Universität Basel

Departement
Altertumswissenschaften

NEO paleography

analysing ancient handwritings in the digital age

JANUARY 27-29, 2020
KOLLEGIENHAUS, REGENZZIMMER 111, PETERSPLATZ 1
BASEL, SWITZERLAND
–
ORGANISED BY DR. I. MARTHOT-SANTANIELLO
D-SCRIBES PROJECT

FNSNF Swiss National Science Foundation

Member of the Swiss Academy of Humanities and Social Sciences www.sagw.ch

SGOA Swiss Society for Ancient Near Eastern Studies

SVAW ASEA

Max Geldner Stiftung

Neo-Paleography – Analysing Ancient Handwritings in the Digital Age

Basel, January 27–29 2020

Monday, January 27

Nachum Dershowitz, Adiel Ben-Shalom, Lior Wolf *in abs.* (Tel Aviv)

Computerized Paleography: Tools for Historical Manuscripts

We present an overview of Tel Aviv University's work on digital paleography. TAU researchers have developed paleography tools for the Cairo Genizah corpus as well as for the Dead Sea Scrolls (DSS) and other datasets.

For the Genizah manuscripts, we showed that automatic handwriting-matching functions, derived from non-specific features obtained from a corpus of writing samples, can be used to identify fragments that originate from the same original work. In addition, we showed that clustering various Genizah documents by script style, without being provided any prior information about the relevant styles, gives classes that agree very closely with the accepted palaeographic taxonomy. Similar results obtained with Tibetan scripts. We can also use the same kind of descriptors to group similar-looking letters together. We experimented with several languages, including Coptic.

In related work, we developed a novel method to evaluate the relative importance of different paleographic descriptors. Subjective evaluation of distinctiveness could potentially be misleading. We used information gain to measure the discriminative power added by each individual descriptor. Then using a powerful tool adapted from computational biology, viz. Gene Set Enrichment Analysis, we determined the relative significance of each character. Considering late Anglo-Saxon miniscule, it was unsurprising that the algorithm found ash (æ) to be most significant, but it was quite surprising to paleographers that the Tironean nota came in second.

Working on the DSS dataset, we presented a novel approach to modeling broken letters from the edges of the ancient Qumran scrolls. This method can be used to assist letter reconstruction by paleographic experts. We also developed a novel method for searching and matching new photographs of individual DSS fragments with the fragments contained in a dataset of many old infrared photographs of plates of fragments. To this end, we developed a deep-learning-based segmentation method and a cascade approach for template matching, based on scale, shape analysis and dense matching.

Finally, we discuss recent results in handwritten text recognition and its implication to computation paleography. Working on Tibetian historical documents corpus, we developed a novel method based on neural networks and domain adaptation for segmentation and recognition of umê, a headless handwritten form of Tibetian script.

Mladen Popović, Lambert Schomaker, Maruf Dhali (Groningen)

Digital Palaeography of the Dead Sea Scrolls for Dating Undated Manuscripts

One of the goals of the ERC project The Hands that Wrote the Bible is to develop for the Dead Sea Scrolls a triple-perspective time-axis calibration, using three information sources: 1. physical, material-based evidence from radiocarbon dating (C14), 2. geometric, writing style-based evidence from computational intelligence methods, and 3. palaeographic evidence from the humanities.

In this presentation, we show how we combine palaeography, artificial intelligence, and radiocarbon dating in order to determine the typological development of Hebrew/Aramaic writing styles in the Dead Sea Scrolls by providing a quantitative basis for the typological estimates of traditional palaeography. We demonstrate the use of state-of-the-art pattern recognition techniques for script-style evolution analysis using both textural features (joint directional probability distributions) and allographic features (grapheme-emission distributions).

The aim is to be able to do tracking or interpolation of writing styles features over time that will allow making estimates for manuscripts that do not have a C14 dating, or are not internally dated (which almost none are), and are only available as a photograph or digital image. In other words, the aim is to be able to date undated scrolls with a precision that was not possible before.

Gemma Hayes, Maruf Dhali (Groningen)

Identifying Dead Sea Scribes: A Digital Palaeographic Approach

This presentation explains how we applied automatic writer identification techniques to fifty-eight of the Dead Sea scrolls for scribal identification. We applied digital tools to these fifty-eight manuscripts to overcome specific challenges with them. The primary challenge that there are both similarities and differences between these manuscripts, but it is difficult to discern if the resemblances are because one scribe penned them, or because numerous scribes penned them in script style.

Our presentation will demonstrate the four steps we have taken to identify the most prolific scribe in the Dead Sea scroll collection.

1) Preprocessing / Binarisation

The binarisation process, one of the vital preprocessing techniques, separates the foreground (meaningful information, in general, the text-ink) from the background (the surface material, non-text area). To apply automatic writer identification algorithms to digital images, effective preprocessing is essential.

2) Hinge Feature

This feature, along with its different variants are developed from the understanding that the distribution of directions in letterforms is unique to a writer's hand. The hinge feature measures on the pixel level the dispersal of hundreds of modes of direction within regions of interest.

3) Fraglet-based features

The Hinge is a textural and angular based feature taping into information over a total scan. A fraglet approach taps into allographic information; not whole allographs, but sub-allographs, cut out by the

computer in a fraglet. It is an intermediate position between full allographic comparison table and sub-allographic boxes that represent strokes and corners.

For writer identification, we are using both textural and allographic features.

4) Probability Curve FAA – FRR

To understand what the distances from feature vectors mean between the manuscripts, we formed a probability curve. Distance is not an accepted currency, but the probability is. There should be a breakpoint with the distances where we have to say this is not the scribe anymore. The probability tool help with this.

Vinodh Rajan Sampath (Hamburg)

Script Analyzer: A Tool for Quantitative Paleography

There exist various machine learning methods that perform classification of scripts, and even writer identification. Most of these techniques are opaque and the quantifications used seemingly impenetrable. Moreover, any custom descriptive analysis of scripts is hard to perform, as they are more adapted to classification. In this context, Script Analyzer is a digital paleographic system that supports quantitative and descriptive analysis of scripts in a transparent manner. It performs a user-aided recovery of the ductus information and decomposes the characters into strokes to create a hierarchically structured representation. The metrics quantifying the scribal handwriting behavior are then derived from the ductus, stroke structure and overall appearance. The metrics directly quantify scribal behavior and, hence, are script agnostic. This makes them applicable to a wide variety of scripts. The entire process is formulated to be human-interpretable and, the metrics, by their very nature, are descriptive (Rajan, 2016). It is also straightforward to define and derive new metrics as necessary, based on the requirements of a paleographer.

Script Analyzer (currently) exists as a Desktop tool running on Python. It requires manual digitization of the exemplar characters into a mathematical spline-based representation for further processing. The subsequent process chain leading to metrics extraction is done in a user-guided semi-automatic manner. The metrics of a character collection is then exported as a CSV file for further analysis and visualization. The tool also allows visualization of the reconstructed ductus and the stroke structure, thereby providing paleographers with visual feedback about the internal processing of the tool. The initial digitization process needs to be done manually and is quite resource intensive. We are considering improving the usability aspects of the tool in the future and porting it into a web-based interface to increase its accessibility.

Timo Korkiakangas (Helsinki)

Quantifying Medieval Latin handwriting with Script Analyzer

My paper will utilize the Script Analyzer tool to quantify early medieval Italian documentary scribes' handwriting and to compare it to the Latin they wrote.

The data is a sample of 50 original documents written in Tuscany between AD 738 and 850. The images come from the Chartae Latinae Antiquiores volumes and the text from the Late Latin Charter Treebank (LLCT). LLCT contains 1,040 scribal documents (480,000 words) written by 220 scribes in Tuscany

in AD 714-897. The documentary Latin displays much variation from writer to writer and differs from Classical Latin in terms of spelling, lexicon, morphology, and syntax.

I will test the hypothesis that the scribes' competences in handwriting and in mastering the Latin language are positively correlated, given that both are important components of a successful scribal performance. I decided to measure the scribes' mastery of Latin simply by way of their spelling. I defined a spelling mistake variable which is operationalized as the percentage of non-standard characters in all the characters of a certain document (Korkiakangas, 2017). Equally simplistically, I decided to examine only one aspect of the scribes' mastery in handwriting, namely the consistency within the instances of the letters <e o p t> in each document. This consistency is measured by the variation coefficients of the metrics given by Script Analyzer for those letters. The coefficients are then compared to the spelling error rate of that same document. To maximize the variation within the sample, I only examined the 25 worst spelled and the 25 best spelled documents.

The preliminary results seem to verify the hypothesis about a positive correlation between handwriting and spelling, suggesting thus that those scribes who knew Latin spelling also knew their job in terms of writing more consistent handwriting than those who made spelling mistakes. However, this result may be at least partly explained away by chronological change in handwriting practices, as the chancery-style cursive (cancelleresca) gains in popularity at the expense of the traditional ordinary cursive.

Elena Nieddu, Serena Ammirati *in abs.* (Rome)

IN CODICE RATIO: a gateway to paleographical thesauri

We are a research group born from the collaboration between computer engineers, paleographers, archivists and historians of Roma Tre University and the Vatican Secret Archives. At the Basel conference we (represented by Elena Nieddu) would like to present our project - In codice ratio (http://www.inf.uniroma3.it/db/icr/) - and its first encouraging results, in the hope that the methodologies applied so far can be useful and interesting in the field of papyrology.

In Codice Ratio is a research project that aims at developing novel methods and tools to support palaeographical analysis and knowledge discovery from large collections of historical documents. The goal is to provide humanities scholars with novel tools to conduct studies over large historical sources.

The project has so far concentrated on medieval papal correspondence registers of the XIII c., but its methodology could hopefully be extended to documents of diverse nature and age. We are developing a full-fledged system to automatically transcribe the contents of the manuscripts. We follow a novel approach, based on character segmentation and minimal training effort. Our idea is to govern imprecise character segmentation by considering that correct segments are those that give rise to a sequence of characters that more likely compose a Latin word. We have designed a principled solution that relies on convolutional neural networks and statistical language models. Preliminary results, also tested on older (X-XII c.) and later (XIV c.) documents, are encouraging. The entire process allows a collection of palaeographical specimina (a sort of 'thesaurus' of graphical variants not only of letters, but also of abbreviation signs, ligatures and so on), which represents an important tool for large-scale paleographical analysis. Once fully developed and appropriately customized to different scripts and languages, we believe that our tool could work for different types of manuscripts, including papyri.

Peter Stokes (Paris)

(Still) Describing Handwriting: With Archetype and Beyond

Digital approaches to palaeography have come a long way in the last ten to fifteen years, with significant advances in many areas of automatic and computer-aided document analysis applied to historical content. One area in this domain that has perhaps received less attention has been explicit or formal models for describing handwriting. The need for such models has been recognised for some time, for instance in an ESF Workshop and Dagstuhl Seminar held in 2011 and 2012, respectively. Some work has been done on this, such as the Archetype model first developed in 2011 and since applied to a relatively wide range of scripts and research projects including Greek, Latin, and Hebrew, as well as experiments in Chinese, Arabic, Mayan and Old Khmer, among others. Although functional and pragmatic, the model (like all models) is of course limited and could usefully be extended. Challenges that have already been discussed include application to cursive script and/or the inclusion of movement of the hand and writing instrument, as well as problems representing multigraphic contexts in which different scripts are found in the same document or corpus. At a more basic level, however, definitions of fundamental concepts seem still to be lacking not only in Archetype but in discussion more widely. Even core terms such as 'letter' lack precise definitions that function across all scripts, let alone more complex concepts such as 'grapheme' or 'alphabet', and palaeographers have debated terms such as 'cursivity' and 'script' without clear conclusions. However, any digital approach depends on a clear model for representing the subject-matter, as well as for communicating it to people across a range of disciplines (palaeography and informatics, for a start). This paper will therefore draw on several case studies to present these problems and the beginnings of some possible solutions.

Simona Stoyanova (Nottingham)

The Python in the letterbox – epigraphic palaeography with Archetype

Archetype, built on its predecessor DigiPal, is an open-source web-based suite of tools for the study of handwriting, palaeographical features and iconography. Although initially designed for medieval handwriting, it is highly generalised and allows the study of various alphabets on various materials (Old English, Hebrew, Latin, Greek), as well as decoration. It is based on annotating images with highly structured descriptions of letterforms.

This paper will discuss my approach to Archetype's customisation for the study of epigraphic palaeography. In the field of epigraphy dedicated palaeographic research has been done only on material from Athens and Rome – both places with long uninterrupted writing tradition. My research focuses on Greek, Latin and bilingual inscriptions from the province of Thrace where the population navigated multilingualism, complex literacies and identities.

Working on lapidary material and two languages requires changes to the Archetype data model. I focus on studying how different scripts potentially influence one another, rather than identifying scribal hands. In epigraphy we very rarely have a second copy of the same text for comparison, so the project does not rely on multiple witnesses. The complex material from a multilingual province allows the exploration and comparison of two languages and two alphabets, with Greek being the domineering one. Influence can be categorised as fashion, statement, status symbol.

I would like to show and put to discussion some of the Archetype customisations I have made so far. They include changes in the terminology for categorising letterforms, to reflect material and disciplinary expectations; changes to metadata formats, e.g. dating and provenance; changes to the interface and search functionality to help organise the material and search results. I am hoping to prompt a conversation about the various types of digital aid in palaeographic research, as well as receive thoughts, suggestions and feedback on my use case.

Lorenzo Sardone (San Marino)

For a Paleography of Demosthenic Papyri

The great number of Demosthenic papyri found in Egypt is a clear sign of this author's success throughout the Roman period. In fact, Demosthenes is one of the most well-known classical authors on papyri, second only to Homer. Among his large *corpus* is *On the Crown*, a true masterpiece in Antiquity, which is the most commonly witnessed, with 30 papyri containing passages from this work. The enquiry into these *specimina* and their lessons could be useful in order to reconstruct the Demosthenic text, as well as the origin and purpose of the public documents quoted within the text. In addition, it could be possible to understand the relations with medieval manuscripts. In recent years, scholars are giving fresh intention to the "material philology", in order to investigate the circulation of this speech and the real physical features of Demosthenes' ancient editions. The main *focus* of this approach is the palaeographical enquiry, with hands identification, scripts classification and the investigation of the quality and function of each item. Fundamental studies in palaeographical and material philology on literary papyri are: W. LAMEERE, *Aperçus de paléographie homérique, à propos des papyrus de l'Iliade et de l'Odyssée des collections de Gand, de Bruxelles et de Louvain*, Paris-Bruxelles 1960; G. CAVALLO, *La papirologia letteraria tra bibliologia e paleografia: un consuntivo del passato e uno sguardo verso il futuro*, in T. DERDA – A. LAJTAR - J. URBANIK - (eds.), *Proceedings of the 27th International Congress of Papyrology*, vol. I (JJP 43, 2013), Warsaw 2015, pp. 277-312; G. CAVALLO – L. DEL CORSO, *1960-2011: mezzo secolo dopo gli Aperçus de paléographie homérique di William Lameere*, in G. BASTIANINI – A. CASANOVA (edd.), *I papiri omerici. Atti del Convegno Internazionale di Studi. Firenze 9-10 giugno 2011*, Firenze 2012, pp. 29-63. After Homeric papyri, it's clear that Demosthenes' fragments can provide us with a perfect basis of enquiry, with more than 200 items, both on papyrus and parchment, parts of ancient *volumina* or *codices*, that show different majuscule, from the most formal ones, to the more cursives.

Yasmine Amory (Ghent)

More than a simple intuition. Towards a categorisation of palaeographical features

In documentary papyrology, paleography still lacks of a proper and defined terminology. Various reasons can be attributed to it: one is the general conception of paleography as an auxiliary tool mainly used to date unpublished documents, another one is the common preconception that paleography is no more than an intuition or an art (see the well-known assertion by W. Schubart in *Palaeographie. Erster teil, Griechische Palaeographie*, 1925, Munich, p. 11). Despite some attempts to fight this long-lasting trend, papyrological editions do not usually contain information on the handwriting, and, when they do, these indications are inconsistent and usually vague. This paper aims to discuss and introduce a possible categorization of paleographical features in documentary papyri. The attempt is to identify objective criteria of what has so far been considered a mere matter of subjectivity and sensitivity to the text.

Loreleï Vanderheyden (Heidelberg)

How to unmask a digraph scribe? Apollos' Greek and Coptic styles in the Aphrodito Byzantine Archive

When we speak about sixth century Egyptian documentation, what we are speaking about is the product of Greek-Coptic contact and interference for more than two centuries: Greek was restrained to administrative and prestige contexts while Coptic was used as the vernacular and daily-life basis for most private documentation. It was common for lettered persons and the elite to write in the two languages, and to easily shift from one language to the other.

Since they share the same alphabet, one can find common patterns, ligatures and/or scribal habits in the documents written in the two languages by the same scribe. Because of its localization and of its precisely dated and identified Greek texts, the Aphrodito papyri from the Dioscorus archive offer a unique body of material for the study of 6th century Coptic digraphism and writing, using Greek texts to identify an unknown Coptic hand or using a signed Coptic letter to give a name to an unidentified Greek scribe.

It gives us the opportunity to find out what we can learn from the case study of Apollos, father of Dioscorus. Considered illiterate by scholars, I will try to show that, thanks to a paleographic methodology he can now be considered as an experienced digraph scribe, thanks to documents where his hand had not been recognized so far.

Anne Boud'hors (Paris)

Identifying hands and styles in the Coptic papyri from Edfu (Papas' archive)

The so-called "archive of Papas", pagarch of Edfu (Upper Egypt) around 660-680, is a collection of several hundreds of Greek and Coptic papyri found in a jar during excavations at Tell Edfu in the 1920's. Apparently during the transportation of the jar to the Institut français d'archéologie orientale in Cairo, the papyri were all so severely damaged that not a single one is complete, and for many of them only a few fragments escaped the destruction. The Coptic documents, still unpublished, by contrast with the Greek ones (published as soon as 1953 by R. Rémondon), are under current study by an international team. Most of them are letters addressed to Papas, either administrative or private. We are facing two interrelated challenges: 1) reconstructing the documents; 2) identifying hands, or at least styles, which is particularly difficult in the case of the administrative letters, as the scripts are very standardized. We are using different approaches, namely general aspect of the script, shape of the letters, space between the lines, colour of the ink and of the papyrus, fiber continuity, as well as rhetorical formulas. I would like to show some of our results and the process to reach them, which is long and empirical, and could perhaps be improved with the help of digital tools.

Esther Garel (Strasbourg)

The Fayyumic Coptic Documentary Papyri: Issues of Palaeography, Formats and Dating

The Coptic Papyri from the Fayyum were acquired by several European collections at the end of the nineteenth century, together with Greek and Arabic documents. Scattered between Vienna, Berlin, Paris or London, they have been understudied so far, due to the difficulty of the dialect in which they are

written. Therefore it has not been possible to identify a coherent ensemble of texts that could constitute a dossier or an archive around a place or a group of people. Striking palaeographical features can however be observed between some of the documents, as well as specific formats, that could lead to the reconstruction of notarial offices. These issues are also related to the dating of the documents, none of which has been dated before the beginning of the 8th century CE, although Fayyumic literary dialect is among the oldest attested. The paper aims at identifying and describing some of the styles of writing and trying to establish a typology.

Christian Askeland (Cambridge)

How to clean up a Papyrustastrophe? Using empirical data and common sense to reconnect shattered fragments

The organization of a shattered collection of papyri begins with organization, sorting fragments by characteristics of ink, writing material, fiber direction and script and ends with reconstruction of those fragments formerly belonging to the same manuscript. The present paper describes the process, considering how digital innovation might supplement or replace traditional methods, concluding with a survey of the Alexandrian Majuscule and its influence on the Biblical Majuscule during the Islamic era.

Katharina Schröder (Münster)

Searching for Relatives: Palaeographical Analysis of Coptic New Testament Manuscripts in the Institute for New Testament Textual Research Münster

The project Editio Critica Maior (http://egora.uni-muenster.de/intf/aecm/aecm_en.shtml) is based in Münster since 2007, working on a new edition of the Greek New Testament that takes into consideration the entire primary tradition. For relevant passages of any given verse, the translations of the New Testament into Latin, Syriac, Ethiopic, Gothic and Coptic are incorporated into the critical apparatus. The organization of Coptic manuscripts hence has had its place in this project from the beginning, taking part in the publication of the Catholic Epistles in 2013, of the Acta Apostolorum in 2017 and currently preparing the Gospel of Mark.

The proposed paper will present the SMR-database of Coptic New Testament Manuscripts (http://intf.uni-muenster.de/smr/). The database aims at holding information on all known Coptic New Testament manuscripts available. It is systematically enlarged with new manuscripts and fragments, according to the book the ECM has in current preparation but also including all other pieces that emerge.

Coptic codices are mostly very fragmentary, their pieces scattered around the collections and museums around the world. Therefore, any new fragment that is to be added to the database is checked for relatives: It could belong to a manuscript that is already registered and, in case that proves true, must not be numbered as an individual item but should be enlisted with the registered manuscript. The paper explains the digital tools used for the process of sorting manuscripts according to metadata and the following palaeographical comparison of the new piece with possible candidates. Case studies will be presented showing where two manuscripts were joined together due to palaeographical characteristics, but also difficult examples for which it was decided against a merging. The method has been successful multiple times since it was developed, so it was possible to digitally join pieces of manuscripts that have been apart for centuries.

Alin Suciu, Ulrich Schmid *in abs*. (Göttingen)

Digital Support for a Paleographical Assessment of the White Monastery Manuscripts

As most of the available Coptic manuscripts have survived in a fragmentary state, their study is notoriously difficult. This is especially true of the Sahidic codices which belonged to the White Monastery, situated in Upper Egypt near modern-day Sohag. Remnants of over 500 manuscripts from this locale have been identified until now. Their reconstruction requires not only expertise in paleography and codicology, but also a good command of the biblical, liturgical, and literary texts that circulated in Coptic. Our paper focuses on the paleographical and codicological methods used in the reconstruction of the White Monastery manuscripts. We will survey the paleographical grounds for dating White Monastery manuscripts and explain the different classes of script attested, advocating for the necessity to group together the manuscripts in scribal corpora. Identifying the codices inscribed by the same copyists will greatly facilitate to establish the age of those manuscripts which are otherwise difficult to date.

To this end, our Virtual Research Environment (VMR) will aid researchers by allowing them to easily create any number of features for tagging either entire manuscripts, individual pages or specific surrogates. It also allows the researcher to associate clips from manuscript images along with the tags that help to create a visual inventory of the tagged items for easy comparison.

Marie Beurton-Aimar, Cecilia Ostertag *in abs.* (Bordeaux)

Re-assembly Egyptian potteries with handwritten texts

In recent years, machine learning and deep learning approaches such as artificial neural networks have gained in popularity for the resolution of automatic puzzle resolution problems. Indeed, these methods are able to extract high-level representations from images, and then can be trained to separate matching image pieces from non-matching ones. These applications have many similarities to the problem of ancient artifacts reassembly from partially recovered fragments. In this work we present a work in progress using deep learning and graph construction to propose possible reconstructions from ostraca fragments. First we designed a siamese neural network to evaluate pairwise matching of image patches, and predict their respective positions for the assembly. This model was tested on 100 unknown patches and gave an accuracy of 81%. Then we built a graph-based pipeline to reassemble a whole image from multiple fragments through iterative pairwise matching.

Vincent Christlein (Nuremberg)

Writer identification in historical document images

In the age of mass digitization of historical documents, automatic or semi-automatic procedures for processing and analysing them have a high priority. For example, an automatic image retrieval system based on an image in question may help humanities scholars to retrieve similar images. In contrast to general image retrieval, where the full image is of importance, in writer identification, we commonly focus only on the handwriting. While writer identification/retrieval in clean benchmark data is very successful, reaching identification rates beyond 95 percent even in multi-script scenarios, it becomes increasingly more difficult in historical data. Historical document images often suffer from artifacts, such as rips, holes or water sparkles. To limit their effect, robust methods are needed. We investigate different methods on different algorithmic stages how to improve writer identification for historical documents and show evaluation results of different datasets, such as a dataset of letters and a dataset of papyrus fragments.

Imran Siddiqi (Islamabad)

Dating of Historical Manuscripts using Image Analysis & Deep Learning Techniques

The last two decades have witnessed a tremendous increase in the digitization of historical manuscripts. This not only allows preserving the priceless manuscripts in digital libraries but also offers substantial research attraction to the pattern recognition community in general and, document and handwriting recognition community in particular. Digitized (scanned or photographed) collections of ancient manuscripts offer a number of interesting research problems for computerized investigations. These include pre-processing tasks (noise removal, document restoration etc.), segmentation tasks (binarization, line and word extraction), handwriting recognition, identification and authenticity of

scribes and predictions on the origin of the document etc. While pre-processing and recognition tasks have been extensively explored, dating of historical manuscripts using image analysis and machine learning techniques remains a relatively less explored area and makes the subject of our present investigation. Features capturing the textural information of writing strokes are computed to characterize the date of a manuscript. A supervised learning approach is employed where features extracted from manuscripts (with known dates of origin) are used to train a classifier. In addition, we also investigate automatic feature learning from the given samples using deep neural networks. More specifically, we employ transfer learning on a number of popular pre-trained Convolutional Neural Network (CNN) models to characterize the date. Experimental study is carried out on the Medieval Paleographical Scale (MPS) dataset and the realized results show significant reduction in the Mean Absolute Error (MAE).

Tanmoy Mondal (Montpellier)

Efficient technique for Binarization, Noise Cleaning and Convolutional Neural Network Based Writer Identification for Papyri Manuscripts

In this work, a novel local threshold binarization method using fast Fuzzy C-Means clustering is proposed. Historical document images with non-uniform background, stains, faded ink are first processed by removing the background using inpainting based method. Then using Fuzzy C-Means clustering is used to cluster out the pixels into three main clusters: sure text pixels, sure background pixels and confused pixels which may or may not be labeled as text. Based on the structural symmetry of pixels (SSP), these confused pixels are then classified into text or background pixels. The SSP is defined as those pixels around strokes, whose gradient magnitudes are big enough and whose directions are opposite. As the gradient map is our basis for computing the SSP, we further propose to estimate the background surface first and to extract potential SSP in the compensated image so as to deal with degradations of document images such as uneven illumination, low contrast and stain.

The next task is to do writer identification. For writer identification, we try to obtain as clean image as possible by removing the noisy background of papyri images. We apply the same impainting based method for background extraction, separately RGB channels of the color image. Then anisotropic diffusion filter is applied on these three background images seperately to smooth and clean noises from these images. Later these three images are merged to generate color image which is followed by modifying those pixel values to zero, which are text pixels (foreground) in binary image. This operation will give us a color image with clear text foreground and noise cleaned (blurred) background.

Then we present a simple framework based on Convolutional Neural Networks (CNNs), where a CNN is trained to classify small patches of text into predefined writer classes. At prediction time, to classify full page or part of it, we average the CNN predictions over densely extracted patches and assign an average predictions over individual patch prediction.

Andreas Fischer (Fribourg)

Recent Advances in Graph-Based Keyword Spotting for Supporting Quantitative Paleography

Keyword spotting aims to automatically retrieve the same character or word from a large collection of scanned manuscripts, which can be very helpful for quantitative paleography.

Most keyword spotting approaches are based on statistical methods, which capture properties of the handwriting in form of a fixed amount of real-valued numbers, e.g. geometrical characteristics, histograms of oriented gradients, or features extracted by deep convolutional neural networks.

Structural methods, on the other hand, offer more flexibility representing handwriting. When using graphs, a variable number of nodes describe parts of the handwriting and a variable number of edges describe relationships among these parts. Both nodes and edges can be labeled with symbols, real-valued numbers, or any other descriptor.

However, the increased representational power of graphs comes at the expense of high computational complexity when comparing two graphs. Graph matching aims to detect similarities and dissimilarities with respect to the graph structure and labels, and thus is the basis for performing keyword spotting.

In this talk, we provide an insight into recent advances in graph-based keyword spotting, which has become feasible in the past decade due to the availability of fast, approximative algorithms for graph matching. We discuss graph-based representations of handwriting, highlight their potential for paleographic studies, present fast matching algorithms for detecting similar graphs in large manuscript collections, and show results of experimental evaluations on benchmark datasets that demonstrate the high potential of structural methods for keyword spotting.

Vlad Atanasiu, Peter Fornaro (Basel)

On the utility of color in computational paleography

Paleography being essentially the study of shapes, it is only natural in both computer science and paleography to disregard color as superfluous. This workshop aims to demonstrate that color can provide information useful in numerous tasks, from reading to computational analysis. It also aims to sensitize participants to the fact that once color images are used, human reading and computational processing are influenced by various aspects related to color science even if users are unaware of it. Beyond the utility of color, the topics covered are basic imaging best practices, interactive image enhancement, and a pipeline of computational processing in color space.

Notes