# HISTORICAL DOCUMENT ANALYSIS: A REVIEW OF FRENCH PROJECTS AND OPEN ISSUES

*Mickael Coustaty, Romain Raveaux and Jean-Marc Ogier*

L3i Labs
University of La Rochelle
Avenue Michel Crepeau, 17042 La Rochelle Cedex 01
phone: + 335 46 45 82 62, fax: + 335 46 45 82 42
email: {mcoustat, rravea01, jmogier}@univ-lr.fr

## ABSTRACT

This subject is on the crossroad of different fields like signal or image processing, pattern recognition, artificial intelligence, man-machine interaction and knowledge engineering. Indeed, each of these different fields can contribute to build a reliable and efficient document interpretation system. This paper points out the necessities and importance of dedicated services oriented to historical documents. In a first step, a bird view approach is adopted describing document specificities and associated projects which deal with the enrichment and the exploitation of heritage documents. This synthesis lead to a set of particular Research Problems. The second part focuses on a set of open issues, which should be tackled by the document analysis community, for the management of the features and the knowledge representation of these ancient 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programs for the cultural heritage, including museum collections, archaeological sites, audiovisual archives, maps, historical documents, and manuscripts.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 management and strategic management tools and no common policy on the management 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 become quickly 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 to ensure the possibility of using automated techniques. 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 the highest as possible 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.

Images of documents are like no-others, documents are made by human for human purpose. 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 changes are due to the amount of data and the poor meta-data surrounding this heap of pixels. Commonly, the question of the organization of the feature space in which the documents are transcribed arisen. 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.

The effort to manage ancient documents so far seems to be in progress. In France, firstly, this idea was generally fragmented. There was a lack of global and strategic management tools and no common policies on handling of ancient document resources and on setting priorities in management. This results in the threat of waste in resources, efforts and investments. Digitization is also costly and needs huge budgets, often based on public funding. Fortunately, from the support of French government and the collaboration of many research laboratories, the projects called MADONNE and NAVIDOMASS[1] were set up for the purpose of preserving and exploiting ancient documents. These pioneer projects opened the way to more and more challenging relations between ICT-HSS communities (Information & Communica-

[1] http://madonne.univ-lr.fr and navidomass.univ-lr.fr

tions Technology - Humanities and Social Sciences). French and European initiatives such as the french digital library GALICA, and the British Library show the engagement for this cause. This effervescence denotes the matter of the digitization of our cultural heritage.

## 2. HISTORICAL DOCUMENT SPECIFIC RESEARCH TOPICS

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 discuss the main research themes in the field of ancient document analysis.

### 2.1 Collection modelling

In the context of large collections of data, 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 adapted image processing tools. 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 MADONNE project, Journet et al. [7] proposed a set of processes allowing to categorize 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 (see figure 1). The future of this work will consist in measuring the similarities between different books, providing the required models for the collection.

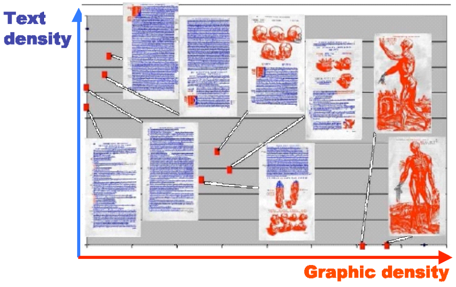
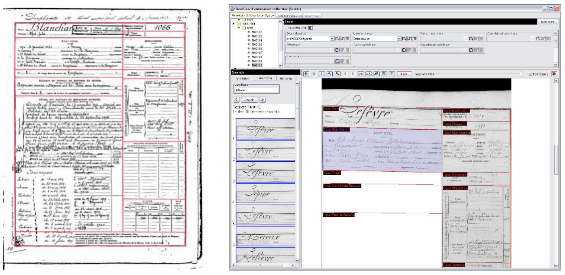


Figure 1: Image categorization as a function of the content

### 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 15$^{th}$, 16$^{th}$ or 17$^{th}$ 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 Madonne project, let us cite the work of Couasnon [5] on the collective annotation process of military registers from the 19th century (see figure 2(a)). This process goes through a very reliable analysis of the structure of the documents, based on 2D grammar techniques integrated in the DMOS 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, graphems, 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 textgraphic 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 [14]. 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.



(a) Example of document (b) Structure analyzed by the system

Figure 2: Structure analysis with the DMos System [5]

### 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. Figure 2(b) 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 [5]. For this, lexical knowledge about the domain of use can be of considerable help.

In the context of the Madonne project, Nicolas *et al.* have proposed a set of processes allowing to help historians to analyze Flaubert's manuscripts layouts [10]. In this context, some relevant signatures are computed in order to check that the spatial organization of the data match features characterizing Flaubert's layout 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 about the spatial organization of the data. Such an analysis leads to characterization of the author's (authentication), but also to the possibility of "reconstructing" the genesis of the writing process through the successive annotations.

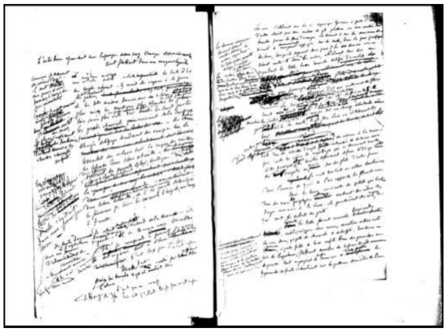


Figure 3: An example of handwritten manuscript with authors annotations

### 2.4 Indexing on Graphical Features

Usually, documents are indexed mainly 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 (see table 1), on which we have focused a lot of work in the MADONNE/NAVIDOMASS projects. There is little know on how to compute invariants for indexing documents on 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 [1], some others want to detect letter [6] (see figure 1), while some others are only interested in global content based retrieval problems (by using query by example searches, like in figure 5).

A first problem to be addressed is to define features to describe these particular images. Different sets of descriptors have been developed in the LIPADE Laboratory, in one hand, and in the L3i laboratory, on the other hand. The first laboratory uses a statistical law (the Zipf Law [11, 4]) to characterize distribution of the pixels within dropcaps. From these characteristics, a decorative style can be identified and used to classify images. The second laboratory characterize dropcaps after a top down segmentation process. This segmentation allows providing a set of layers on each of which a signature is computed as in [18], or specific elements are extracted as in [6]. These signatures present the advantage to take into account the spatial organization of the data.

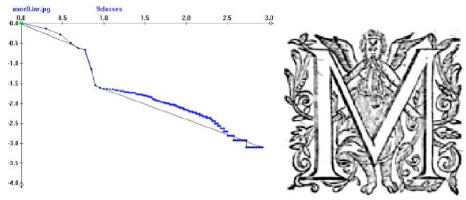
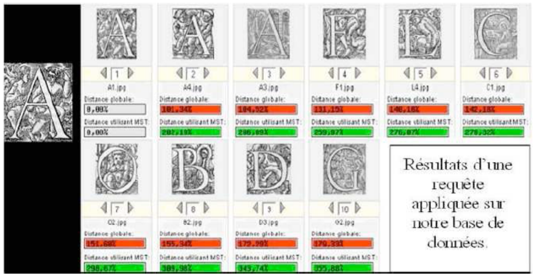


Figure 4: Zipf Law of a specific DropCap [11]

As the letter is important for historians, two approaches [9, 6] have been developed for this specific problem. The first approach, developed by the LORIA's group, used a combination of different shape descriptors based on a behaviour study of a learning set. 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 using a ranking measure. The second approach, developed by the L3i's team, extract simplified shapes using a combination of two decomposition. Once the shapes have been separated, they are described using shape features. A Selection rule is applied allowing to only keep the letters. Some results can be observed in figure 1.

All these processes allow us to implement two content based image retrieval system[2], the results of which are very encouraging in terms of recall/precision.



1 A1.jpg Distance globale: 0,00% Distance utilisant MST: 0,00%

2 A4.jpg Distance globale: 101,34% Distance utilisant MST: 202,19%

3 A3.jpg Distance globale: 104,52% Distance utilisant MST: 206,09%

4 F1.jpg Distance globale: 131,15% Distance utilisant MST: 259,97%

5 L4.jpg Distance globale: 140,18% Distance utilisant MST: 276,07%

6 C1.jpg Distance globale: 142,18% Distance utilisant MST: 279,32%

7 O2.jpg Distance globale: 151,68% Distance utilisant MST: 298,67%

8 B2.jpg Distance globale: 155,34% Distance utilisant MST: 309,98%

9 D3.jpg Distance globale: 172,29% Distance utilisant MST: 343,74%

10 G2.jpg Distance globale: 179,33% Distance utilisant MST: 355,88%

Résultats d'une requête appliquée sur notre base de données.

Figure 5: Drop Cap indexing [18]

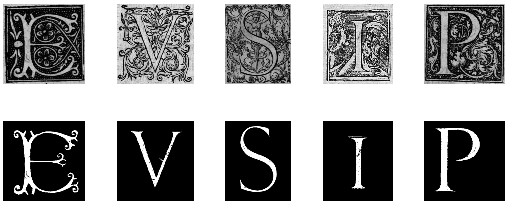


Table 1: Letters extracted from dropcaps and used in retrieval and ranking applications[6]

## 3. ACHIEVEMENTS AND OPEN PROBLEMS

As one can see through these different points, the preservation of cultural heritage documents requires to combine vari-

[2] The first one is displayed in figure 4 and the second one can be found at http://navidomass.univ-lr.fr/LettrineRequestWeb.html

ous methods from the document image analysis field : image processing, handwriting recognition, document layout analysis, graphics recognition, etc.

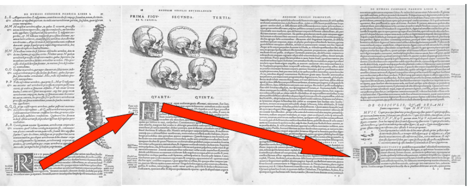


Figure 6: Hyper text navigation

However, many problems remain open, and require more works. Of course, this huge amount of data raises new problems, specifically in relation with ”content based” operations. For graphics in old manuscripts, some new signatures have to be developed in order to order offer hyper-text navigation based on meta-data obtained using word spotting or graphic spotting methods (see figure 6). These problems may be organized as follows:

1. Content characterization : the different examples illustrated in the previous parts highlight the necessity to work on the definition of relevant signatures for the indexing process. Depending on the kind of information to characterize, and depending of the user requirements, some of them can deal with structure, handwritten, or graphic indexing
   (a) In the context of manuscript documents, as one can see in the part dealing with handwritten document, many aspects should be considered for the authentication/transcription of a document. Some signatures integrating texture features and/or spatial based characteristics should be considered. These signatures should integrate statistical descriptors, combined in the context of structural signatures. Concerning indexing services, some new researches should be considered about word spotting techniques, in order to provide relevant shape based words descriptors, allowing to retrieve a particular word, without running any OCR system. Some interesting issues can be found in the works of [8]

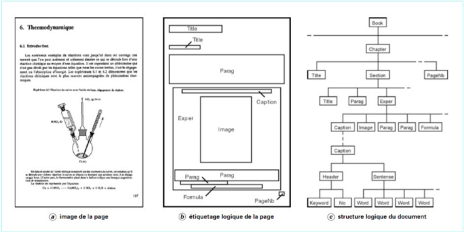
   (b) In the context of structure based indexing, some researches should be considered, in order to define spatial based signature for structure characterization. These signatures should be based on preliminary segmentation stages, in order to distinguish all the layers of information of the document : printed, manuscript, graphic,... This point highlights the necessity to work on image segmentation techniques, in the context of historical documents. Computer vision based techniques should be re-visited in order to analyse the approaches that may be adapted to the specific context of documents images. Some structural signatures, combining statistical descriptors should also be considered. Some interesting works issuing from [13] should be mentioned as interesting approaches for this kind of problem.

   (c) In the context of graphic images, some new content based characterization techniques should be developed. Indeed, in the context of ancient documents, most of the time, one has some to consider illustrations that had been printed thanks to wood stamps, ant the resulting images are generally images of strokes. As a consequence, these images have so specific features that the reuse of ”classical” images processing techniques is not so obvious. Different approaches are possible for characterizing such images like segmentation based techniques [18], Zipf law based techniques for content characterisation [12], key-points detection characterization or Wavelets decomposition. Whatever classical technique is considered, this one should be re-analysed in order to see how it has to be re-configured for being adapted to ancient document images.

2. Scale resistance : this concerns mainly the problem of scaling the recognition approaches, because of the variability of representation that is one of the specific features of ancient documents. This aspect appears to be one of the most fundamental aspects since it is based on the assumption that the number of objects in the learning database is very small. On this basis, the problem is here to characterize which features are relevant for the user requirements, and which are generic enough for representing the class of object to be indexed or recognized, from a generalization point of view. Some interesting approaches dealing with this problem can be found in [16].

3. Masses management: most of the time, each object to be retrieved is ”summarized” through a signature, that can be based either on a statistical or a structural description. When dealing with huge amount of document images, the problem is thus to be able to retrieve an object or a part of an object. In the context of a huge repository of images, an exhaustive and sequential comparison of the query with all the objects of the database is not reasonable, because of obvious complexity problems. As a consequence, in order to avoid such sequential research, the problem of structuring the features space becomes a crucial problem. Depending on the nature of the signature (statistical vs structural), this problem can be tackled by using different strategies. Considering statistical approaches, the indexing and clustering techniques in the context of high dimensional features vector should be explored, in order to structure the features spaces within a hierarchical manner, so that the access could be naturally indexed. Some interesting issues can be found in [17] In the context of structural description, the problem is often to structure graphs spaces, so that it is not necessary to run an exhaustive comparison between the query and all the graphs summarizing the images. Some interesting issues dealing with median graphs descriptions, spectral graphs and graphs probing approaches are interesting from this point. [3] (See figure 7)

4. Use interaction and Knowledge modelling: Another topic is the problem of modelling the domain knowledge, in order to assist the user for producing a relevant scenario when dealing with a specific subject. Most of the time, the lack of user requirements is a recurrent problem. Considering this point, three aspects seem important to develop. First, the necessity to provide human interfaces allowing the user to interactively construct image analysis scenarii, according to their objectives. This kind of strategy can be implemented by proposing simple and



ⓐ image de la page ⓑ étiquetage logique de la page ⓒ structure logique du document

Figure 7: Graph analysis for document image description

editable scenarii that can be easily experimented by the user. Some interesting issues can be found in [15], with AGORA system. Second, in the context of information retrieval, the problem is to provide user-friendly interfaces allowing the user to interactively modify the results provided by an automatic system, and such that the system is able to "learn" the requirements of the user. This aspect raises fundamental problems related to relevance feedback and incremental learning, which are probably the most difficult points (and one of the less significantly explored) of this research problems. Third, the problem of formal modelling of the user knowledge appears to be a fundamental aspect of ancient document indexing. Indeed, domain experts are generally able to express the criteria on which their requirements are based on. On this basis, the ontology based models should be developed in order to provide generic system, allowing to dynamically generate analysis scenarii. This research aspect, in interaction with relevance feedback questions is a difficult point that has to be considered in the future.

## 4. CONLUSION

This paper presents a short synthesis of different problems that have been tackled in the context of research projects funded by the French government: MADONNE and NAVIDOMASS. Some of the contributions of this research consortium have been presented and research issues have also been suggested. These orientations highlight this interest of working with other communities, since many of these suggestions concern new human interfaces, ontology modelling, and graph clustering.

## REFERENCES

[1] E. Baudrier, G. Millon, F. Nicolier, and S. Ruan. Binary-image comparison with local-dissimilarity quantification. *Pattern Recognition*, 41(5):1461–1478, 2008.

[2] A. Bensefia, T. Paquet, and L. Heutte. A writer identification and verification system. *Pattern Recogn. Lett.*, 26:2080–2092, October 2005.

[3] H. Bunke and T. Caelli. Graph matching in pattern recognition and machine vision. *Special Issue of Int. Journal of Pattern Recognition and Art. Intelligence*, 18(3):261–263, 2004.

[4] H. Chouaib, F. Cloppet, and N. Vincent. Graphical drop caps indexing. In *Graphics Recognition. Achievements, Challenges, and Evolution*, LNCS.

[5] B. Coasnon and I. Leplumey. A generic recognition system for making archives documents accessible to public. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 228–232, 2003.

[6] M. Coustaty, R. Pareti, N. Vincent, and J.-M. Ogier. Towards historical document indexing: extraction of drop cap letters. *International Journal on Document Analysis and Recognition*.

[7] N. Journet, J.-Y. Ramel, R. Mullot, and V. Eglin. Document image characterization using a multiresolution analysis of the texture: application to old documents. *IJDAR*, 11(1):9–18, 2008.

[8] Y. Leydier, F. Le Bourgeois, and H. Emptoz. Serialized unsupervised classifier for adaptive color image segmentation: application to digitized ancient manuscripts. In *ICPR 2004*, volume 1, pages 494 – 497 Vol.1, 2004.

[9] B. Naegel and L. Wendling. Combining shape descriptors and component-tree for recognition of ancient graphical drop caps. In *VISAPP'09: Fourth International Conference on Computer Vision Theory and Applications*, pages 297–302, Lisboa, Portugal, 2009.

[10] S. Nicolas, T. Paquet, and L. Heutte. Enriching historical manuscripts: the bovary project. In *Workshop on Document Analysis Systems (DAS)*, volume 3163 of *Lecture Notes in Computer Science (LNCS)*, pages 135–146, 2004.

[11] R. Pareti and N. Vincent. Global discrimination of graphics styles. In *Workshop on Graphics Recognition (GREC)*, volume 3926 of *Lecture Notes in Computer Science (LNCS)*, pages 121–132, 2006.

[12] R. Pareti, N. Vincent, S. Uttama, J.-M. Ogier, J.-P. Salmon, S. Tabbone, L. Wendling, and S. Adam. On defining signatures for the retrieval and the classification of graphical drop caps. In *DIAL*, pages 220–231. IEEE Computer Society, 2006.

[13] R. Qureshi, J.-Y. Ramel, D. Barret, and H. Cardot. Spotting symbols in line drawing images using graph representations. In *Graphics Recognition. Recent Advances and New Opportunities*.

[14] J. Ramel and S. Leriche. Segmentation et analyse interactives documents anciens imprimes. *Traitement du Signal (TS)*, 22(3):209–222, 2005.

[15] J. Y. Ramel, S. Busson, and M. L. Demonet. AGORA: the interactive document image analysis tool of the BVH project. In *DIAL'06*, page 145155, Washington, DC, USA, 2006. IEEE Computer Society.

[16] J. Salmon, L. Wendling, and S. Tabbone. Improving the recognition by integrating the combination of descriptors. *International Journal on Document Analysis and Recognition*, 9:3–12, 2007.

[17] S. Tabbone and D. Zuwala. An indexing method for graphical documents. In Flavio Bortolozzi and Robert Sabourin, editors, *ICDAR'07*, volume 2, pages 789 – 793. IEEE Computer Society, 2007.

[18] S. Uttama, P. Loonis, M. Delalandre, and J. Ogier. Segmentation and retrieval of ancient graphic documents. In *Workshop on Graphics Recognition (GREC)*, volume 3926 of *LNCS*, pages 88–98, 2006.