Automatic Algorithms for Medieval Manuscript Analysis

Ruggero Pintus [a] , Ying Yang [b] , Holly Rushmeier [c] and Enrico Gobbetti [a]

[a] *CRS4, Visual Computing Group, Italy*

[b] *Fujian Provincial Key Laboratory of Information Processing and Intelligent Control*
Minjiang University, Fuzhou, China

[c] *Graphics Lab, Department of Computer Science, Yale University, USA*

Abstract. Massive digital acquisition and preservation of deteriorating historical and artistic documents is of particular importance due to their value and fragile condition. The study and browsing of such digital libraries is invaluable for scholars in the Cultural Heritage field, but requires automatic tools for analyzing and indexing these datasets. We will describe a set of completely automatic solutions to estimate per-page text leading, to extract text lines, blocks and other layout elements, and to perform query-by-example word-spotting on medieval manuscripts. Those techniques have been evaluated on a huge heterogeneous corpus of illuminated medieval manuscripts of different writing styles, languages, image resolutions, amount of illumination and ornamentation, and levels of conservation, with various problematic issues such as holes, spots, ink bleed-through, ornamentation, and background noise. We also present a quantitative analysis to better assess the quality of the proposed algorithms. By not requiring any human intervention to produce a large amount of annotated training data, the developed methods provide Computer Vision researchers and Cultural Heritage practitioners with a compact and efficient system for document analysis.

Keywords: Semantic Feature Extraction, Medieval Manuscript, Document Layout Analysis, Word spotting, Multi-spectral analysis

1 Introduction

Centuries of humanities research have generated a huge amount of data relevant to humanistic inquiry. Much of this research requires the study of cultural heritage (CH) items, which includes the analysis of objects with very different shapes, materials, ages and conservation statuses. Each group of them (or even, in some cases, each single item) has its own environmental conditions and preservation procedures. This makes their monitoring and restoration hard and challenging tasks to address. The vast majority of such research has been historically carried out manually by direct visual analysis. In the last decades, the rapid spread of digital technologies and large-scale multi-modal digitization efforts have led to new ways of doing scholarship. Digital tools enable automatic semantic feature extraction, as well as fast worldwide access to study CH material. This yields new insights through faster quantitative and repeatable analyses [1, 2, 3], and improves cross-disciplinary collaboration [4], as well as validation and dissemination of results.

Among the large variety of CH items, historical handwritten texts represent a vast invaluable corpus, crucial for the creation and understanding of human culture and identity. Data-driven research on digital collections of texts requires the ability to flexibly and efficiently perform data-intensive operations on a massive corpus of digital elements, including mono- and multi-modal digitization (e.g., from simple digital colour copies to multispectral/multilayer images revealing materials or invisible/hardly visible information). Data processing is necessary for the generation of better/smarter digitally-born representations of cultural documents, as well as for the direct support to quantitative humanistic inquiries at a large scale.

Here we summarize our recent research on tools for automatic analysis of medieval manuscripts. We present the developed techniques to automatically extract document layout elements (section 2), such as main text leading, capital letters or other special components (i.e., semantically meaningful elements that are not strictly plain text), text blocks, text lines. We show how to use a part of this information to facilitate and speed-up query-based word spotting on handwritten documents (section 3). Finally, we describe multi-modal analysis techniques (section 4), which exploit multi-spectral signals to perform color-based reasoning to understand document structure and label its elements accordingly.

2 Document Layout Analysis

Recovering the geometric and logical layout of digitized historical documents is a prerequisite to associate semantics to their content, as well as to index and browse them within large databases in terms of their textual or pictorial content.

A common problem in the document structure analysis is the initial tuning of algorithm parameters, which are typically related to an estimation of the text scale. In particular, computing the text leading can provide an automatic guess on those values, which will guide the rest of the analysis pipeline. Due to their clear inter-line spacing, it is easy to find this parameter for printed books/pages, while it becomes challenging when descenders and ascenders overlaps or when the spacing is narrow (e.g., in medieval manuscripts). We propose a text leading extraction approach [13] that works on a per-page basis. It takes a manuscript image with a quasi-horizontal text; if this is not the case, we can add one of the state-of-the-art pre-processing steps for page alignment correction [9]. An N-level multi-scale representation is produced, where at level n, we split each original image in 2^{2n} small sub-images. Then, we analyze these levels separately. For each of their sub-images we compute the normalized autocorrelation function (NACF), and we integrate this signal to obtain its y-axis projection profile (y_{pp}). We find the main periodicity of the y_{pp} by applying the Discrete Fourier Transform (DFT), and we use the information corresponding to the highest DFT coefficient from all sub-images to compute, for that particular level n, an estimation of the text leading in terms of a probability mass function (PMF). Finally, we exploit the coherence between levels to find the final estimation of the page text leading, by accumulating all the PMFs from all levels.

We then use the knowledge on per-page text leading computation to automatize the extraction of other layout elements of a manuscript, e.g., text blocks and text lines. We propose two approaches. One is based on local feature extraction [11] (e.g., SIFT [10]), while the other is based on image binarization coupled with connected component analysis and template matching [18].

In the first case [11], the algorithm is given an entire book as a set of images. As a pre-processing step, for each page we compute the average text leading [12, 13], and the SIFT features [10]. By exploiting the information of the text leading distribution across the book and the hue histogram of each image, we select the most salient pages, and the corresponding local image descriptors. With a modification of the frequency based approach presented before [13], we compute a coarse two-class segmentation of this subset, and we assign to each feature one of the following labels: *text* or *non-text*. We use this rough classification to automatically train a Support Vector Machine (SVM) with Radial Basis Function (RBF), and we then re-launch a prediction step to all original SIFTs to obtain a fine text keypoints segmentation. Since the predicted points across the entire manuscript are a sparse representation of text positions, we employ a kdtree data structure to estimate the influence radius of each point, and we extract dense text regions and the associated rectangular text blocks. Finally, for each block, we compute the projection profile by integrating only the contribution of the text keypoints. The size of profile bins, the algorithm to extract the maxima of the profile, and the final segmentation of each single line, all are strongly dependent of the previously computed text leading value.

While the aforementioned solution work with an entire input book, conversely, the second approach [18] exploits the text leading computation to automatically set the parameters for the extraction of text blocks, text lines and other layout components on a per-page basis. This method relies on a completely and orthogonal rationale; unlike feature based techniques (e.g., SIFT-based) that work on grayscale images, this approach starts by computing a binary version of the page. By analyzing the projection profile of the binary signal, it produces a rough text block segmentation. Then it refines the text block using an operator that assumes a rectangularly shaped text (this assumption is generally true for old English and Latin manuscripts) and an image template matching strategy [19]. Then, the text line extraction is pretty straightforward; projection profile techniques are applied to each block rather than to the whole image. In this case, spotting other layout components as figures and capital letters is based on the assumption that their color and shape is different from text pixels; it is then possible to compute some feature vectors based on the color and the 2D geometry, and formulate the entire problem as a clustering task. Those features, presented by Yang et al. [19], are composed of 60 components associated with color and statistical characteristics, such as the mean, standard deviation, skewness, energy and entropy of the color information.

We applied the proposed methods to medieval books provided in the database of the Yale Digital Collection Center [7] (a set of scripts is available [6] to download a subset of the book database), the Oxford University's Bodleian Library, the Florence's Biblioteca Nazionale Centrale, the Walters Art Museum, the Admont's Stiftsbibliothek, the Köln's Erzbischöfliche Diözesan- und Dombibliothek, the Ripoll's Biblioteca Lambert Mata, the St. Gallen's Stiftsbibliothek, and the London's Wellcome Library. The text leading extraction has been tested on 34 Medieval manuscripts and 19 Arabic handwritten books (15552 pages), while the text block/line extraction has been evaluated on 8 medieval handwritten books, with 2724 pages, 3558 text blocks and 65697 lines. The data are very heterogeneous, in terms of layout structure (e.g., number of columns and text density), conservation (e.g., aging, ink bleed-through and noise), resolution, and writing styles. We obtain a very high rate of good estimations of the per-page text leading, i.e., 98% of good pages. For text region segmentation and text block extraction we measured *Precision*/*Recall* values over 90%. The text line extraction exhibits a Recall value of about 98%, and a lower *Precision* of 70%.

3 Word spotting

We present a completely automatic and scalable framework to perform query-by-example word-spotting on medieval manuscripts [14]. The techniques presented in section 2 are used here to feed the algorithm with the automatically segmented text lines. The main idea is to get rid of all non-text data, and to reduce the system in-core memory footprint, making the framework, designed to be compliant with the standards of document layout analysis web services, scalable to very large image databases. We have executed this pipeline not only on a single-manuscript, but also in a cross-manuscript scenario, and its high success rate has been proven with a heterogeneous set of medieval manuscripts, that includes a variety of writing styles, languages, image resolutions, levels of conservation, noise and amount of illumination and ornamentation.

Given an entire book as a set of images, and a sub-region from one of those pages that contains the user selected query word, we first automatically compute the average text leading of the book [13], then we apply an automatic segmentation step to extract text lines from the input manuscript [11]. We then scale query and line images to a canonical size. In a decolorization step we convert them to a high contrast monochromatic signal, and we compute their compressed HOG descriptors. Then, for each query word, the word-spotting algorithm (a modification of the approach of Almazán et al. [5]) takes those grayscale images and returns a ranked sequence of lines, which contain the query word, together with the region of the line surrounding that word. The automatically computed text leading value is used to adaptively set parameters in all the steps related to the text line extraction, scaling, decolorization and word-spotting.

The word-spotting pipeline has been evaluated on the handwritten George Washington (GW) dataset [15, 16], the Lord Byron (LB) dataset [17], and on a set of medieval books from the Yale Digital Collection Center [7] (about 270 pages and 5000 lines). Since the pipeline returns a ranked list of word occurrences, in order to take into account this ranking, we measure the performance by using the *Average Precision* (AP) metrics. Resulting AP statistics ranges from 52% to 85%, which is comparable to, and even better than, the state-of-the-art methods (see for instance statistics in Almazán et al. [5]), which have AP values ranging from 30% to 60% for the handwritten GW documents, and, more important, from 40% to 80% for the machine printed LB dataset. Moreover, our algorithm is capable of doing a crossmanuscript search with good AP scores.

4 Color analysis

When multi-modal signals are available for the digital documentation of a handwritten medieval book, a series of specific analyses arises. These investigations are important for the study of material properties across the corpus, to compute the variety of used colors, to infer some clues about pigment distribution, and to try to solve ink-related issues. Such multi-modal based automatic procedures to process massive numbers of historical documents might be useful to reveal possible pigment changes, which indicate changes in scribal hand, to identify initials and differentiate between types of them, and to derive page classes (e.g., full-page drawings, calendar pages, litanies).

Our recent technique [19] is capable of automatically finding the number of multi-modal clusters used for *non-text* regions in a book; it uses both RGB or multi-spectral signals. It is based on the observation that *text* pixels in a manuscript page generally have the same color (perceptually), but are colored distinctively from *non-text* ones. The algorithm also relies on a text segmentation routine that exploits the particular and repetitive shape of text strokes, a specific characteristic of medieval writings, together with a template matching strategy. The parameter estimation is again based on the initial computation of the per-page text leading [13]. Given a set of candidate pixels from the previous segmentation/classification steps, for each of them a series of N-dimensional features is computed. For RGB images, the first nine elements of the feature vector are the Hue, Saturation, and Value of the pixel, followed by the six RGB and normalized RGB components. Then, other 45 elements are appended that represent the multi-modal statistics in neighbor regions of the analyzed pixel taken at various different size; those statistical measurements includes the mean, the standard deviation, the skewness, the energy and the entropy. The final six components are related to the statistical distribution of zero-valued pixels in the segmented binary image. In the case of multi-spectral values, RGB will be computed as a chromatic dimensionality reduction, while the multi-spectral values will be appended at the end of the feature vector. The feature vectors are the input of a unsupervised clustering approach who finds the best solution by minimizing the number of cluster, their overlap, and maximizing a separation metrics. A Davies-Bouldin criterion [8] has been used to meet the requirement of a good clustering.

The proposed method has been evaluated on the images from the Yale Digital Collection Center [7]. This dataset consists in 1027 RGB and 72 8-band multi-spectral images of pages from 7 different manuscripts, as well as their 1099 variants produced by modifying their spatial and spectral resolutions. The test results show that the proposed algorithm is capable of finding expected number of color clusters with a high likelihood (up to 70%). Moreover, the application of color clustering to element segmentation (e.g., capital letters, line fillers) produced retrieval statistics with both *Precision* and *Recall* higher than 90%.

5 Conclusions

We have presented a series of tools to allow scholars to compute a digital representation of the structure of both entire medieval manuscripts or single pages. The computational framework is automatic, being completely driven in an adaptive way through pre-processing steps that calculate the initial values of the algorithm parameters. By not requiring any human intervention to produce the large amount of annotated training data, developed methods provide Computer Vision researchers and Cultural Heritage practitioners with a compact and efficient system for document analysis. The outcomes are some page or book attributes that relate to layout (text leading, blocks, lines), special book components (capital letters, pictorial elements, line fillers), query-based word classification, and color or multi-spectral clusters. We have evaluated all the aforementioned techniques on a huge heterogeneous corpus of illuminated medieval manuscripts [13, 11, 14, 19, 18], showing how they exhibit a high success across different writing styles, languages, image resolutions, amount of illumination and ornamentation, and levels of conservation, with various problematic issues such as holes, spots, ink bleed-through, ornamentation, and background noise.

Acknowledgments. This work was partially supported by a Yale University and CRS4 research agreement, as well as by the Scan4Reco project funded by European Union's Horizon 2020 Framework Programme for Research and Innovation under grant agreement no 665091. We also acknowledge the contribution of Sardinian Regional Authorities.

References

[1] Yale University, DESMM project. `digitalhumanities.yale.edu/projects/desmm`, 2013.

[2] tranScriptorium project. `transcriptorium.eu`, 2015.

[3] READ project (Recognition and Enrichment of Archival Documents). `www.caa.tuwien.ac.at/cvl/project/read`, 2017.

[4] T-PEN, Tool for digital humanities. `t-pen.org/TPEN`, 2017.

[5] J. Almazán, A. Gordo, A. Fornés, and E. Valveny. Segmentation-free word spotting with exemplar svms. *Pattern Recognition*, 47(12):3967–3978, 2014.

[6] Beinecke. 21 book database - download scripts - http://hdl.handle.net/10079/cz8w9v8, 2013.

[7] Beinecke. Beinecke rare book and manuscript library - http://beinecke.library.yale.edu/, 2013.

[8] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[9] N. Journet, R. Mullot, J.-Y. Ramel, and V. Eglin. Ancient printed documents indexation: a new approach. *Pattern Recognition and Data Mining*, pages 580–589, 2005.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] R. Pintus, Y. Yang, E. Gobbetti, and H. Rushmeier. A TaLISMAN: Automatic text and line segmentation of historical manuscripts. In *The 12th Eurographics Worhshop on Graphics and Cultural Heritage*, pages 35–44, October 2014.

[12] R. Pintus, Y. Yang, and H. Rushmeier. ATHENA: Automatic text height extraction for the analysis of old handwritten manuscripts. In *Digital Heritage International Congress (DigitalHeritage), 2013*, volume 1, pages 605–612. IEEE, 2013.

[13] R. Pintus, Y. Yang, and H. Rushmeier. ATHENA: Automatic text height extraction for the analysis of text lines in old handwritten manuscripts. *ACM Journal on Computing and Cultural Heritage*, 8(1):1:1–1:25, 2015.

[14] R. Pintus, Y. Yang, H. Rushmeier, and E. Gobbetti. An automatic word-spotting framework for medieval manuscripts. In *Proc. Digital Heritage*, pages 5–12, September 2015.

[15] T. M. Rath and R. Manmatha. Word image matching using dynamic time warping. In *CVPR*, volume 2, pages II–521, 2003.

[16] T. M. Rath and R. Manmatha. Word spotting for historical documents. *IJDAR*, 9(2-4):139–152, 2007.

[17] M. Rusinol, D. Aldavert, R. Toledo, and J. Llados. Browsing heterogeneous document collections by a segmentation-free word spotting method. In *ICDAR*, pages 63–67, 2011.

[18] Y. Yang, R. Pintus, E. Gobbetti, and H. Rushmeier. Automatic single page-based algorithms for medieval manuscript analysis. *ACM Journal on Computing and Cultural Heritage*, 10(2):9:1–9:22, 2017.

[19] Y. Yang, R. Pintus, H. Rushmeier, and E. Gobbetti. Automated color clustering for medieval manuscript analysis. In *Proc. Digital Heritage*, pages 101–104, September 2015.