricated. Since texts 2 and 12, are rubricated with the same ink as *Parzival*, we can conclude that *Artusnotiz* was the last text added to Codex germ. 6.

5. What new insights can be derived from this conclusion?

- The results of the examination using XRF spectrometry show that *Artusnotiz* was not transcribed immediately after *Parzival.* It can be assumed on this basis that *Artusnotiz* was not combined with *Parzival* in the original manuscript given to Jordan to copy draft, but that he deliberately decided to add the short text.
- Since the composition of the red ink used in *Artusnotiz* differs considerably from all the other red inks detected in Cod. germ. 6, it is quite possible that the short, 17-line text was added quite some time after completion of the penultimate text. It is not even possible to rule out that Jordan added *Artusnotiz* after the manuscript had been bound.
- When a text as short as *Artusnotiz* is added to a manuscript, it is possible that it simply served as a filler. It is not very likely that this applies to *Artusnotiz* in Cod. germ. 6, however, since the following page, 366 (which is the last one of the 15th quire), is blank. Jordan could have written *Artusnotiz* or another short text on this blank page, but he chose not to. He most likely wanted to position *Artusnotiz* in relation to the text of *Parzival.* What could have been the purpose of placing the texts in this way?

Artusnotiz tells us that in 464, King Arthur ruled *Franckerich* (France), *Flandern* (Flanders), *Norwegen* (Norway), *Dacia* and lots of other kingdoms. It subsequently gives an account of how the king was mortally wounded and went to an island to recover and of how his people never knew whether or not he would return. This is an abridged passage from the *Chronicon pontificum et imperatorum* by Martin von Troppau, a chronicle with a rich tradition which synoptically renders the reigns of popes and emperors.[16] Jordan quite obviously tried to historicize the text of *Parzival* by incorporating *Artusnotiz* immediately after *Parzival*. The very different data collected from the red ink measured in *Artusnotiz* makes it likely that it was added after the manuscript had been bound. The fact that Jordan added something (that he probably discovered subsequently) to a manuscript that was actually already finished shows that historicizing *Parzival* was extremely important to him.

This example shows us that the first XRF study of Cod. germ. 6 was able to help answer a number of codicological questions. However, the new findings which indicate that there were seven different red inks used in the manuscript without a proper system also raise a few additional questions, especially with regard to the date of transcription of texts 11 and 12. Further research will help us gain greater insight into the complicated history behind the creation of this codex.

[15] It is possible to view a scan of Cod. germ. 6 online. For more information, see below under 'References'.

[16] Von den Brincken 1987, 161–162.

ACKNOWLEDGEMENTS

We gratefully acknowledge the funding support provided by the German Research Foundation (DFG) for the SFB 950 'Manuscriptkulturen in Asien, Afrika und Europa' / Centre for the Study of Manuscript Cultures (CSMC), University of Hamburg.

PICTURE CREDITS

Fig. 2: © Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky.

REFERENCES

Von den Brinken, A.-D. (1987),'Martin von Troppau', in Kurt Ruh et al. (eds), *Die deutsche Literatur des Mittelalters: Verfasserlexikon,* vol. 6 (Berlin: de Gruyter), 158–166.

Bronk, H., Röhrs, S., Bjeoumikhov, A., Langhoff, N., Schmalz, J., Wedell, R., Gorny, H.-E., Herold, A., Waldschläger, U. (2001), 'ArtTAX®: A new mobile spectrometer for energy dispersive micro X-ray fluorescence spectrometry on art and archaeological objects', *Fresenius J. Anal. Chem.,* 371: 307–316.

Bumke, J. (2004), *Wolfram von Eschenbach* (Stuttgart: Metzler).

Putzo, C. (2002), 'Cod. germ. 6', in E. Horváth and H.-W. Stork (eds.), *Von Rittern, Bürgern und von Gottes Wort. Volkssprachige Literatur in Handschriften und Drucken aus dem Besitz der Staats- und Universitätsbibliothek Hamburg* (Kiel: Ludwig), 64–67, and 136–141.

Rabin, I., Schütz, R., Kohl, A., Wolff, T., Tagle, R., Pentzien, S., Hahn, O., Emmel, S. (2012), 'Identification and classification of historical writing inks in spectroscopy', *Comparative Oriental Manuscript Studies Newsletter,* 3: 26–30.

Schanze, F. (1985): 'König Artus' Horn', in Kurt Ruh et al. (eds.), *Die deutsche Literatur des Mittelalters: Verfasserlexikon*, vol. 5 (Berlin: de Gruyter), 69–70.

—— (1985), 'Luneten Mantel', in Kurt Ruh et al. (eds.), *Die deutsche Literatur des Mittelalters: Verfasserlexikon,* vol. 5 (Berlin: de Gruyter), 1068–1069.

Wennerhold, M. (2005), *Späte mittelhochdeutsche Artusromane. 'Lanzelet', 'Wigalois', 'Daniel von dem Blühenden Tal', 'Diu Crône'. Bilanz der Forschung 1960–2000* (Würzburg: Königshausen & Neumann).

Permanent link to the scan of Codex germanicus 6: http://resolver.sub.uni-hamburg.de/goobi/HANSh496 or: www.sub.uni-hamburg.de > Digitalisierte Bestände > Abendländische Handschriften > Sammelhandschrift: Meisterlieder – Wolfram von Eschenbach 'Parzival' – Wirnt von Grafenberg 'Wigalois' – Chronikauszüge – Kleinepik

Article

A Modular Workbench for Manuscript Analysis

Arved Solth, Rainer Herzog, and Bernd Neumann | Hamburg

1. Introduction

This article presents ongoing work towards developing a modular workbench for the visual analysis of manuscripts. The workbench is known as AMAP (*Advanced Portal for Manuscript Analysis*). We will briefly describe existing tools for manuscript analysis, then introduce the system structure of our workbench and present three modules in detail. Unlike most other systems, AMAP will provide a broad repertoire of interactive tools enabling manuscript researchers to determine palaeographic features in large datasets. Further tools will support advanced tasks such as layout analysis, word spotting and writer identification.

2. Related work

Several projects dealing with specific tasks in manuscript research have been commenced in recent years. The functionalities developed in these projects range from tools that help scholars to annotate and transcribe digital images of handwritten documents to systems for writer identification and verification. Although many of the systems mentioned in this section are still under development, their performance in the respective fields of application already looks promising. The majority of these systems are more or less one-off custom solutions for specific problems, however. With AMAP, our goal is to develop an integrated modular system that provides manuscript researchers with a broad range of functionalities and can be accessed through an internet portal. At the time of writing, the workbench included basic tools for handling manuscript images, defining and annotating details, obtaining geometric measurements and visualising statistics. As examples of more advanced tools, the system contains modules for layout analysis, grapheme retrieval and determining character features. Additional modules will be added based on the requirements arising from the ongoing research work at the Centre for the Study of Manuscript Cultures (CSMC), Hamburg. In the following, we will discuss the existing systems that are most relevant for our approach.

Diptychon[1] is a web-based tool for the transcription of medieval handwriting, which is currently under development by the Artificial Intelligence Research Group led by Björn Gottfried at the University of Bremen, Germany. The tool guides the user through a semi-automatic transcription process. In a first step, *Diptychon* provides the manuscript researcher with an over-segmentation of the text in a given digital image of a folio of a manuscript. The individual segments can then be merged or split into letters or ligatures after analysis by human experts. For each region obtained this way, a transcription can be entered, enabling the user to search the manuscript for similar samples of previously transcribed letters, ligatures or words.

Another system which provides electronic transcription aids is *tranScriptorium*.[2] This is designed for indexing, searching and transcribing scans of handwritten historical documents. The approach chosen here is based on interactive segmentation-free HTR techniques (handwritten text recognition). *TranScriptorium* aims at providing the results of the automatic and semi-interactive transcription process through web portals by attaching the transcribed text to scanned images of the historical manuscripts.

Transcription is just one of many tasks performed by palaeographers when working with manuscripts, hence there are also other tools in development which address further aspects of manuscript research. One of these new tools is the DivaDia[3] system. DivaDia is a document image analysis (DIA) framework that offers semi-automatic layout analysis for digital images of historical manuscript pages. It applies machine-learning (ML) techniques to obtain a model of the layout in digital images of manuscripts which have been annotated by manuscript researchers. The layout model obtained in this way is then used to determine layout

[1] *Diptychon*: http://www.tzi.de/~bjoerng/Diptychon.htm (last accessed 15.04.14).

[2] *transScriptorium*: http://transcriptorium.eu/ (last accessed 03.09.14).

[3] DivaDia: http://diuf.unifr.ch/hisdoc/divadia (last accessed 15.04.14).

properties of large sets of data. The user can accept or dismiss the automatically computed results and thus refine the layout detection performance of the DivaDia framework.

The *Monk*[4] system developed at the University of Groningen, Holland, was primarily conceived for accessing and analysing manuscripts in the Royal Dutch Library. Through a web interface, researchers and volunteers can manually annotate single words in historical texts – for which OCR techniques are not applicable. Classifiers trained on these samples and their annotations allow word retrieval and word recognition, enabling researchers to search through large, handwritten archives. The system is currently being extended to include various historical records such as a section of the Dead Sea Scrolls.

3. The AMAP workbench for manuscript analysis

While the tools mentioned above perform well in their respective areas, they are all specific solutions for specific problems. By providing a one-click interface, they act as 'black box' systems and do not show in detail how the obtained results are computed. Unlike these systems, the AMAP workbench will also offer interactive methods for the typical manual tasks conducted by palaeographers when examining and analysing manuscripts. These tasks can range from simple activities such as measuring geometric features of pages and script properties to more complex procedures such as statistical evaluation and comparing the features of multiple manuscripts. In addition, AMAP will incorporate fully automatic modules which provide the user with tools for selecting specific tasks, in particular layout analysis and grapheme retrieval, a segmentation-free variant of word spotting.

Similar to the systems introduced above, the workbench basically uses a client-server structure (see fig. 1).

On the client side, modern standard web technologies such as HTML5 and JavaScript are used to provide a user interface for organising and viewing manuscripts and performing real-time interaction such as manual measurements and statistical evaluations. The workbench is integrated with a repository at CSMC, where users keep digitised images of their manuscript pages.

The server performs the 'heavy-weight' image-processing tasks and implements the Web Server Gateway Interface

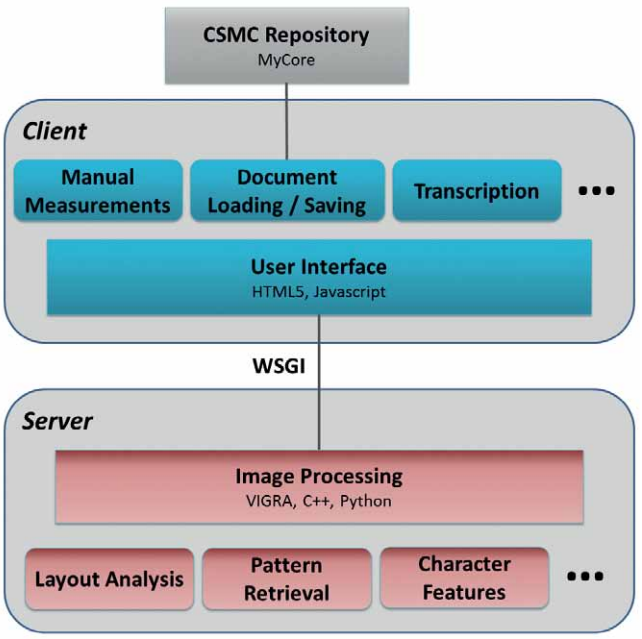


Fig 1: AMAP's system components: a modular approach.

(WSGI) to communicate with the client. Tasks such as image segmentation, interest point computation for pattern retrieval, and frequency analysis for layout segmentation are conducted on the server and use the VIGRA image library, which was developed at the University of Hamburg.[5]

The system can be enhanced with additional features at any time, exploiting its modular structure. In the following sections, we will describe three components which are already available in the current version. The descriptions will illustrate the type of functionalities which are provided for the user.

4. Layout analysis

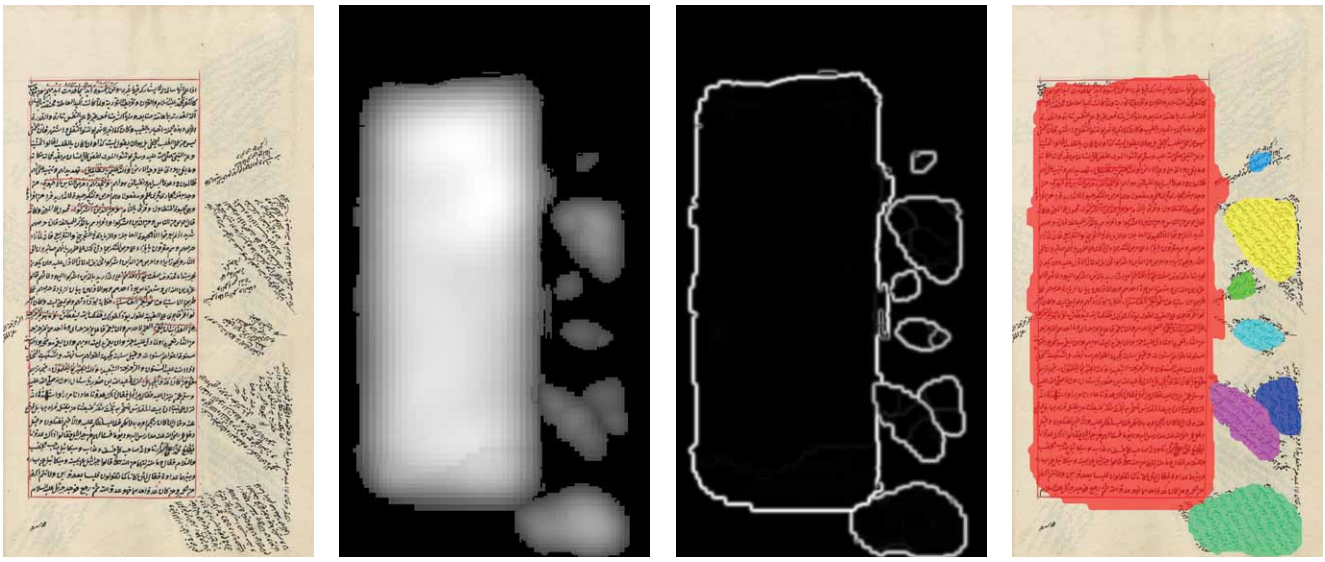
One topic of general interest in manuscript research is the layout of manuscript pages. By counting the number of text lines and measuring the size of the margins and the dimensions of the main blocks of text and paratexts, scholars can form assumptions about the age of a manuscript or the region where it was produced. In this section, we describe our approach and give a short overview of related work.

Postl used Fourier analysis and simulated skew scans to detect the orientation of skewed lines in scanned printed documents.[6] Since then, the problem of layout analysis has practically been solved with regard to printed pages containing

[4] *Monk*: http://www.ai.rug.nl/~lambert/Monk-collections-english.html (last accessed 15.04.14).

[5] Köthe 2000.

[6] Postl 1986.



a) b) c) d)

Fig. 2a: University Library of Leipzig, Ms B. or. 002 / p. 84, 2b: magnitude of detected frequencies above the threshold, 2c: boundaries of homogeneous frequency sections, 2d: false-colour representation of determined homogeneous text blocks.

rectangular-shaped text blocks, tables and images. When it comes to handwritten documents, however, layout analysis is much more challenging. The problem has been approached, e.g. by Bulacu et al.[7] for historical Dutch manuscripts, by Garz et al.[8] using SIFT features to segment initials, headlines and text and by Bukhari et al.[9] using machine-learning techniques to discern between main texts and side-note texts.

Our approach[10] is not limited to specific layouts such as horizontally written text, nor does it anticipate a specific writing system or presuppose a main text region or a text of a certain size. We use the Gabor transform (GT) to determine the frequency of periodic line structures of text blocks. GT is a windowed Fourier transform where a Gaussian window is shifted across the image on a two-dimensional grid. The frequency and orientation with the greatest magnitude are stored for each position. The best results, i.e. the optimal compromise between accurate spatial resolution and reliable frequency magnitudes – are achieved with a window size covering about 6–12 lines of text. In order to discern between text and non-text regions, a threshold is applied to all collected magnitudes of the GT response.

In a second step, a gradient magnitude image is compiled by combining the gradient information of the frequency and the normalised orientation vector, representing the discontinuities of line distance and orientation. In a final step, text regions are determined by applying the watershed principle for segmentation to the gradient magnitude image. Each region can be described by means of position, contour, average line distance and orientation.

This binarisation-free method is able to locate text blocks consisting of at least three lines. It can also discern between blocks written in different orientations (up to 180 degrees) or with different line spacing, even if the text blocks touch each other. Text blocks do not need to be rectangular in shape and can be written using any kind of writing system. The resulting information can also be used to support line segmentation methods or serve as guidance for the script-retrieval process described in the next section.

5. Script retrieval

Using our approach for script retrieval, a scholar can retrieve instances of script patterns or graphemes which are visually similar to a given target pattern. Script patterns may be whole words, parts of words or single characters. Searching for a specific word can be useful for several purposes, for example, to determine all occurrences of the word in a large dataset of manuscript images or to compare the frequency of its occurrence in different manuscripts. Palaeographers

[7] Bulacu, van Koert, Schomaker, and van der Zant 2007.

[8] Garz, Sablatnig, and Diem 2011.

[9] Bukhari, Breuel, Asi, and El-Sana 2012.

[10] Herzog, Solth, and Neumann 2014.

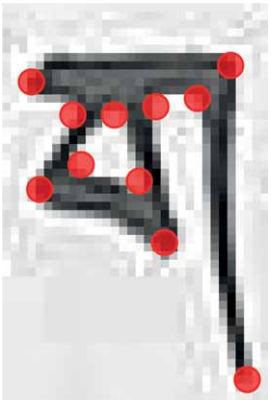


Fig. 3a: Tibetan character with IPs.

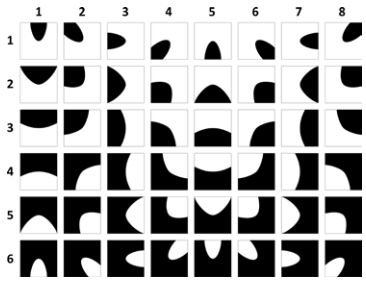


Fig. 3b: Code chart of IP types.

might be interested in analysing a specific part of a script like a ligature or radical of a Chinese character. Our spotting method can efficiently provide them with numerous instances of such segments for further analysis.

Some approaches to word spotting are based on SIFT features[11], whereas others use zones of interest[12] or interest points[13]. The most recent research uses segmentation-free methods, while with Rothfeder's method, words need to be segmented in advance in order to compare them.[14] We have developed a segmentation-free method for our workbench based on Harris corners, which can be applied to any system of writing. The following section describes our approach, which uses a new matching method.[15]

The first step in the process is to determine Harris corners as interest points (IPs). They are the basis for word spotting in any manuscript. These IPs mark locations of strong cornerness or angularity, which are typically found at the end of strokes or at points where strokes cross (fig. 3a). The IPs are then classified by an SVM according to a code table of 48 different types of corner neighbourhoods (fig. 3b). A specific configuration of typed IPs can be understood as a very compact description of the underlying script pattern. The locations and types of corners are then stored in a database where they can be used as a quick index for each query.

Fig. 4: Two examples of the same Chinese character (left); the white pixels represent non-matching regions before and after warping (right).

To specify a query, a user selects a segment of a manuscript image where the IPs and their types are determined as described above. To determine a possible target location for a query, one IP from the query is matched with a type-compatible IP from the data. Using this location as a reference, all other query IPs are checked for type-compatible and location-compatible partners in the data. Acceptance of a target is controlled by a hypothesis test based on probabilities for type and location deviations.

If a strong variability is expected within the script, the comparison can be refined by computing binarised versions of the query image and the segment at a possible target location. The target image is then warped according to the deviations shown by corresponding IPs in the query image, and the number of non-overlapping pixels is taken as a dissimilarity measure.

While one central idea regarding this method was to avoid using features prominent in only a few writing systems, we obtained the best results for Chinese manuscript images. Chinese characters in regular script typically feature a large number of corners, resulting in a correspondingly large number of IPs and good retrieval performance. The approach was also tested successfully on Amharic, Sanskrit and Tibetan manuscripts. Since the majority of the researchers at CSMC deal with Asian writing systems, an evaluation of European scripts has not been carried out yet. The method is going to be optimised in future to improve the results for less compact structures such as longer words.

6. Character features

Challenging palaeographic tasks such as writer identification and verification require a comparison of individual hands.[16] One way of facilitating this is to provide computer support in determining character and stroke features such as the width-height ratio, slant, vertical and horizontal ink distribution and compactness of characters as well as the direction, curvature and length of strokes.

[11] Rusiñol, Aldavert, Toledo, and Lladós 2011; Rothacker, Rusiñol, and Fink 2013.

[12] Leydier, Ouji, LeBourgeois, and Emptoz 2009.

[13] Rothfeder, Feng, and Rath 2003.

[14] Ibid.

[15] Herzog, Solth, and Neumann 2013.

[16] Richter 2006.

Disparities in the digitisation process such as different scanning resolutions or imprecise alignment of manuscript pages can cause problems when comparing absolute values of individual elements such as stroke length or writing orientation in multiple manuscripts. It can therefore prove advantageous to use relational properties between two or more elements. Relational features are less prone to the digitisation problems stated. The distance between text lines, characters or strokes or the angle between strokes and the ratio of their lengths are some examples of meaningful relational features of multiple elements.

AMAP offers various methods for detecting and comparing visual manuscript features. One way to obtain feature values is to measure visual elements manually by means of a digital ruler and compass. Using these tools, manuscript researchers can add geometrical annotations to arbitrary elements in digitised manuscript pages that, in turn, can be evaluated using a statistical module in the workbench. Other character features such as compactness, ink distribution and slant can be computed automatically by applying well-established image-processing methods such as edge detection, pixel projections and gradient histograms.

Alternatively, characters can be segmented automatically into individual strokes using an integrated stroke-extraction module. This module deploys a new stroke-extraction algorithm that extracts moderately straight strokes from character patterns using subpixel watershed segmentation and constrained Delaunay triangulation.[17] Individual features of strokes such as their length and orientation are computed automatically in the process and can be combined to obtain relational features between multiple strokes such as their angle and length ratio.

7. Outlook

This article has presented our approach to devising a modular workbench providing modules for layout analysis, script retrieval and character-feature analysis in the context of manuscript research. The workbench represents work in progress and will be extended by adding further modules in future. Our central idea is to offer a toolset which supports manuscript researchers with traditional palaeographic analysis, but also with advanced tasks. The tools are not limited to any particular type of script, but are designed to be used in an omni-lingual environment, given that the research groups at CSMC deal with almost every type of writing system found on the Asian, African and European continents. A first prototype of the workbench has been developed and will be evaluated and refined in co-operation with members of CSMC.

[17] Solth, Neumann, and Stelldinger 2009.

REFERENCES

Bukhari, Syed Saqib, Breuel, Thomas M., Asi, Abedelkadir, and El-Sana, Jihad (2012), 'Layout Analysis for Arabic Historical Document Images Using Machine Learning', *Proc. 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012)*, 639–644 (doi:10.1109/ICFHR.2012.227).

Bulacu, Marius, van Koert, Rutger, Schomaker, Lambert, and van der Zant, Tijn (2007), 'Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen', *Proc. 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, 357–361 (doi:10.1109/ICDAR.2007.154).

Garz, Angelika, Sablatnig, Robert, and Diem, Markus (2011), 'Layout Analysis for Historic Manuscripts Using SIFT Features', *Proc. 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, 508–512 (doi:10.1109/ICDAR.2011.108).

Herzog, Rainer, Solth, Arved, and Neumann, Bernd (2013), 'Using Harris Corners for the Retrieval of Graphs in Historical Manuscripts', *Proc. 12th International Conference on Document Analysis and Recognition (ICDAR 2013)*, 1295–1299 (doi:10.1109/ICDAR.2013.262).

——, ——, and —— (2014), *Text Block Recognition in Multi-Oriented Handwritten Documents*, Report FBI-HH-B-301/14, Department of Informatics, University of Hamburg

Köthe, Ullrich (2000), *Generische Programmierung für die Bildverarbeitung,* Dissertation, (University of Hamburg, Department of Informatics).

Leydier, Yann, Ouji, Asma, LeBourgeois, Frank, and Emptoz, Hubert (2009), 'Towards an omnilingual word retrieval system for ancient manuscripts', *Pattern Recognition*, 42.9, 2089–2105 (doi:10.1016/j.patcog.2009.01.026).

Postl, Wolfgang (1986), 'Detection of Linear Oblique Structures and Skew Scan in Digitized Documents', *Proc. 8th Int. Conf. Pattern Recognition* (Paris), 687–689.

Richter, Matthias (2006), 'Tentative Criteria for Discerning Individual Hands in the Guodian Manuscripts', *Rethinking Confucianism: Selected Papers from the Third International Conference on Excavated Chinese Manuscripts*, 132–147.

Rothacker, Leonard, Rusiñol, Marçal, and Fink, Gernot A. (2013), 'Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents', *Proc. 12th International Conference on Document Analysis and Recognition (ICDAR 2013)*, 1305–1309 (doi:10.1109/ICDAR.2013.264).

Rothfeder, Jamie L., Feng, Shaolei, and Rath, Toni M. (2003), 'Using Corner Feature Correspondences to Rank Word Images by Similarity', *Proc. Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2003)*, 30 (doi:10.1109/CVPRW.2003.10021).

Rusiñol, Marçal, Aldavert, David, Toledo, Ricardo, and Lladós, Josep (2011), 'Browsing heterogeneous document collections by a segmentation-free word spotting method', *Proc. 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, 63–67 (doi:10.1109/ICDAR.2011.22).

Solth, Arved, Neumann, Bernd, and Stelldinger, Peer (2009), *Strichextraktion und -analyse handschriftlicher chinesischer Zeichen. Report FBI-HH-B-291/09* (Department of Informatics, University of Hamburg) (http://kogs-www.informatik.uni-hamburg.de/publikationen/pub-solth/Strichextraktion.pdf).

PICTURE CREDITS

Fig. 1, fig. 2b-d, fig. 3, fig. 4: © Authors.

Fig. 2a: © University Library of Leipzig.

Contributors

Athina Alexopoulou
Technological Educational Institute of Athens
Department for Conservation of Antiquities and Works of Art
Ag. Spyridonos Str, Egaleo
12210 Athens, Greece

athfrt@teiath.gr

Ayoub Al-Hamadi
Otto-von-Guericke-University Magdeburg
Institute for Information Technology and Communications (IIKT)
P. O. Box 4210, Germany
39016 Magdeburg, Germany

ayoub.alhamadi@ovgu.de

Christian Brockmann
University of Hamburg
Institut für Griechische und Lateinische Philologie
Von-Melle-Park 6
20146 Hamburg, Germany

SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

christian.brockmann@uni-hamburg.de

Ana Čamba
Vienna University of Technology
Institute of Computer Aided Automation
Computer Vision Lab
Favoritenstr. 9/1832
1040 Vienna, Austria

acamba@caa.tuwien.ac.at

Kai Chen
University of Fribourg
DIVA Group (Document, Image and Voice Analysis)
Department of Informatics
Boulevard de Pérolles 90
1700 Fribourg, Switzerland

kai.chen@unifr.ch

Daniel Deckers
University of Hamburg
Institut für Griechische und Lateinische Philologie
TEUCHOS – Zentrum für Handschriften- und Textforschung
Von-Melle-Park 6
20146 Hamburg, Germany

daniel.deckers@uni-hamburg.de

Laslo Dinges
Otto-von-Guericke-University Magdeburg
Institute for Information Technology and Communications (IIKT)
P. O. Box 4210 Germany
39016 Magdeburg, Germany

laslo.dinges@ovgu.de

Roger Easton
Rochester Institute of Technology
Chester F. Carlson Center for Imaging Science
Rochester, NY, USA
USA

easton@cis.rit.edu

Nicole Eichenberger
University of Fribourg
Germanistische Mediävistik
Avenue de l'Europe 20
1700 Fribourg, Switzerland

nicole.eichenberger@unifr.ch

Sherif El-Etriby
Menoufia University
Faculty of Computers and Informations
Gamal Abd El-Nasir
Shebeen El-Kom, Menufia, Egypt

el_etriby100@yahoo.com

Moftah Elzobi
Otto-von-Guericke-University Magdeburg
Institute for Information Technology and Communications (IIKT)
P. O. Box 4210 Germany
39016 Magdeburg, Germany

moftah.elzobi@ovgu.de

Gabriele Ferrario
Cambridge University Library
West Road
Cambridge, CB3 9DR, UK

gf275@cam.ac.uk

Stefan Fiel
Vienna University of Technology
Institute of Computer Aided Automation
Computer Vision Lab
Favoritenstr. 9/1832
1040 Vienna, Austria

fiel@caa.tuwien.ac.at

Gernot A. Fink
TU Dortmund University
Department of Computer Science 12
Otto-Hahn-Str. 8
44221 Dortmund, Germany

gernot.fink@udo.edu

Michael Friedrich
University of Hamburg
Asien-Afrika-Institut
Edmund-Siemers-Allee 1, Flügel Ost
20146 Hamburg, Germany

SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

michael.friedrich@uni-hamburg.de

Angelika Garz
University of Fribourg
DIVA Group (Document, Image and Voice Analysis)
Department of Informatics
Boulevard de Pérolles 90
1700 Fribourg, Switzerland

angelika.garz@unifr.ch

Melanie Gau
Vienna University of Technology
Institute of Computer Aided Automation
Computer Vision Lab
Favoritenstr. 9/1832
1040 Vienna, Austria

mgau@caa.tuwien.ac.at

Mirjam Geissbühler
University of Bern
Institut für Germanistik
Länggasstr. 49
3000 Bern, Switzerland

mirjam.geissbuehler@germ.unibe.ch

Leif Glaser
Deutsches Elektronen-Synchrotron
Notkestr. 85
22607 Hamburg, Germany

Leif.Glaser@desy.de

Björn Gottfried
University of Bremen
Fachbereich Mathematik und Informatik
Am Fallturm 1
28359 Bremen, Germany

bg@tzi.de

Oliver Hahn
BAM Federal Institute for Materials Research and Testing, Berlin
Division 4.5
Unter den Eichen 44-46
12203 Berlin, Germany

University of Hamburg
SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

oliver.hahn@bam.de

Rainer Herzog
University of Hamburg
SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

herzog@informatik.uni-hamburg.de

Fabian Hollaus
Vienna University of Technology
Institute of Computer Aided Automation
Computer Vision Lab
Favoritenstr. 9/1832
1040 Vienna, Austria

holl@caa.tuwien.ac.at

Rolf Ingold
University of Fribourg
DIVA Group (Document, Image and Voice Analysis)
Department of Informatics
Boulevard de Pérolles 90
1700 Fribourg, Switzerland

rolf.ingold@unifr.ch

Andreas Janke
University of Hamburg
Graduate School 'Manuscript Cultures'
Centre for the Study of Manuscript Cultures (CSMC)

andreas.janke@uni-hamburg.de

Agathi Kaminari
Technological Educational Institute of Athens
Department for Conservation of Antiquities and Works of Art
Ag. Spyridonos Str, Egaleo
12210 Athens, Greece

agathakam@yahoo.com

David Kelbe
Rochester Institute of Technology
Chester F. Carlson Center for Imaging Science
Rochester, NY, USA

kelbe@cis.rit.edu

Josep Lladós
Universitat Autònoma de Barcelona
Computer Vision Center
Campus UAB, Edificio O
08193 Bellaterra (Cerdanyola)
Barcelona, Spain

josep@cvc.uab.cat

Mathias Lawo
The Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
Monumenta Germaniae Historica
Jägerstraße 22/23
10117 Berlin, Germany

lawo@bbaw.de

Marcus Liwicki
University of Fribourg
DIVA Group (Document, Image and Voice Analysis)
Department of Informatics
Boulevard de Pérolles 90
1700 Fribourg, Switzerland

marcus.liwicki@unifr.ch

Claire MacDonald
University of Hamburg
SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

claire.macdonald@uni-hamburg.de

Bernd Neumann
University of Hamburg
Fachbereich Informatik
Vogt-Kölln-Straße 30
22527 Hamburg, Germany

SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

neumann@informatik.uni-hamburg.de

Ben Outhwaite
Cambridge University Library
West Road
Cambridge, CB3 9DR, UK

bmo10@cam.ac.uk

Ira Rabin
BAM Federal Institute for Materials Research and Testing, Berlin
Division 4.5
Unter den Eichen 44-46
12203 Berlin, Germany

University of Hamburg
SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

ira.rabin@bam.de

Leonard Rothacker
TU Dortmund University
Department of Computer Science
Otto-Hahn-Str. 8
44221 Dortmund, Germany

leonard.rothacker@udo.edu

Marçal Rusiñol
Universitat Autònoma de Barcelona
Computer Vision Center
Campus UAB, Edificio O
08193 Bellaterra (Cerdanyola)
Barcelona, Spain

Université de La Rochelle
L3i, Laboratoire Informatique, Image et Interaction
Pôle Sciences & Technologie
Avenue Michel Crépeau
17042 La Rochelle Cedex 1, France

marcal@cvc.uab.cat

Robert Sablatnig
Vienna University of Technology
Institute of Computer Aided Automation
Computer Vision Lab
Favoritenstr. 9/1832
1040 Vienna, Austria

robert.sablatnig@caa.tuwien.ac.at

Arved Solth
University of Hamburg
SFB 950 'Manuskriptkulturen in Asien, Afrika und Europa'
Centre for the Study of Manuscript Cultures (CSMC)

solth@informatik.uni-hamburg.de

Marianna Spano
The Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
Regesta Imperii - Regesten Kaiser Friedrichs III.
Jägerstraße 22/23
10117 Berlin, Germany

spano@bbaw.de

Daniel Stökl Ben Ezra
École Pratique des Hautes Études (EPHE), Paris
Section des Sciences historiques et philologiques
4-14 rue Ferrus
75014 Paris, France

stoekl@mmsh.univ-aix.fr

Christopher Stokoe
Cambridge University Library
West Road
Cambridge, CB3 9DR, UK

cms93@cam.ac.uk

Marius Wegner
University of Bremen
Fachbereich Mathematik und Informatik
Am Fallturm 1
28359 Bremen, Germany

mwegner@cs.uni-bremen.de

Hao Wei
University of Fribourg
DIVA Group (Document, Image and Voice Analysis)
Department of Informatics
Boulevard de Pérolles 90
1700 Fribourg, Switzerland

hao.wei@unifr.ch

Lior Wolf
Tel Aviv University
The Blavatnik School of Computer Science
Schreiber Building
P.O.B. 39040
Ramat Aviv, Tel Aviv 69978, Israel

liorwolf@gmail.com

Studies in Manuscript Cultures (SMC)

Ed. by Michael Friedrich, Harunaga Isaacson, and Jörg B. Quenzer

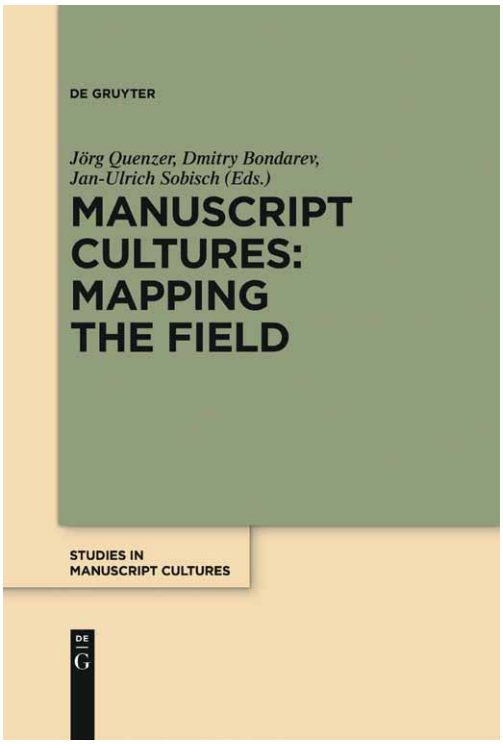
Writing is one of the most important cultural techniques, and writing has been handwriting throughout the greater part of human history, in some places even until very recently. Manuscripts are usually studied primarily for their contents, that is, for the texts, images and notation they carry, but they are also unique artefacts, the study of which can reveal how they were produced and used. The social and cultural history of manuscripts allows for 'grounding' the history of human knowledge and knowledge practices in material evidence in ways largely unexplored by traditional scholarship.

With very few exceptions, the history of the handwritten book is usually taken to be the prehistory of the (printed Western) book, thus not only denying manuscripts their distinct status as carrier medium, but also neglecting the rich heritage of Asian and African manuscript cultures from which, according to conservative estimates, more than ten million specimens survive until today.

The series *Studies in Manuscript Cultures (SMC)* is designed to publish monographs and collective volumes contributing to the emerging field of manuscript studies (or manuscriptology) including disciplines such as philology, palaeography, codicology, art history, and material analysis. SMC encourages comparative study and contributes to a historical and systematic survey of manuscript cultures.

Publisher: de Gruyter, Berlin

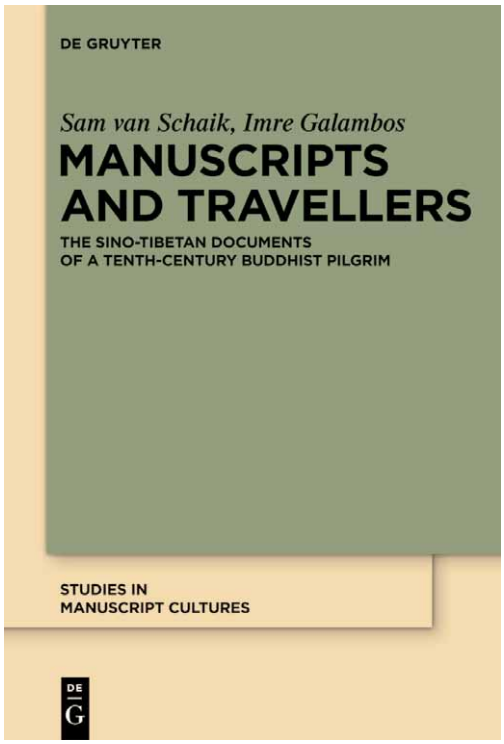
DE GRUYTER



1 - Manuscript Cultures: Mapping the Field

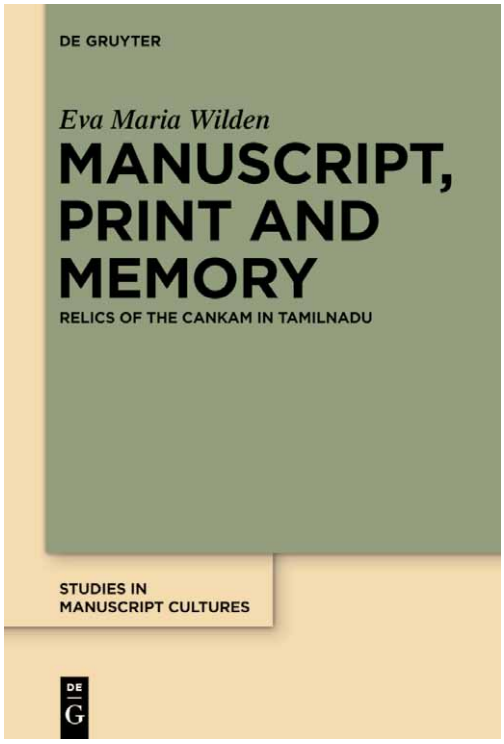
Ed. by Jörg B. Quenzer, Dmitry Bondarev, Jan-Ulrich Sobisch

Script and writing were among the most important inventions in human history, and until the invention of printing, the handwritten book was the primary medium of literary and cultural transmission. Although the study of manuscripts is already quite advanced for many regions of the world, no unified discipline of 'manuscript studies' has yet evolved which is capable of treating handwritten books from East Asia, India and the Islamic world equally alongside the European manuscript tradition. This book, which aims to begin the interdisciplinary dialogue needed to arrive at a truly systematic and comparative approach to manuscript cultures worldwide, brings together papers by leading researchers concerned with material, philological and cultural aspects of different manuscript traditions.



2 - Manuscripts and Travellers - The Sino-Tibetan Documents of a Tenth-Century Buddhist Pilgrim
by Sam van Schaik, Imre Galambos

This study is based on a manuscript which was carried by a Chinese monk through the monasteries of the Hexi corridor, as part of his pilgrimage from Wutaishan to India. The manuscript has been created as a composite object from three separate documents, with Chinese and Tibetan texts on them. Included is a series of Tibetan letters of introduction addressed to the heads of monasteries along the route, functioning as a passport when passing through the region. The manuscript dates to the late 960s, coinciding with the large pilgrimage movement during the reign of Emperor Taizu of the Northern Song recorded in transmitted sources. Therefore, it is very likely that this is a unique contemporarytestimony of the movement, of which our pilgrim was also part. Complementing extant historical sources, the manuscript provides evidence for the high degree of ethnic, cultural and linguistic diversity in Western China during this period.



3 - Manuscript, Print and Memory - Relics of the Cankam in Tamilnadu
by Eva Maria Wilden

The ancient Tamil poetic corpus of the Cankam ('The Academy') is a national treasure for Tamilians and a battle-ground for linguists and historians of politics, culture and literature. Going back to oral predecessors probably dating back to the beginning of the first millennium, it has had an extremely rich and variegated history. Collected into anthologies and endowed with literary theories and voluminous commentaries, it became the centre-piece of the Tamil literary canon, associated with the royal court of the Pandya dynasty in Madurai. Its decline began in the late middle ages, and by the late 17th century it had fallen into near oblivion, before being rediscovered at the beginning of the print era. The present study traces the complex historical process of its transmission over some 2000 years, using and documenting a wide range of sources, in particular surviving manuscripts, the early prints, the commentaries of the literary and grammatical traditions and a vast range of later literature that creates a web of inter-textual references and quotations.

ISSN 1867–9617

"Manuskriptkulturen in Asien, Afrika und Europa"
Universität Hamburg
Warburgstraße 26
D-20354 Hamburg

www.manuscript-cultures.uni-hamburg.de