Published in Microchemical Journal 125 (2016) 185-189

Virtual reading of a large ancient handwritten science book

F. Albertin[a], A. Patera[b], I. Jerjen[b,c], S. Hartmann[d], E. Peccenini[e,f], F. Kaplan[g], M. Stampanoni[b,c], R. Kaufmann[d], G. Margaritondo[a]

[a] Faculté des Sciences de Base, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[b] Swiss Light Source, Paul Scherrer Institut (PSI), Villigen, Switzerland

[c] Institute for Biomedical Engineering, ETHZ, Zurich, Switzerland

[d] Center for X-ray Analytics, Swiss Federal Laboratories for Materials Science and Technology (EMPA), Dübendorf, Switzerland

[e] Department of Physics, University of Bologna, Italy

[f] Museo Storico della Fisica e Centro Studi e Ricerche "E. Fermi", Roma, Italy

[g] Laboratoire d'humanités digitales, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract

We present a fundamental development step of a new technique to read and digitize ancient handwritten documents. Chemical analysis by x-ray fluorescence and x-ray tomography enabled us to decipher words and drawings from inside a closed, 200-pages 18th century handwritten book. The ink chemistry is essential: tomographic reading is feasible thanks to the iron present in ancient inks (iron gall) over one millennium — whereas carbon or organic inks do not provide sufficient x-ray contrast. The results presented are a key progress towards the ultimate goal of the technique: non-invasive reading of fragile and/or unopenable documents.

1. Introduction

We applied the new x-ray tomography "virtual reading" technique to read inside a large ancient handwritten book without opening it. Positive results were obtained using a compact, laboratory-based radiology system.

This work is based on preliminary investigations that included extensive chemical analysis of ancient manuscripts over six centuries [see Fig. 2[b] and [1]]. The objective was to verify the iron content of black inks in ordinary specimens, e. g., private and administrative records. This element and its quantity are crucial for the technique: it's the x-ray absorption of the ink that provides the necessary x-ray contrast not only for radiology but also for tomographic reconstruction of the writings.

Previous tests in the similar direction also included text recognition by tomography on small fragments of ancient and modern manuscripts using synchrotron radiation [1] and laboratory based [2] x-ray sources. The investigations provided the capability of the technique to detect characters and words on small samples and also assessed the use of contrast mechanisms based on phase effects rather than on attenuation.

All these pioneering efforts aimed to address the key issue: can the technique work for real, large-size books with hundreds of pages — still using a laboratory-based equipment suitable

for future applications in the manuscript collections sites? This was the objective of the present investigation, and the results are positive.

The development of the virtual reading technique is primarily inspired by the Venice Time Machine (VTM) project [6]. This is an ongoing collaboration between the Ecole Polytechnique Fédérale de Lausanne (EPFL) and two institutions in Venice: the University Ca' Foscari and the “Archivio di Stato”. The Archivio is an historical collection containing almost 100 km of handwritten documents covering ten centuries of the administrative and legal life of Venice. But, as for all ancient collections, their exploitation by scholars is problematic for conservation and logistic reasons: without massive digitization, deciphering, indexing and storage, they are almost unusable. Such are the tasks targeted by the VTM project.

The project also includes the development of novel digitization techniques, since the present ones would require up to 20 years to complete the task. Among the new approaches, a leading one is virtual reading: the use of x-ray imaging to analyze specimen without opening them. The approach is based on pioneering research of other authors [10, 4, 11, 7, 15, 14, 17, 3] and on the wide experience of Swiss institutions in x-ray techniques. Virtual reading is not limited to conventional tomography based on x-ray absorption but also involves phase approaches [12, 13, 8, 18].

2. Materials and methods

2. 1. Chemical analysis of ancient inks

The book was chemically analyzed with x-ray fluorescence (XRF) spectroscopy in order to assess the chemical composition of the black ink (no color inks were present) and predict the x-ray contrast. Fig. 1 shows the head of the spectroscopy instrument during the analysis.
Fig. 2[a] presents XRF spectra of ink in text and drawings. The results emphasize the iron content fluctuations, which could be explained by the frequent use of new inks and/or by the handwriting process. These features are also systematically found for all other manuscripts under investigation. Fig. 2[b] shows indeed iron content data for a large set of manuscripts, covering six centuries, from religious parchment to books.

Fig. 1. The XRF instrument during the chemical analysis of inks and paper.

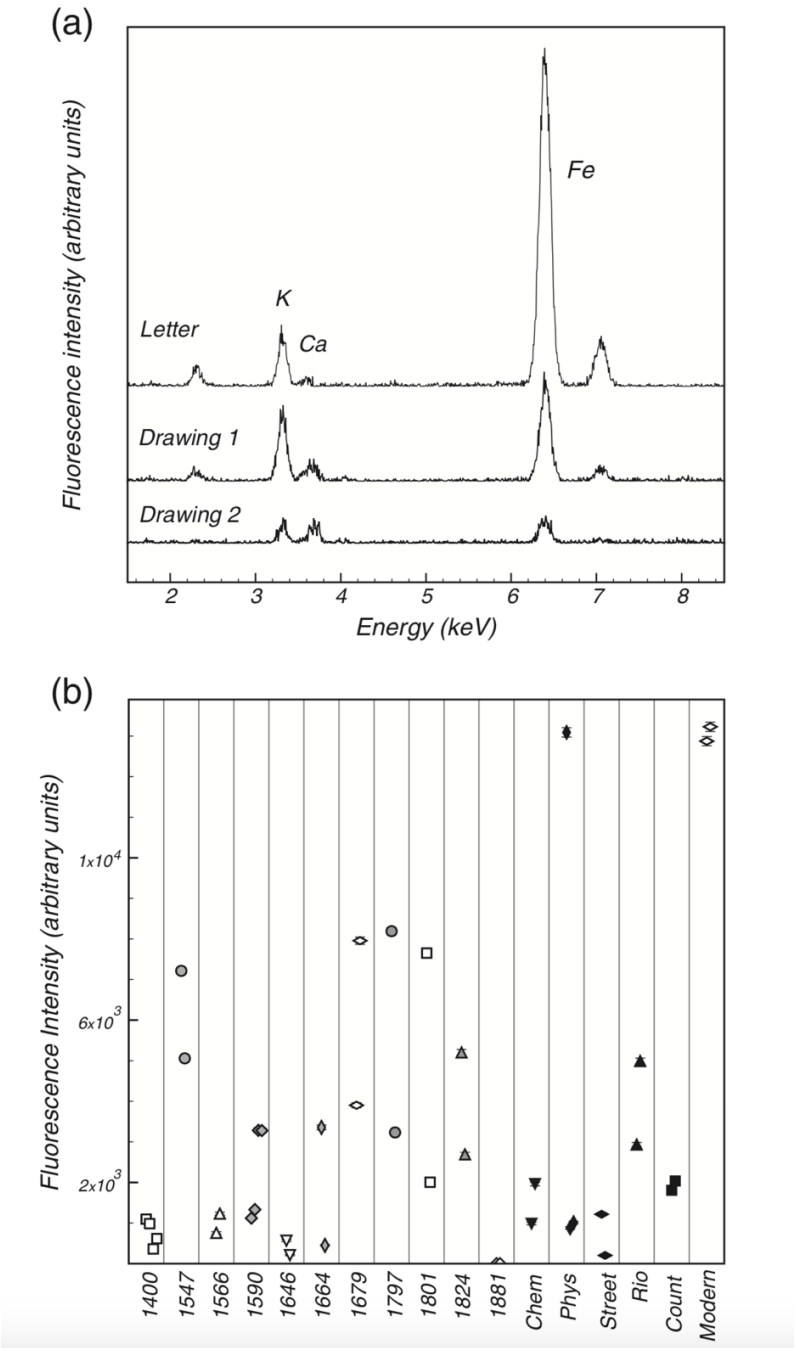

Fig. 2. [a]: XRF spectra for different ink spots in the 200-page manuscript. Note the presence of iron and the large fluctuations in its relative density. [b]: Survey of iron content in the inks of many different ancient manuscripts over six centuries. The labels identify the specimen age or its nature. The last six specimens were written after the mid-18th century: a chemistry books, a physics book (the present specimen), two legal books, a short novel and a simulated stack ("modern") written with iron gall ink.

The x-ray fluorescence data were taken with a portable μ-XRF spectrometer Artax (mod. 400, Brucker) using a 200 μm collimator. Even if not fully quantitative, the results can be compared to each other thanks to the use for all measurements of the same experimental conditions and geometrical set-up.

2. 2. X-ray tomography

Computer tomography (CT) consists in acquiring a large set of projection radiographs at equidistant angles with respect to the source-detector system, carrying enough x-ray absorption information to perform three-dimensional reconstructions. Fig. 3 shows the

imaging system, with the plexiglas book holder placed on the motorized translational and rotational sample stage that accurately control the specimen position.

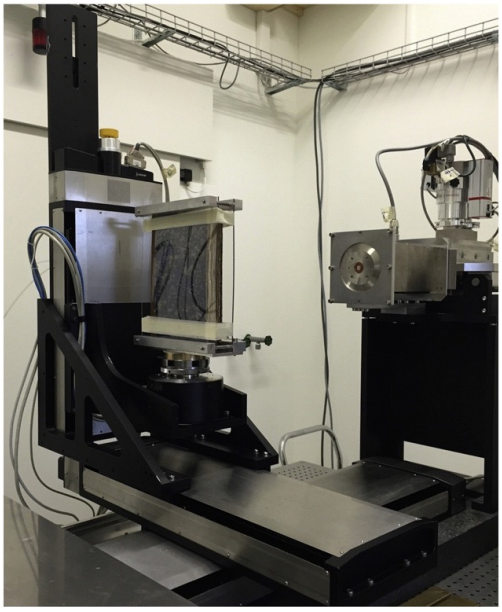

Fig. 3. The instrument for acquisition of the raw images for x-ray tomography.

The experimental μCT instrument available at the Center for X-ray Analytics at EMPA in Dübendorf, Switzerland [5], consists of a microfocus tube (VISCOM XT9160-TXD), operating at 80 keV with an electrical current of 250 μA and a flat panel detector (Perkin Elmer XRD 1621 CN3 ES), 2048 × 2048 pixels and 409. 6 × 409. 6 mm^2, with a maximum image data range of 16 bit. Considering the magnification and the geometrical setup, the final voxel size was 50 μm.

After detector calibration, performed taking into account dark and flat field images, 2001 projections were acquired over 360 degrees. Each projection was the result of an averaged acquisition of 16 frames, with an exposure time of 400 ms per frame.

The dataset was reconstructed using the tomographic reconstruction software Octopus Imaging Software. Source–detector as well as source–object distances were taken in account as cone beam reconstruction parameters, an optimal center of rotation (COR) was evaluated and finally used for reconstruction.

The paper book analyzed was a handwritten scientific manuscript, dated 1790–1800, approximately 27 × 19 cm^2, with a cardboard cover. The author Giulio Mancini, an Italian scientist from Citta' di Castello (near Perugia), wrote a N 200 pages text with notes about different branches of physics, also including several inserts with drawings and graphs.

During the tomographic scanning particular attention was taken in continuously monitoring the specimen for possible damage by x-ray exposure. This was done by a visual comparison before and after each acquisition set, and by a more detailed inspection after the entire experiment. Actually, no damage was expected due of the nature of x-ray interaction with the involved materials — and, indeed, it was not observed.

To prevent any damage we decided to sharply limit the x-ray source emission; this increased the total acquisition time up to 5 h. Future work will provide an optimization of the source intensity and exposure time to drastically reduce the total acquisition time.

3. Results and discussion

Fig. 4 shows a typical example of individual x-ray projection. Clearly, the overlapping of many pages makes impossible to directly use the radiographic images for our text recognition objective. A posteriori extraction of bi-dimensional pictures from the three-dimensional tomographic reconstructed volume solves this problem.

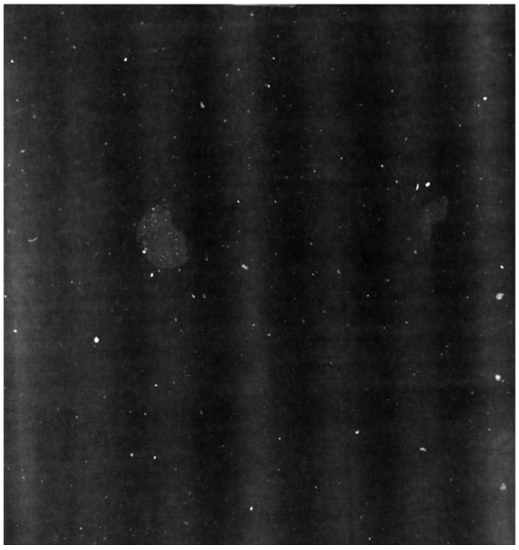

Fig. 4. Example of projection radiograph, with no identifiable contributions from individual pages or characters

Fig. 5 shows an example of reconstructed volume: the three-dimensional structure of the $10 \times 10 \times 5$ cm^3 portion of the book investigated. We can clearly see page side edges.

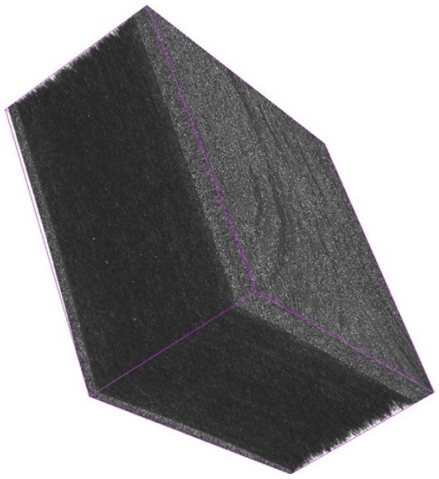

Fig. 5. Three-dimensional tomography volume reconstructed ($10 \times 10 \times 5$ cm^3).

The most significant tomography reconstructions are those concerning individual pages. Figs. 6 and 7 show several examples of words extracted from inside pages compared to the visible pictures of the texts. Fig. 8 shows similar result for a drawing from an insert

sheet. In agreement with the preliminary tests of Refs. [1, 2], we can positively confirm that with x-ray tomography characters and words extraction from ancient handwritten documents is feasible.

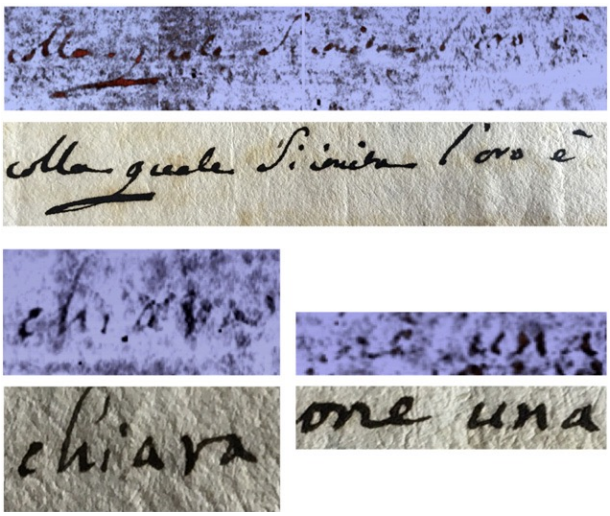

Fig. 6. Several examples of tomography reconstructed inner page portions revealing words and sentences (top), compared to visible pictures (bottom).

Fig. 7. Additional examples of results similar to Fig. 6.

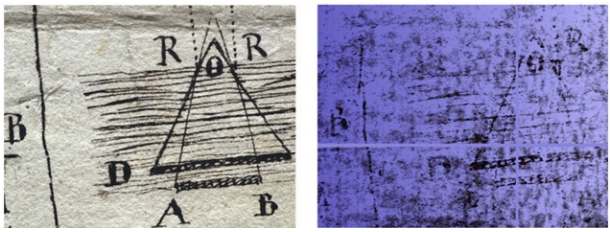

Fig. 8. Results like those of Figs. 6 and 7, for a drawing illustrating refraction from a prism (visible image of the left).

Nevertheless, the extraction is not trivial without an ad hoc segmentation algorithm. Fig. 9 shows an example of page extraction over the full field of view acquired: although individual words can be seen, the reconstructed image contains words from more than one page — as indicated in particular by the character orientation. One of the most challenging efforts

in the algorithm development is the warping of the ancient pages. The task is easier when significant air gaps exist between the pages: in such cases, we can be optimistic about the feasibility of automated page extraction.

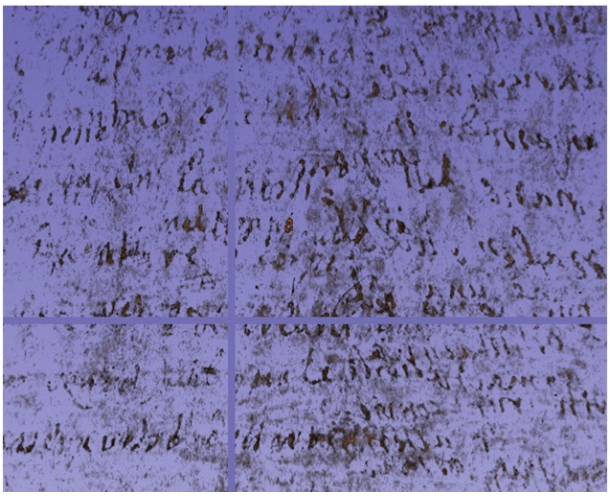

Fig. 9. Tomography reconstruction of an extended area, revealing words from different pages.

Besides letters and words, tomography can also reveal other hidden features from the inside of the book. Fig. 10 shows indeed the seal of a private letter, written to the author by a colleague and inserted in the volume. Note the microscopic cracks in the seal, revealed by the good spatial resolution of the tomographic reconstruction (50 μm^3).

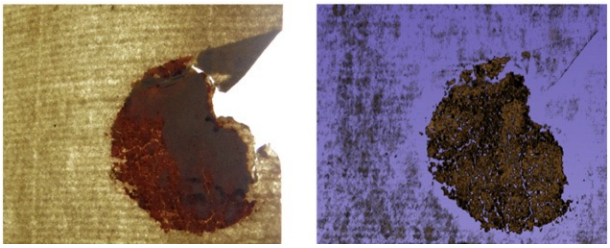

Fig. 10. Visible (left) and tomography images of the seal of a private letter inserted in the book. Note the folded portion of the paper near the seal, visible in both pictures.

4. Conclusions

Our investigations confirm the potentiality of the non-invasive, lab-based technique to "virtually read" large ancient handwritten manuscripts. However generally positive, results also reveal some critical obstacles. In particular, to become a competitive technique, the acquisition of the tomographic dataset must be accelerated. This will require additional tests to progressively increase the source current while carefully monitoring possible signs of damage.

Moreover, the standard algorithm for tomographic reconstruction should be optimized for the particular structure of our object. But the most important issue is the development of an ad hoc segmentation algorithm to separate and extract warped pages.

In the long run, our approach will be affected by a general problem of massive digitization programs trying to preserve for future generations our invaluable document patrimony: while

books can be read for centuries, how we guarantee the same for digital images? A standard strategy must be found to made images readable and searchable for many years, using a “non-copyrighted” format, continuously adapted to the new reading technologies — for example the FITS (Flexible Image Transport System) format [16]. In parallel, efforts based on web technologies seek solutions for long-term storage of images and related data [9].

Finally, our present tests confirm the basic message of the previous experiments: a future systematic use of the technique cannot rely on reconverted instrumentation but requires systems specifically designed for this task.

Acknowledgments

The authors are grateful to Patrick Aebischer for his leadership of the VTM project, to the Verbantiqua bookstore (www.copernicum.it) for their generous donation of antique manuscript specimens, to Ferruccio Petrucci for his assistance in the chemical analysis and for several stimulating discussion and to the staff of the Center for X-ray analytics, EMPA, for the excellent technical assistance. The research was supported by the Venice Time Machine Project and the Center for Biomedical Imaging (CIBM).

References

1. F. Albertin, A. Astolfo, M. Stampanoni, E. Peccenini, Y. Hwu, F. Kaplan, G. Margaritondo, Ancient administrative handwritten documents: X-ray analysis and imaging, J. Synchrotron Radiat. 22 (2015) 446–451.
2. F. Albertin, E. Peccenini, Y. Hwu, T. -T. Lee, E. B. Ong, J. H. Je, F. Kaplan, G. Margaritondo, The Venice Archivio di Stato: Innovating Digitization with X-ray Tomography, Proceeding of Digital Heritage 20152015.
3. R. Baumann, D. C. Porter, W. B. Seales, The use of micro-CT in the study of archaeological artifacts, Proc. of 9th Int Conf. on NDT of Art 2008, pp. 1–9.
4. P. D. Carmine, L. Giuntini, W. Hooper, F. Lucarelli, P. Mandó, Further results from PIXE analysis of inks in galileos notes on motion, Nucl. Instrum. Methods B 113 (1996) 354–356.
5. EMPA, Center for X-ray Analytics, http://www. empa. ch/x-ray2015.
6. EPFL, Digital Heritage Venice, http://dhvenice. eu/ 2015.
7. J. Gunneweg, A. Adriaens, J. Dik, Holistic Qumran: Trans-disciplinary Research of Qumran and Dead Sea Scrolls, Online Publication: Brill2010.
8. Y. Hwu, H. H. Hsieh, M. J. Lu, W. L. Tsai, H. M. Lin, W. C. Goh, B. Lai, J. H. Je, C. K. Kim, D. Y. Noh, H. S. Youn, G. Tromba, G. Margaritondo, Coherence-enhanced synchrotron radiology: refraction versus diffraction mechanisms, J. Appl. Phys. 86 (1996) 4613–4618.
9. E. Hyvönen, Publishing and using cultural heritage linked data on the semantic web. Synthesis Lectures on the Semantic Web, Theory Tech. 2 (1) (2012) 1–159.
10. Library Special Collections Conservator Unit, Preservation Department, Y. U., Medieval Manuscripts, Some Ink and Pigment Recipes, Yale University, 2012.
11. F. Lucarelli, P. Mandó, Recent applications to the study of ancient inks with external-PIXE facility, Nucl. Instrum. Methods B 109 (1996) 644–652.
12. G. Margaritondo, Elements of Synchrotron Light for Biology, Chemistry, and Medical Research, New York: Oxford, 2002.
13. G. Margaritondo, G. Tromba, Coherence-based edge diffraction sharpening of x-ray images: a simple model, J. Appl. Phys. 85 (7) (1999) 3406–3408.
14. D. Mills, O. Samko, P. Rosin, K. Thomas, T. Wess, G. Davis, Apocalypto: Revealing theUnreadable, Proc. SPIE 8506, VIII:85060A2012.
15. V. Mocella, B. Emmanuel, C. Ferrero, D. Delattre, Revealing letters in rolled Herculaneum papyri by x-ray phase-contrast imaging, Nat. Commun. 6 (2015) 5895.
16. NASA, FITS — The Astronomical Image and Table Format, http://fits. gsfc. nasa. gov/ 2014.
17. W. Seales, W. Griffioen, R. Baumann, M. Field, Analysis of Herculaneum papyri with x-ray computed tomography, Proceedings of 10th International Conference on NDT of Art, Jerusalem 2011, pp. 1–9.
18. T. Weitkamp, A. Diaz, C. David, F. Pfeiffer, M. Stampanoni, P. Cloetens, E. Ziegler, X-ray phase imaging with a grating interferometer, Opt. Express 13 (2005) 6296–6304.