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 (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]

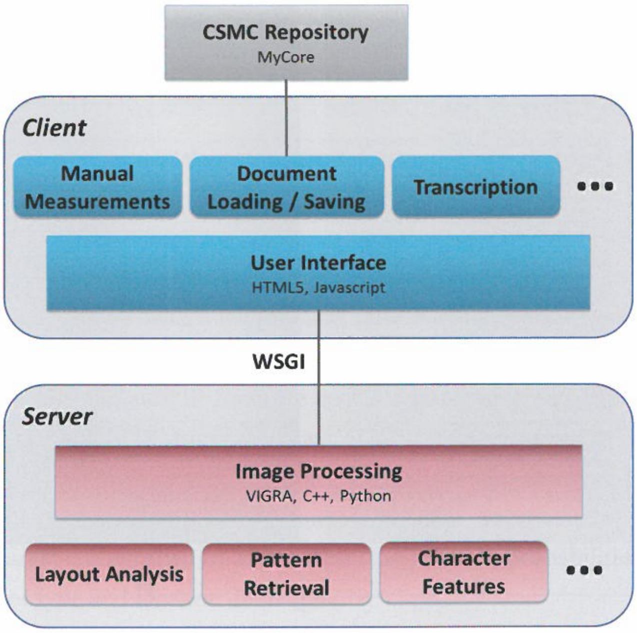

Fig 1: AMAP's system components: a modular approach.

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.

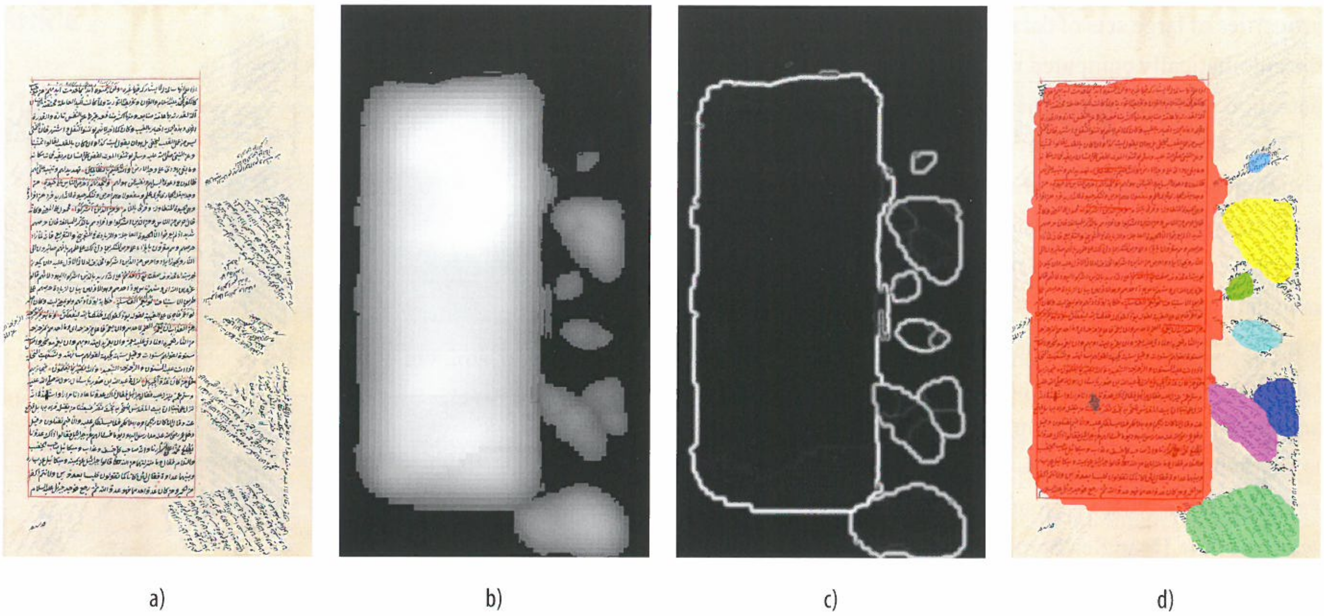

Fig. 2a: University Library of Leipzig, Ms B. or. 002 / p. 84, 2b: magnitude of detected frequencies above the threshold, 2c: boundaries of homogeneous frequency sections, 2d: false-colour representation of determined homogeneous text blocks.

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.

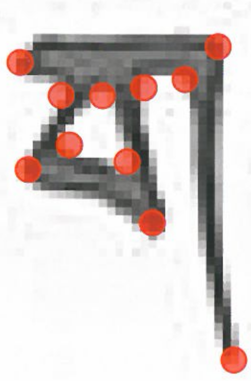

Fig. 3a: Tibetan character with IPs.

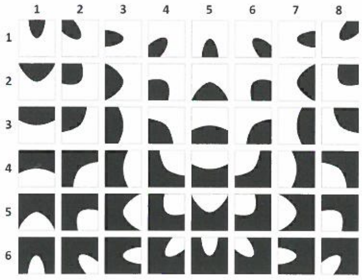

Fig. 3b: Code chart of IP types.

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.

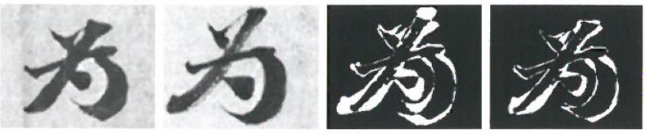

Fig. 4: Two examples of the same Chinese character (left); the white pixels represent non-matching regions before and after warping (right).

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.