

Madonne: Document Image Analysis Techniques for Cultural Heritage Documents

Jean-Marc Ogier* and Karl Tombre†

Abstract. *This paper presents the Madonna project, a French initiative to use document image analysis techniques for the purpose of preserving and exploiting heritage documents.*

1. Introduction

Experts plead for strong actions guaranteeing a lasting preservation of our cultural and scientific resources, which represent a living and collective memory of our societies. The evolution of our economies towards a model based on *digital content* has a deep impact on this preservation; the challenge is to make this impact a benefit and not a drawback. Large resources have been invested on digitization programs for the cultural heritage, including museum collections, archaeological sites, audiovisual archives, maps, historical documents, and manuscripts. However, several factors can become a hindrance in optimizing the management of these resources.

First, the approach is often *fragmented*, with a lack of global and strategic management tools and no common policy on the handling of already digitized resources and on setting priorities; hence the threat of waste in resources, efforts and investments. Digitization is also *costly* and needs huge investments, often based on public funding. Some kind of “return on investment” is expected, at least from the point of view of lasting availability and usability of the digitized resources. But the technologies and standards chosen and used today may quickly become obsolete and inadequate. *Intellectual and industrial property rights* also lead to various problems: Many partners have obviously rights and claims on the digitized content, which need to be acknowledged and taken into account. There is a strong need for common solutions for handling these rights in the cultural domain.

During the whole acquisition process—from scanning the paper and all the way to indexing the digital documents—many precautions must absolutely be taken if automated techniques are to be used. One typical example is the fact that many institutions produce highly compressed files, e.g. using JPEG, which sometimes hinders the use of automated image processing techniques. Thus, institutions which do not consider all the constraints relative to the global “valorization process” produce more or less unusable data, from the point of view of automation. Among the fundamental constraints, let us cite the resolution of the images, that must be at least around 200 or ideally 300 dpi for a long term exploitation strategy. This highlights the necessity of having a close dialogue between different communities, from social and human sciences researchers to computer science specialists.

*L3i, Université de la Rochelle, Avenue Michel Crépeau, 17042 La Rochelle CEDEX 1, France, email: Jean-Marc.Ogier@univ-lr.fr

†LORIA-INPL, École des Mines de Nancy, Parc de Saurupt, 54042 Nancy CEDEX, France, email: Karl.Tombre@loria.fr

From the point of view of pattern recognition in general and document image analysis more specifically, we are in the presence of a classical problem involving image processing techniques as well as computing of invariants used for indexing, and database management issues. Compared to classical document image analysis problems, the main differences are due to the amount of data, which raises new research problems. This huge amount of data produces problems with respect to the organization of the feature space in which the documents are transcribed. Another important difference with classical recognition problems is the wide variability of representation of the information that can be found in ancient documents. The fact that the images are often degraded by noise adds to the difficulty. Finally, and this is probably the most important difficulty, the problem of having an exhaustive expression of the future usage of the indexed documents raises the question of how to structure the information and of the cues that have to be extracted from the images.

In this general context, the *Madonne* research project, funded by the French government in the general *ACI Masse de données* program, aims at designing methods for going beyond plain digitization projects for historical documents, as such projects tend to yield large databases of images with very little *structure* for navigating and indexing these databases. The project investigates the use of document image analysis methodology for providing useful browsing and indexing features in these large collections. The *Madonne* project is thus focused on indexing, organization and incremental enrichment of heritage data, in order to provide general services to the users, including researchers in human and social sciences, and to work towards interoperability of data and browsing tools.

The *Madonne* project started at the end of 2003 and will end at the end of 2006. In addition to our groups at the L3i laboratory at University of La Rochelle and at LORIA in Nancy, the partners were from the PSI lab at University of Rouen, the LI lab at University of Tours, the LIRIS lab at University of Lyon, the CRIP5 lab at University Paris 5, and IRISA in Rennes. These research partners had a strong and complementary background in document image analysis, thus allowing us to build a large set of generic services. This paper presents an overview of the main results obtained by the partners of the *Madonne* project, and especially by B. Coüasnon, V. Eglin, L. Heutte, N. Journet, T. Paquet, R. Pareti, J.-Y. Ramel, J.-P. Salmon, S. Tabbone, S. Uttama, N. Vincent and L. Wendling.

2. Specific Research themes

To reach the objectives of providing structured access and browsing capabilities to large sets of cultural heritage documents, we need to index these sets using the various features which can be of interest for searching. This includes illustrations, text, styles, various kinds of symbols, handwritten annotations, etc. This leads us to the need for close *cooperation* between various document analysis expertise areas, as none of these areas answers the requirements on its own. In the following, we will detail some of the research themes we addressed in the *Madonne* project.

2.1. Collection modelling

In the context of large collections of data¹, one can observe a strong homogeneity in the way the information is structured, depending on the different collections. Collections modeling consists in extracting as automatically as possible the features that characterize a collection or a set of collections, in order to assist the analysis of the images, by applying appropriate image processing tools.

¹One of our major data providers is the Center for Higher Renaissance Studies at University of Tours, see <http://www.cesr.univ-tours.fr/>

This question raises the problem of automatically discovering the similarities concerning the structures of the books, in order to construct a relevant model of the corresponding collection. In the context of the Madame project, Journet et al. [4] proposed a set of processes for categorizing the pages of a book according to the spatial organization of the data. The extraction of some features describing the layout and the structure of the document allowed them to structure the collections of books, in term of similarities between the spatial organization of their contents. For that purpose, Journet proposed a function based on autocorrelation for the extraction of the features (Fig. 1). The future of this work will consist in measuring the similarities between different books, providing the required models for the collection.

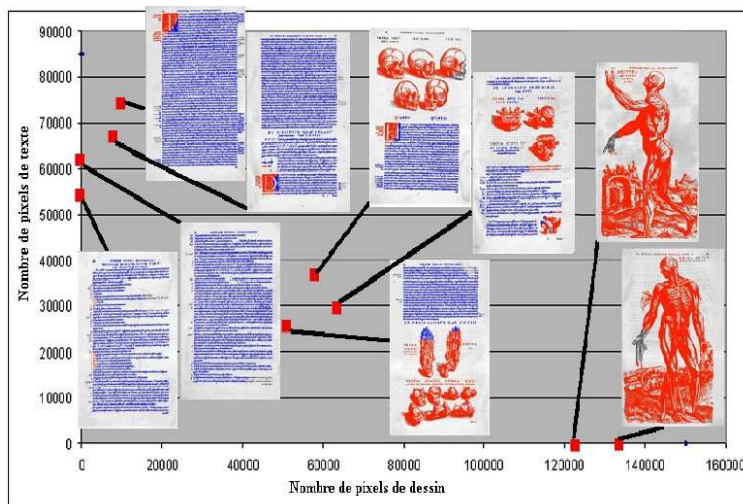


Figure 1. Categorization of the image as a function of their content.

2.2. Document Layout Analysis

The structure of a document is usually relative to a presentation and organization model and aims at helping the user in understanding the information provided by the document. Specific challenges appear in old collections, as the typical documents from the 15th, 16th or 17th century dealt with in our project. In a number of cases, the layout itself conveys precious information for browsing the documents.

The analysis of the elements of the layout may be an excellent guide for content based information retrieval, by using full text search of similarities measurements, applied to specific zones yielded by the layout analysis. It may also guide us in finding the information which can be made readily available to the general public, as opposed to information which is protected by privacy, confidentiality or property rules. In the context of the Madame project, let us cite the work of Couasnon [3] on the collective annotation process of military registers from the 19th century (Fig. 2). This process goes through a very reliable analysis of the structure of the documents, based on 2D grammar techniques integrated in the DMOS², system, allowing to detect each cell of the military register even if the structure of the document is degraded. Thanks to this fine detection of the cells, the system proposes a similarity measurement system allowing to browse military registers on handwritten names with textual queries without OCR. The similarity measure is based on the extraction of low level primitives,

²Acronym for “Description and MOdification of Segmentation”, an analysis system designed at IRISA, Rennes.

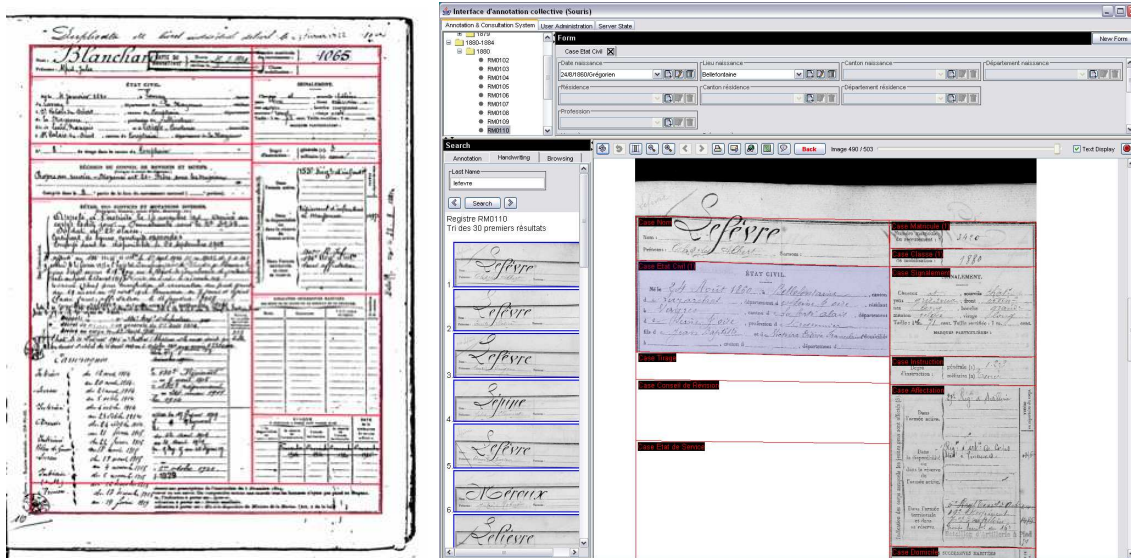


Figure 2. Left: Structure analysis with the DMOS system, by Couasnon [3]. Right: Access by handwritten names after integration into a platform for annotations.

graphemes, the organization of which permits to provide a measure of similarity between two handwritten models. The difficult points encountered here are relative to the overlapping of graphic layers, for which text-graphic segmentation techniques may be useful. This system has been validated on 165,000 pages.

Another contribution has been the building of a system called Agora for the interactive analysis of document layout [8]. Depending on the needs (extraction of ornamental letters, of marginal notes, of titles...), the user can thus build scenarios allowing to label, to merge or to remove the extracted blocks. The scenarios can be stored, modified and applied to other sets of images in batch processing mode.

2.3. Handwritten documents

The processing of handwritten manuscripts from the cultural heritage leads to specific questions which are far away from usual handwriting recognition analysis as addressed in postal or banking applications, for instance. The aim is rarely to recognize the handwriting but rather to characterize and identify different writers [2], or to date some documents. Fig. 3 is a typical example of what we aim at working on. Indexing based on visual information features is therefore one of the main keywords for us. In specific cases (handwritten name registers for instance) global shape recognition techniques can lead to classification according to shape similarities and even to limited handwriting recognition for indexing purposes [3]. For this, lexical knowledge about the domain can be of considerable help.

In the context of the Madame project, Paquet et al. have proposed a set of processes allowing to help historians to analyze Flaubert's manuscripts layouts³ [5]. In this context, some relevant signatures are computed in order to check that the spatial organization of the data match features characterizing Flaubert's handwriting style. In this case, Hidden Markov Models, as well as dynamic programming, are used for the segmentation and modeling process. The results highlight that the relevant features that have to be considered in such a process combine handwriting features and structural information

³Gustave Flaubert, French author, 1821–1880.

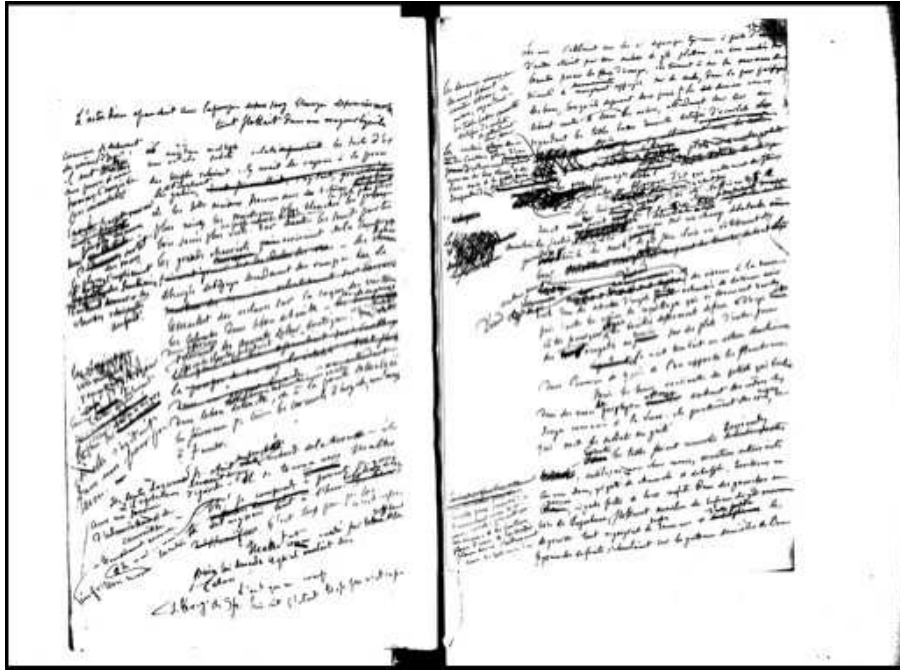


Figure 3. A handwritten manuscript with annotations by the author.

about the spatial organization of the data. Such an analysis leads to characterization of the author's style (authentication), but also to the possibility of "reconstructing" the genesis of the writing process through the successive annotations.

2.4. Indexing on Graphical features

Usually, documents are mainly indexed on text. However, heterogeneous sets of historical documents often contain features which are graphical in nature, although they represent text. This is especially the case with illustrated dropcaps associated with artwork (Fig. 4), on which we have focused a



Figure 4. Example of illustrated dropcap.

lot of work in the Madonne project [6]. There is little knowhow on how to compute invariants for this kind of features. Actually, our experience with our CESR partners highlights the diversity of requirements that may be expressed by the users. Some historians want to detect slight differences between dropcaps in order to be able to date them, while some others are only interested in global content based retrieval problems (find similar dropcaps).

A first problem to be addressed with these features is that of the document image segmentation problem. The images to be processed are noisy and we have used an adaptive image-smoothing filter which is more robust to different noise levels than existing methods (Fig. 5).

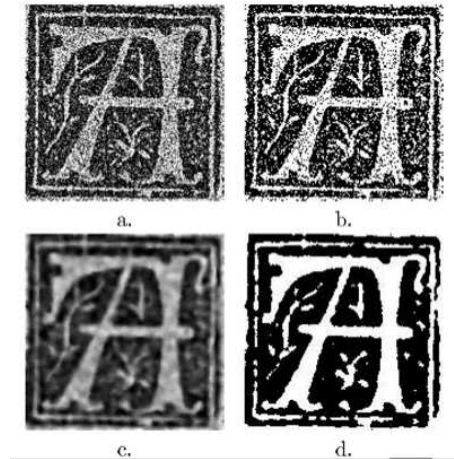


Figure 5. Noise filtering.

Complementary sets of descriptors have been developed in the CRIP5 Laboratory and in the L3i laboratory. The first is based on a statistical modeling of the distribution of the pixels within a dropcap, using the Zipf law [7]. This allows us to classify the dropcaps as a function of their style, thanks to the analysis of specific ruptures according to the Zipf law (Fig. 6). The second is based on a top down

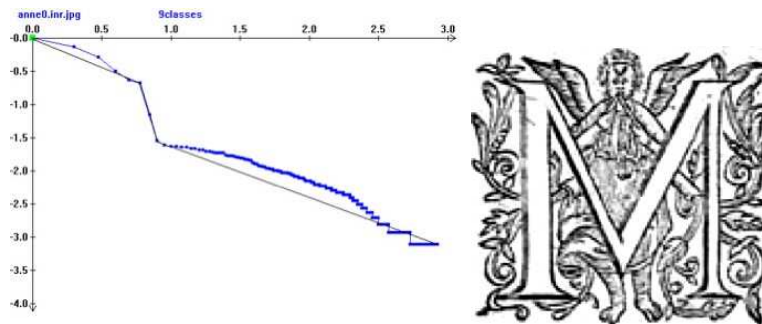


Figure 6. Zipf law of a specific dropcap [7].

segmentation process providing a set of layers on each of which a signature is computed [10], in order to characterize the spatial organization of the data. This process allows us to implement a content based image retrieval system, the results of which are very encouraging in terms of recall/precision (Fig. 7).

We have also defined a new method for combining shape descriptors based on a behavior study of a learning set [9]. Each descriptor is computed on several clusters of objects or symbols. For each cluster and for any descriptor, an appropriate mapping is directly carried out from the learning database. Then, existing conflicts are assessed and integrated into a map. Such a combination of descriptors improves the recognition rates and the ranking obtained on dropcaps like those in Fig. 8.

2.5. Similarity retrieval for document compression

The digitization of cultural documents also leads to difficulties in terms of storage and transmission on a bandwidth-limited network. Only lossy compression with an acceptable perceptible loss of information may reduce the weights of images. The existing compression formats like JPEG, DJVU



Figure 7. Content based image retrieval [10].

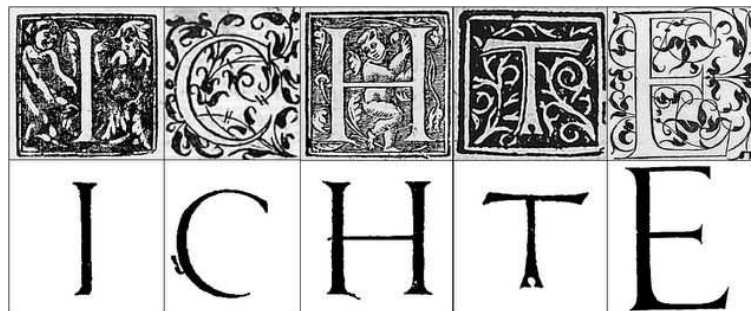


Figure 8. Letters extracted from dropcaps and used in retrieval and ranking applications [9].

or DEBORA are unfortunately not effective on handwritten documents images, due to the great complexity of handwritten shapes and to the difficulty of localizing them precisely (see Fig. 3). Within the Madonna project, a handwritten text compression methods has been proposed [1], consisting in separating the text and the background by similarity retrieval. Basically, the localization of redundancies is based on a decomposition of the handwriting text into oriented segments with invariant contours points and can be extended to any part of the image which presents distributed similarities.

3. Achievements and open problems

As one can see through these different points, the preservation of cultural heritage documents requires to combine various methods from the document image analysis field: image processing, handwriting recognition, document layout analysis, graphics recognition, etc. However, many problems remain open, and require more work. This concerns mainly the problem of scaling the recognition approaches, because of the variability of representation that is one of the specific feature of ancient documents. Another topic is the problem of modeling the domain knowledge, in order to assist the user producing a relevant scenario when dealing with a specific subject. Of course, this huge amount of data raises new problems, specifically in relation with “content based” operations. For graphics in old manuscripts, for instance, some new signatures have to be developed in order to design word spotting or graphic spotting methods (Fig. 9), similar to what can be achieved through hyper-text navigation.

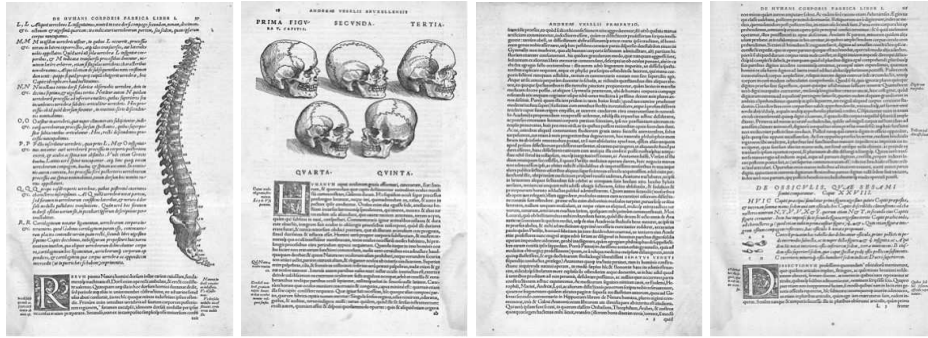


Figure 9. Perspective: Navigation through graphic spotting technique.

References

- [1] A. El Abed, V. Eglin, and F. Lebourgeois. Frequencies decomposition and partial similarities retrieval for patrimonial handwriting documents compression. In *Proceedings of 8th International Conference on Document Analysis and Recognition, Seoul (Korea)*, pages 996–1000, 2005.
- [2] A. Bensefia, T. Paquet, and L. Heutte. A writer identification and verification system. *Pattern Recognition Letters*, 26(13):2080–2092, October 2005.
- [3] B. Coüasnon and I. Leplumey. A Generic Recognition System for Making Archives Documents Accessible to Public. In *Proceedings of 7th International Conference on Document Analysis and Recognition, Edinburgh (Scotland, UK)*, pages 228–232, August 2003.
- [4] N. Journet, V. Eglin, J.-Y. Ramel, and R. Mullot. Text/Graphic Labeling of Ancient Printed Documents. In *Proceedings of 8th International Conference on Document Analysis and Recognition, Seoul (Korea)*, pages 1010–1014, September 2005.
- [5] S. Nicolas, T. Paquet, and L. Heutte. Enriching Historical Manuscripts: The Bovary Project. In *Proceedings of the 6th IAPR International Workshop on Document Analysis Systems, Florence, (Italy)*, volume 3163 of *Lecture Notes in Computer Science*, pages 135–146, September 2004.
- [6] R. Pareti, S. Uttama, J.-P. Salmon, J.-M. Ogier, S. Tabbone, L. Wendling, and N. Vincent. On defining signatures for the retrieval and the classification of graphical dropcaps. In *Proceedings of 2nd IEEE International Conference on Document Image Analysis for Libraries*, pages 220–231, Lyon, France, April 2006.
- [7] R. Pareti and N. Vincent. Global Discrimination of Graphics Styles. In *Proceedings of 6th IAPR International Workshop on Graphics Recognition, Hong Kong*, August 2005.
- [8] J.-Y. Ramel and S. Leriche. Segmentation en analyse interactives de documents anciens imprimés. *Traitement du Signal*, 22(3):209–222, November 2005.
- [9] J.-P. Salmon, L. Wendling, and S. Tabbone. Improving the Recognition by Integrating the Combination of Descriptors. *International Journal on Document Analysis and Recognition*, 2006. Accepted for publication.
- [10] S. Uttama, J.-M. Ogier, and P. Loonis. Top-down segmentation of ancient graphical drop caps: lettrines. In *Proceedings of 6th IAPR International Workshop on Graphics Recognition, Hong Kong*, pages 87–96, August 2005.