Report from Dagstuhl Perspectives Workshop 12382

Computation and Palaeography: Potentials and Limits

Edited by

Tal Hassner[1], Malte Rehbein[2], Peter A. Stokes[3], and Lior Wolf[4]

1 **Open University – Israel, IL,** hassner@openu.ac.il
2 **Universität Würzburg, DE,** malte.rehbein@uni-wuerzburg.de **/ University of Nebraska Lincoln, US,** malte.rehbein@unl.edu
3 **King's College London, GB,** peter.stokes@kcl.ac.uk
4 **Tel Aviv University, IL,** wolf@cs.tau.ac.il

Abstract

This report documents the program and outcomes of Dagstuhl Seminar 12382 'Perspectives Workshop: Computation and Palaeography: Potentials and Limits'. The workshop focused on the interaction of palaeography, the study of ancient and medieval documents, with computerized tools, particularly those developed for analysis of digital images and text mining. The goal of this marriage of disciplines is to provide efficient solutions to time-consuming and laborious palaeographic tasks. It furthermore attempts to provide scholars with quantitative evidence to palaeographical arguments, consequently facilitating a better understanding of our cultural heritage through the unique perspective of ancient and medieval documents. The workshop provided a vital opportunity for palaeographers to interact and discuss the potential of digital methods with computer scientists specializing in machine vision and statistical data analysis. This was essential not only in suggesting new directions and ideas for improving palaeographic research, but also in identifying questions which scholars working individually, in their respective fields, would not have asked without directly communicating with colleagues from outside their research community.

Seminar 18.–21. September, 2012 – www.dagstuhl.de/12382
1998 ACM Subject Classification I.7.5 Document Capture, J.5 Arts and Humanities
Keywords and phrases Digital Palaeography, Document Analysis
Digital Object Identifier 10.4230/DagRep.2.9.184

1 Executive Summary

Tal Hassner
Malte Rehbein
Peter A. Stokes
Lior Wolf

The Schloss-Dagstuhl Perspectives Workshop on "Computation and Palaeography: Potentials and Limits" focused on the interaction of palaeography, the study of ancient and medieval documents, and computerized tools developed for analysis of digital images in computer vision. During the workshop, the interaction between domain experts from palaeography and computer scientists with computer vision backgrounds has yielded several very clear themes for the future of computerized tools in palaeographic research. Namely,

- Difficulties in communication between palaeographers and computer scientists is a prevailing problem. This is often reflected not only in computerized tools failing to meet the requirements of palaeography practitioners but also in the terminology used by the two disciplines. Better communication should be fostered by joint events and long-term collaborations.
- Computerized palaeographic tools are often black boxes which put the palaeography scholar on one end of the system, only receiving a systems output, with little opportunity to directly influence how the system performs or to communicate with it using natural palaeographic terminology. The long-term desire is to have the scholar at the center of the computerized system, allowing interaction and feedback in order to both fine-tune performance and better interpret and communicate results. This is crucial if palaeography is to become a truly evidence-based discipline. To this end the use of high-level terminology, natural to palaeography, should be integrated into computerized palaeographic systems.
- Palaeographic data, scarce to begin with, is even more restricted by accessibility and indexing problems, non-standard benchmarking techniques and the lack of accurate meta-data and ground truth information. Multiple opportunities were identified for acquiring data and disseminating it both in the palaeographic research community and outside to the general public.
- Palaeographic research is largely restricted to the domain of experts. Making palaeography accessible to non-experts by using computerized tools has been identified as an effective means of disseminating valuable cultural heritage information while at the same time potentially giving rise to crowdsourcing opportunities, such as those proved successful in other domains.

The manifesto which resulted from this work elaborates on the existing challenges and limitations of the field and details the long-term recommendations that have emerged from the workshop.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Three Pattern-Recognition Approaches to the Automatic Identification of the Writers of Ancient Documents

Dimitris Arabadjis and Micalis Panagopoulos (National TU – Athens, GR)

Joint work of Papaodysseus, Constantinl; Arabadjis, Dimitris; Rousopoulos, Panayiotis; Giannopoulos, Fotios; Blackwell, Chris

Dating the content of ancient documents is absolutely crucial for History and Archaeology. For example one of the most prominent historians, Professor Christian Habicht, has recently written that proper historical use of inscriptions can only be made if they can be dated. However, writers of ancient inscriptions and manuscripts, as a rule, did not sign or date their documents. So far, dating the content of ancient inscriptions and manuscripts is a very difficult task and is based on scholars' instinct and frequently subjective considerations. One main goal of the work that was presented is to perform quantitative analysis on the scribal hands, so that the relationships among these volumes and their relative dates of production are obtained. This will be achieved by means of writer identification, since the working careers of most ancient writers covered about 20 to 25 years. So, if one could attribute a document to a writer, then the content of the document gains a date immediately, which is clearly the time period during which the writer was active. Hence, three different approaches for the identification of the writer of ancient documents were outlined.

The first approach estimates ideal representatives of selected alphabet symbols for each document separately and then compares these representatives. The second approach introduces a new mathematical notion, a kind of two dimensional curvature, so that the various alphabet symbols realizations can be optimally matched and compared after proper associated transformations. The third approach uses exhaustive comparisons based on classical curvature and set-theoretic similarity measures. In addition, a number of cases were presented where identification of the writer(s) of the ancient documents is of great importance for the analysis of their content.

3.2 Multi-Source and Multi-View 3D Data Exploration

Matthieu Exbrayat (Université d'Orleans, FR)

Joint work of Exbrayat, Matthieu; Martin, Lionel

Multi-source data consists of data, coming from various producers, that share a common object of interest. During the 2008-2011 Graphem Project, we have been studying the spatial visualisation of medieval writing samples, by the means of an interactive spatial projection tool named Explorer3D that we developed at LIFO. In this tool we take as an input the description of the writing samples, in the form of a set of numerical features. Based on these features, we use 3D projection techniques, such as Principal Component Analysis, to create a 3D space in which each sample is represented by a point, so that the distance between points in the 3D space reflects the proximity (or distance) of the writing samples according to the input features. We offer various interaction extensions, for instance to visualise the writing samples or to modify the 3D projection based on the user's visual analysis of the relevance of this projection.

The feature sets we based our projections on during Graphem were produced by the other computer science partners of this project. Several sets of features have been proposed and evaluated. Nevertheless each of them has been studied independently. Arguing that a visual comparative study of these feature sets might help understanding their respective strengths and weaknesses, we have extended Explorer3D in order to load and visualise a set of writing samples using several feature sets simultaneously, and thus using several 3D scenes. Various interactive tools have been developed or adapted in order to compare how writing samples appear similar or dissimilar across the available 3D scenes, and thus across the underlying feature sets.

3.3 Computerized Paleography of Hebrew Writing from the First Temple Period

Shira Faigenbaum (Tel Aviv University, IL)

The only texts from the First Temple Period in Israel and Judah that endured the harsh local climate were written in ink on pieces of pottery (ostraca). The discipline of Iron Age epigraphy classically involves manual labor in analyzing the inscriptions and establishing a comparative typology of characters. However, this approach may unintentionally mix documentation and interpretation. We introduce image processing and pattern recognition methods to the field of First Temple epigraphy, minimizing the epigrapher's involvement in activities prone to subjective judgment. Our work comprises various image acquisition techniques, image quality assessment, image binarization, and letter comparison metrics.

3.4 Modern Technologies for Manuscript Research

Melanie Gau and Robert Slabatnig (TU Wien, AT)

This paper on multispectral imaging and image enhancement addressed the following:

- Making faded-out text legible
- Unmixing and inpainting techniques for palimpsests
- Stroke and Character Analysis
- Information about writing tools
- Degraded Character Recognition and OCR
- Layout Analysis and Text Line Detection

3.5 Experiments in the Digital Humanities

R. Manmatha (University of Massachusets – Amherst, US)

Joint work of Manmatha, R.; Rath, Toni; Rothfeder, Jamie; Lavrenko, Victor; Yalniz, I. Zeki; Can, Ethem
Main reference I. Yalniz, E. Can, R. Manmatha, "Partial duplicate detection for large book collections," in Proc. of the 20th Conf. on Information Knowledge and Management (CIKM'11), pp. 469–474, ACM, 2011.
URL http://doi.acm.org/10.1145/2063576.2063647
URL http://ciir.cs.umass.edu/irdemo/hw-demo/

Searching historical handwritten manuscripts is a challenging task given that there is a way to go before handwriting recognition is reasonably accurate. I described two approaches to doing this. The first is word spotting – where the query is a word image – and the system looks for similar word images in a set of documents. Word spotting has now been investigated by a large number of researchers for handwritten and printed documents. The second is based on relevance models where the word images are automatically annotated with vocabulary words and their probabilities using a joint probability model. I showed a demonstration of such an automatic system on a sample of George Washington's documents. Such an automatic system requires a number of other automatic steps including automatic word segmentation and I described a technique for doing this.

In the second part of the talk I described an efficient automatic approach to linking old printed (and scanned) books by finding partial duplicates. For example, in a large collection of books one can find different versions of Shakespeare's Othello or Virgil's Aeneid. One version may just have the main text, a second scholarly version may have a lot of footnotes while a third may have substantial additional text in the form of an introduction and endnotes. By representing a book as a sequence of unique words one can find partial duplicates efficiently. A similar approach may be used to find translations of books without explicitly translating a book.

3.6 Challenges in Palaeography for which Computer Sciences Might Offer Some Solutions

Wendy Scase (University of Birmingham, GB)

The aim of my presentation was to facilitate identification of kinds of important palaeographical problem that computer sciences might make a major contribution to solving in the next few years. I illustrated a variety of problems that I have encountered in my own projects on Middle English manuscripts and offered a rudimentary analysis of some of the different kinds of problem palaeographers and manuscripts specialists are faced with. The problems discussed were of three kinds.

1) Problems in the analysis of digital images of medieval manuscripts that might be amenable to computer vision methods. Digitisation of manuscripts is proceeding fast and has transformed researchers' access to manuscripts but to maximise the benefit of digitised manuscript images for research, I proposed, we need to provide researchers with computer-assisted ways to search and analyse the images. I illustrated the method used for my Vernon manuscript project (Oxford, Bodleian Library, MS Eng.poet.a.1) where a full transcription and detailed manuscript description link to digital images of the entire manuscript greatly

assisting searching and analysis of this very large and important codex. I proposed that datasets such as this might provide training data for the development of computer-assisted recognition and searching so that such metadata could be created in a less labour-intensive way in future.

2) Problems in the linking of manuscript metadata and a possible solution. Much metadata about manuscripts is in digital form; often this is in discrete and distributed datasets. In recent years strides have been made to link this data (the example given was the Manuscripts Online project hosted at the University of Sheffield) opening up the possibility of conducting research across much larger datasets in future. However problems remain because different datasets use different vocabularies and languages. Tackling this by conventional means would be a huge task (as attempts have shown) and I suggested that it might be possible to semi-automate the making of dictionaries and thesauri of synonyms and related terms. I proposed it might be possible to harness the search activity of expert searchers to develop thesauri of synonyms and related terms to aid the search function. I proposed that perhaps the problem might be addressed by technology similar to that which underlies commercial search engines where systems harvest user activity and build up data on related and synonymous search terms that they then offer to users.

3) Problems in the Public Understanding of Manuscripts. I proposed that one of the most challenging and urgent problems is to improve public understanding and valuing of medieval manuscripts if they are to be a useful cultural resource in future and not to remain only available to a very small elite as they were when they were first made. I described successful forms of public engagement with manuscripts (particularly the Vernon MS) and asked for suggestions of how gamification and other computer-based strategies might be harnessed for this purpose.

3.7 Bringing the Digital to Palaeography: Some Background and Challenges

Peter A. Stokes (King's College London, GB)

The purpose of this talk was to set the scene for the following days' discussion. It provided a brief overview of the recent history of the field, in an attempt to identify the status quo. There was no attempt to canvass all the projects to date, but instead to analyse the problems that they have addressed and the problems that still remain. In particular, it was argued that there is still a significant distance between 'computational' projects, lead principally by computer scientists, and 'palaeographical' projects lead by humanities scholars. Not only do they take different approaches, but the questions being asked by the two groups are often very different as well, and as a result neither group has had very much influence in the day-to-day research of the other. These difficulties are in addition to ones that have been previously identified in the literature, such as questions of trust, transparency and verifiability.

In order to start discussion a number of suggestions were made regarding:

- Which sorts of question are most amenable to computational methods from the position of a humanities scholar, and why?
- Which sorts of questions are interesting to humanities scholars but are not yet being addressed computationally (but could be)?

- What is the future of ‘computational’ vs other forms of ‘digital’ palaeography?
- Can related discussions in other branches of Digital Humanities help?

3.8 The Graphem Research Project

Dominique Stutzmann (CNRS – Paris, FR)

Joint work of Stutzmann, Dominique; Smith, Marc; Muzerelle, Denis; Gurrado, Maria; Eglin, Véronique; Bres, Stéphane; Lebourgeois, Frank; Joutel, Guillaume; Daher, Hani; Vincent, Joutel; Leydier, Yann; Exbrayat, Mathieu; Martin, Lionel; Moalla, Ikram; Siddiqi, Imran

Main reference D. Muzerelle, M. Gurrado, (eds), “Analyse d’image et paléographie systématique: travaux du programme ‘Graphem’,” Paris, Association ‘Gazette du livre médiéval’, 2011.

The research program GRAPHEM (Grapheme based Retrieval and Analysis for PalaeograpHic Expertise of medieval Manuscripts, 2008–2011), funded by the French National Research Agency, aimed at improving the data mining and image processing techniques applied to medieval scripts and their classification with several methods: outline directions, generalized cooccurrences, stroke categorization. The results of the unsupervised categorization showed too much overlapping to be used properly by academic end users from the humanities. The supervised methods reached very satisfying results in assigning a random handwriting to a category of script and to a century. Nevertheless the global approach from the methods of cooccurrences or curvelets makes it very difficult to relate the results back to the visual differences that the palaeographer can observe. Even in the chain code method (stroke categorisation), the calculated features cannot be traced back to the morphological ones that the palaeographers are used to observing. These features are observed through other methods: metrology (human-based measurements of writing features such as writing angle, density, word spacing), graphonomics, and (allo)graphetic transcriptions, which are more directly relevant to state the script evolution and also may have positive incidence on image analysis and text recognition.

3.9 The Ongoing Effort to Reconstruct the Cairo Genizah

Lior Wolf (Tel Aviv University, IL)

Many significant historical corpora contain leaves that are mixed up and no longer bound in their original state as multi-page documents. The reconstruction of old manuscripts from a mix of disjoint leaves can therefore be of paramount importance to historians and literary scholars. In collaboration with the The Friedberg Genizah Project, we showed that visual similarity provides meaningful pair-wise similarities between handwritten leaves and then went a step further to suggest a semi-automatic clustering tool that helps reconstruct the original documents. The proposed solution is based on a graphical model that makes inferences based on catalog information provided for each leaf as well as on the pairwise similarities of handwriting. Several novel active clustering techniques were explored, and the solution has been applied to a significant part of the Cairo Genizah, where the problem of joining leaves remains unsolved even after a century of extensive study by hundreds of human scholars.

4 Working Groups

4.1 Acquisition of Images

Dimitris Arabadjis, Shira Faigenbaum, Robert Sablatnig, and Timothy Stinson

Standards for digital image acquisition need to be clearly articulated and the same protocol followed by all digital imaging projects when possible. These include practices such as:

1. Using color bars and grey cards
2. Documenting illumination used (e.g., how many lamps, their positioning, diffuser used)
3. Including references to size of original objects
4. Documenting information about photographic equipment used
5. Using shared standards for metadata descriptions of digitized objects
6. Including information that links multiple names and catalogue records when original objects have no single identifier (e.g., a manuscript with shelfmarks that change over time and that is also referred to by other common names in scholarly literature)
7. Establishing file naming conventions in order to facilitate the creation of good metadata and their proper sequence of images when books or other documents are being digitized.

Additionally, if one takes several images of the same object (e.g., jpeg, tiff, multiple sizes, multispectral), it is important that metadata indicates that these are images of the same object.

It would be helpful to have a set of guidelines articulating how to capture digital and analogue images across a wide range of technologies – e.g., scanning objects and photographic negatives, using digital and analogue cameras, digitizing microfilm.

Copyright or contractual use restrictions on photographs of cultural heritage items create many barriers for researchers. In many cases, tax-funded or state-supported research projects must expend significant financial and human resources on negotiating and paying for reproduction rights, even if those rights are being obtained from state repositories. Furthermore, rights tend to be granted only to scholars or research groups on a one-by-one basis, which frustrates large-scale studies of collections of manuscript images. Making large sets of images more easily available at an international scale would greatly facilitate the pursuit of significant new research questions (e.g., large-scale comparative studies of handwriting that map regional and national developments of hands across time).

It might be useful to call attention to libraries and museums with progressive policies that help researchers, such as the Austrian State Library, which makes images paid for by one project freely available to subsequent researchers needing those images.

4.2 Tools

Nachum Dershowitz, Matthieu Exbrayat, Eyal Ofek, Micalis Panagopoulos, and Ségolène Tarte

What tools are needed to progress in this field? The assets of computers are their ability to deal with big data, using memory, distinction/identification of fine differences, and rare occurrences. The assets of humans are dealing with complex data, making sense of the data, and gestalt questions. It is vital that these two sets of assets are combined through semi-automatic and interactive tools, not through 'black boxes': we must always keep humans in the loop! This includes

- Provide training data / annotated data
- Online training / expert-in-the-loop
- Crowd-sourcing

Rather than a single product, we also need a collection of tools that contribute to each other: a toolbox to account for the different needs of different researchers.

Low-level tools:

- Binarization – segmentation – alignment / matching / registration (for later comparison) – physical feature extraction – expert feature extraction (angles, curvatures, strokes. . .) – similarity measures (for comparison between characters, words, texts, fragments, documents, corpora)

Medium-level tools:

- Clustering – classification – character recognition – word spotting – searching (text via string – text via image – image via text – image via image – characters) – image-text correspondence
- Databases: organisation of data in a way that allows fast queries of metadata, transcripts, text qualities, etc.

Higher-level tools:

- Interfaces, ergonomy (CHI) – searches of combinations of characters/words (bigrams, trigrams) – correspondence of expert vocabularies – inferences of paraphrases and synonyms for searches through metadata
- A transcription tool to make the connection between text as shape and text as meaning

Other principles for development:

- Feedback loops and cognitive triggers: drawing/touch screen technologies – simple interactive image enhancements – visualization aspects of interactions with these tools (of results, of databases) – interactive visualisations (e.g. time varying graphs) – customizable visualisations – multiple languages – rationale building support, tracking of expert hypotheses in interpretation building – statistical tools with tests of significance – information sharing – sounding the texts
- Web-services to provide access to such tools via internet?

This topic has potential links with medical imaging, cognitive sciences, CHI, and NLP, all of which should be explored in future work.

4.3 Content and Context

Melanie Gau, R. Manmatha, Ophir Münz-Manor, Wendy Scase, and Dominique Stutzmann

Scholars need links from image to text (and vice versa): many manuscripts are already imaged but are not accessible in any way except to look at. Linking then becomes the key issue. This should ideally be done automatically and involving multiple forms of content, including not only images and transcripts but also other metadata such as contextual information, art references, articles and papers, content/semantics, codicology, textology, other discreet distributed datasets, named entities, descriptions, and so on.

The question, then, is how. A broad variety and combination of technical approaches and tools is required, e.g. word spotting, finding named entities (both using underlying dictionaries and also via more visionary approaches such as crawling the internet), spotting symbols, controlled vocabularies, text alignment between different versions of the same text such as old or even possibly faulty transcripts (also for acquiring training data), reference corpora, standardised datasets, handwriting recognition, alignment techniques, automatic creation of thesauruses to support queries/resource discovery. It should be possible to find all instances of a word in all images and texts, and to map vocabulary relations/keywords/concepts applicable to and mixable with different data sets, languages, collections, and intersections of information.

We would also like to recommend a note on an EU-wide harmonisation of copyright given the very wide range of policies and freedoms/restrictions across different institutions let alone countries.

4.4 Challenges and Limitations

Dimitris Arabadjis, Melanie Gau, and Ségolène Tarte

We face challenges, rather than limitations, in that the issues discussed here are not necessarily insurmountable. Technical limitations are not reviewed here because, in the light of the potential communications problems, they seem largely surmountable. The discussions and round table in this workshop has revealed that that a lot more is possible than single experts could predict, so any prognosis of technical limitations could have the risk of pre-emptive delimitations.

Current problems include computational limitations, access to data, issues of data retrieval, flexibility of searches (too flexible or too rigid – precision and recall). Major bottlenecks include communication, namely differences of terminology not only between disciplines, but also within disciplines due to different traditions in various specialities (eg. classics, slavonic studies, medieval studies all have different traditions). The tradition in computer science for image processing vs data mining has expert vocabularies (within a given field) which are a very abstract way of formulating problems that might not translate well into formal language. Mid-level features might be a useful compromise – a slowing down approach. This has the disadvantage of likely constraining the potential of each discipline, but better alternatives have not yet been found. A meeting ground is needed. Computer science has a convention of not deriving natural interpretation from the methodology. What is excluded from systematic analysis at the moment is context and meaning, which are crucial (indeed, the whole point) for palaeographers. The output needs further cognitive processing to be interpreted, and computer science doesn't really have ways to do that, nor a tradition to do that. Instead there is need for systematisation and formatting of approaches: this will lead to better exchange but at the expense of less room for creativity both for palaeography and computer science. Nevertheless a common language is needed, for example for features in image processing vs features in palaeography. This in turn leads to issues of trust vs anxiety about black-boxes. Mutual education is also needed for this: an understanding on both sides of the main principles if not of detailed methodologies. For this, the middle-person/translator becomes vital. There is need either for an extra person for the role, or for one (at least) of the experts to be trained. This is the end of the age of the lone scholar in the humanities as well as in computer science.

4.5 Relevance to Society

Wendy Scase, Eyal Ofek, and Ophir Münz-Manor

Manuscripts are one of the major sources of knowledge of human culture and society for most of history. All of the world's written heritage produced before the invention of printing is handwritten. Much written heritage dating from after the introduction of printing is also in manuscript form. Unlike printed texts, manuscript sources are often highly inaccessible. They pose challenges of legibility, of interpretation, of language, and of subject matter. Owing to these challenges manuscript materials are often accessible for only a very small number of highly- trained expert groups. They also pose challenges of discovery and physical accessibility. There are hundreds of thousands of manuscript materials. They are scattered across the world in libraries, archives, museums, and private collections and no single catalogue or list exists to discover this material. Each manuscript source is unique and requires specialist curation and conservation. Exposing these materials to too much handling could result in damage and destruction.

For these reasons, despite their importance to knowledge of human history and culture, manuscripts have remained a largely untapped cultural resource. Exceptions to this neglect, such as the Book of Kells (now in Trinity College, Dublin), a book that has inspired art, regional tourism, and has become iconic of a culture, show how manuscripts can be sources of economic activity and creativity. Another example is the Rothschild Codex, a prayerbook decorated in the Flemish style. One of its iconic illustrations has been used to create an i-pad cover. Digitisation and other computer-assisted research opens up the possibility of tapping into this huge, unused cultural and economic resource to benefit society. An example is the Vernon manuscript (Bodleian Library, Oxford). A recent digitisation and research project on this manuscript has enabled it to become known to a much wider audience and to connect people with their regional literary and linguistic heritage.

Finding solutions to the problems involved in making manuscript culture more accessible is expected to have technological benefits beyond the heritage domain. Research into these problems, such as how to search digital images using computer vision methods, is tackling problems at the edge of what technology today can achieve. This work can be expected to yield results with applications in all of the other fields where computer vision could make a difference.

5 Open Problems

The workshop participants have identified a multitude of research questions and open problems. These are itemized below and are further explored in the manifesto which resulted from the workshop.

- There are different techniques (text recognition, word-spotting, image analysis) and different questions (writer identification, classification): the question is how to make better use of them.
- Wordspotting is very appealing to cultural heritage institutions since it may prove very useful for indexing large collections, but, research remains to be done on:
  - Pre-processing of images (background and foreground)
  - Typing words, creating an ideal image of what the user is searching, and then searching in a very user friendly way (although this needs a lot of research to be carried out across collections and data-sets with very different script families).
  - Above all, taking the variability of the graphical system into account. If the end user is typing the letters, the system has to manage all allographs to find the different forms of a word.
- Prior knowledge:
  - We need to include more textual resources, so that the computer can have a better separation of words (current dictionaries are still not enough).
  - We need to combine different techniques and prior knowledge efficiently. How can we automatically align available digital images with available texts (even if not direct transcriptions)?
  - We need to create a system for aligning, overcoming textual variability, extracting forms and giving the possibility of monitoring all data at different levels (word/letter) and adding new information (abbreviations, allographs).
- Combining techniques: we need to incorporate text recognition for the alignment of images. This
  - would create a complete dictionary of all forms
  - would allow major synoptic editions
  - would create a standard data-set (much bigger than the IAM data-set)
  - would enable real research on script history
- Image analysis: creating 'mid-level' features. We need research for creating new features, inspired by human expertise
- New features:
  - Strokes (identification of different strokes; analysis of them in a palaeographically accurate way)
  - Allographs (intra-allographic and inter-allographic variability)
  - At a letter scale, to be combined with text recognition (cf. 'Efficiently combine different techniques', above, since text recognition is not a solved problem).
- Image analysis: matching existing features with visual cognition

- Doing research to interpret the features that already exist. (NB it is clear that the features do not measure what the human eye perceives, but what they perceive is probably correlated to formal phenomena that the human eye/brain can be aware of). Combining techniques and prior knowledge: use the content at the same time (identify the letters and see how they influence the measurements of features)
- Ergonomics, cognitive approach, and visualization
- Visualization of data, and presentation of data that adressess the pre-attentive perception of the researchers is not only an efficient way to promote dialog with Humanities: visualization also helps also computer science for validation during the research process. It is efficient, accurate, and offers a cross-validation since it confronts the results with another semiotic (e.g. analysis of contours, if you visualize them, you can tell better than through mathematical cross-examination if the results are possible).
- Enhanced visualization is a way to provide researchers in computer science with feedback of experts and researchers from other sciences and to support interaction of researchers
- Research remains to be done on visualization and human-computer interface and the cognitive needs to improve the comprehension of the results => user groups studies
- Hyperspectral imaging could be more efficient if a program were added into the camera to set the parameters automatically
- Visualization is also a way to efficiently introduce mid-level features

6 References

1 Tanya Clement, Sara Steger, John Unsworth, and Kirsten Uszkalo. *How Not to Read a Million Books*. http://people.lis.illinois.edu/~unsworth/hownot2read.html

2 F. Cloppet, H. Daher, V. Eglin, H. Emptoz, M. Exbrayat, G. Joutel, F. Lebourgeois, L. Martin, I. Moalla, I. Siddiqi, N. Vincent. New tools for exploring, analysing and categorising medieval scripts. *Digital Medievalist* 7, 2011. http://digitalmedievalist.org/journal/7/cloppet/

3 Tom Davis. The practice of handwriting identification. *The Library* 8:251–276, 2007. doi: 10.1093/library/8.3.251

4 Albert Derolez, *The Palaeography of Gothic Manuscript Books*, 2003. Cambridge University Press.

5 Matthieu Exbrayat and Lionel Martin. *Explorer3D*. http://www.univ-orleans.fr/lifo/software/Explorer3D/

6 David Ganz. 'Editorial palaeography': One teacher's suggestions. *Gazette du livre médiévale* 16:17–20, 1990. http://www.palaeographia.org/glm/glm.htm?art=ganz

7 Martin Jessop. Digital visualisation as scholarly activity. *Literary and Linguistic Computing* 23:281–293, 2008. doi: 10.1093/llc/fqn016

8 Lionel Martin, Matthieu Exbrayat, Guillaume Cleuziou and Fréderic Moal. Interactive and progressive constraint definition for dimensionality reduction and visualization. *Advances in Knowledge Discovery and Management* Vol. 2 (AKDM-2), pp. 121–136, 2012. Springer

9 Wendy Scase. Medieval manuscript heritage: Digital research challenges and opportunities. *Safeguard of Cultural Heritage: A Challenge from the Past for the Europe of Tomorrow: COST Strategic Workshop, 11-13 July, 2011, Florence*, pp. 97–99, 2011. Florence University Press.

10 Wendy Scase, ed. *The Vernon Manuscript: A Digital Facsimile Edition of Oxford, Bodleian Library, MS Eng.poet.a.1*, Bodleian Digital Texts 3, 2012. Oxford.

11 B. Sculley and D.M. Pasanek. Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing* 23:409-424, 2008. doi: 10.1093/llc/fqn019

12 Smith, M.H. *Les formes de l'alphabet latin, entre lécriture et lecture*, 2011. Paris.

13 Peter A. Stokes. Palaeography and image processing: Some solutions and problems. *Digital Medievalist* 3, 2007/8. http://www.digitalmedievalist.org/journal/3/stokes/

14 Peter A. Stokes. Computer-aided palaeography, present and future. In *Kodikologie und Paläographie im Digitalen Zeitalter — Codicology and Palaeography in the Digital Age*, ed. by M. Rehbein *et al.*, 2009, pp. 313-42. Books on Demand. urn:nbn:de:hbz:38-29782

15 Stutzmann, D. Paléographie statistique pour décrire, identifier, dater…Normaliser pour coopérer et aller plus loin? In *Kodikologie und Paläographie im Digitalen Zeitalter 2 — Codicology and Palaeography in the Digital Age 2*, ed. by F. Fischer *et al.*, pp. 247–277, 2009. Books on Demand.

16 *DigiPal: Database and Resource of Palaeography, Manuscripts and Diplomatic.* http://www.digipal.eu

17 *Late Medieval English Scribes.* http://www.medievalscribes.com

18 *ManCASS C11 Database of Script and Spelling.*
http://www.arts.manchester.ac.uk/mancass/C11database/

19 *Manuscripts Online: Written Culture 1000 to 1500*
http://manuscriptsonline.wordpress.com/

20 *Mapping the Republic of Letters.* https://republicofletters.stanford.edu

21 *The Vernon Manuscript Project*, University of Birmingham.
http://www.birmingham.ac.uk/vernonmanuscript

Participants

- Dimitris Arabadjis
National TU – Athens, GR
- Nachum Dershowitz
Tel Aviv University, IL
- Matthieu Exbrayat
Université d'Orleans, FR
- Shira Faigenbaum
Tel Aviv University, IL
- Melanie Gau
TU Wien, AT
- Tal Hassner
Open University – Israel, IL
- R. Manmatha
University of Massachusets – Amherst, US
- Ophir Münz-Manor
The Open University of Israel – Raanan, IL
- Eyal Ofek
Microsoft Res. – Redmond, US
- Micalis Panagopoulos
Ionian University – Corfu, GR
- Robert Sablatnig
TU Wien, AT
- Wendy Scase
University of Birmingham, GB
- Timothy Stinson
North Carolina State Univ., US
- Peter A. Stokes
King's College London, GB
- Dominique Stutzmann
CNRS – Paris, FR
- Ségolène Tarte
University of Oxford, GB
- Lior Wolf
Tel Aviv University, IL