Enriching Historical Manuscripts : The Bovary Project

Stéphane Nicolas, Thierry Paquet, and Laurent Heutte

Laboratoire PSI
CNRS FRE 2645 - Université de Rouen
Place E. Blondel
UFR des Sciences et Techniques
F-76 821 Mont-Saint-Aignan cedex

Abstract. In this paper we describe the Bovary Project, a manuscripts digitization project of the famous French writer Gustave FLAUBERT first great work, which should end in 2006 by providing an online access to an hypertextual edition of "Madame Bovary" drafts set. We first develop the global context of this project, the main objectives, and then focus particularly on the document analysis problem. Finally we propose a new approach for the segmentation of handwritten documents.

1 Introduction

Libraries and museums contain collections of a great interest which can not be shown to a large public because of their value and their state of conservation, therefore preventing the diffusion of knowledge. Today with the development of the numerical technologies, it is possible to show this cultural patrimony by substituting the original documents by numeric high quality reproductions allowing to share the access to the information while protecting the originals. These last years, numerous libraries have started digitization campaigns of their collections. Faced to the mass of the numerical data produced, the development of digital libraries allowing access to these data and the search for information becomes a major stake for the valuation of this cultural patrimony. However such task is difficult and expensive, so requires prior studies concerning the technical means to use. Although they have a great interest for the study and the interpretation of literary works, modern manuscripts have been adressed by few digitization programs because of the complexity of such documents and the lack of adapted tools. We present in this paper the Bovary Project, a digitization project of modern manuscripts concerning especially FLAUBERT's manuscripts and we discuss the underlying principles, difficulties and technical aspects related to such a project. In section 2 we present the global context of this project and the aimed objectives. We discuss in section 3 the requirements of critical publishing and we overview in section 4 the few critical editions available! in electronic format and the projects dedicated to the digitization of manuscripts. Discussions on related problems and possible solutions are adressed in section 5, and we focus particularly on document analysis in section 6. Finally we propose a new approach for handwritten document segmentation in section 7.

2 General presentation of the project

Recently the municipal library of ROUEN has begun a program for digitizing its collections. For this purpose an efficient system of digitization allowing a high resolution display of the digitized documents has been purchased. One of the first aims of this program is the digitization of a manuscript folder compound of almost 5 000 original manuscripts issued from "Madame Bovary", a well known work of the french writer Gustave FLAUBERT. This set of manuscripts constitutes the genesis of the text, i.e. the successive drafts which highlight the writing and rewriting processes of the author. This digitization task is still in progress, and will be completely achieved in a couple of month. The final aim of this program is to provide an hypertextual edition[1] allowing an interactive and free web access to this material. Such an electronic edition will be of great interest for researchers, students, and anyone who wants to see FLAUBERT's manuscripts, especially because there is no critical edition of a full literary work of this author available on the web. This project called "Bovary Project" is a multidisciplinary project which implies people from different fields of interest: librarians, researchers in literary sciences and researchers in computer science. Flaubert's drafts have a complex structure (fig.1 left), they contain several blocks of text not arranged in a linear way, and numerous editorial marks (erasures, words insertion,...). So these manuscripts are very hard to decipher and interpret. Providing an electronic version of such a corpus is a challenging task which have to respect some requirements in order to meet the users needs.

3 Genetic edition requirements

In literary sciences, the study of modern manuscripts is known as genetic analysis. This analysis concerns the graphical aspect of the manuscripts and the successive states of the textual content. In fact the nature of a manuscript is dual. A manuscript can be considered as a pure graphical representation or as a pure textual representation. A manuscript is a text with graphical interest [1]. As modern manuscripts reflect the writing process of the author, they may have a complicated structure and may be difficult to decipher. So in a genetical edition, transcriptions are generally joined to the facsimile of the manuscripts. A transcription allows an easier reading of the manuscript (fig.1 right). One can distinguish two transcription types: the linear one and the diplomatic. A linear transcription is a simple typed version of the text, which uses an adapted coding to transcribe, in a linear way, complex editorial operations of the author (deletation, insertion, substitution) sometimes located over one or several pages. Diplomatic transcriptions respect, as far as possible, the physical layout of the original manuscript, and for this reason are helpfull to the reading of complex drafts.

Provide of a genetic folder consists in locating and dating, ordering, deciphering, and transcribing all pre-text material. A genetic publishing presents

[1] we propose a prototype at www.univ-rouen.fr/psi/BOVARY

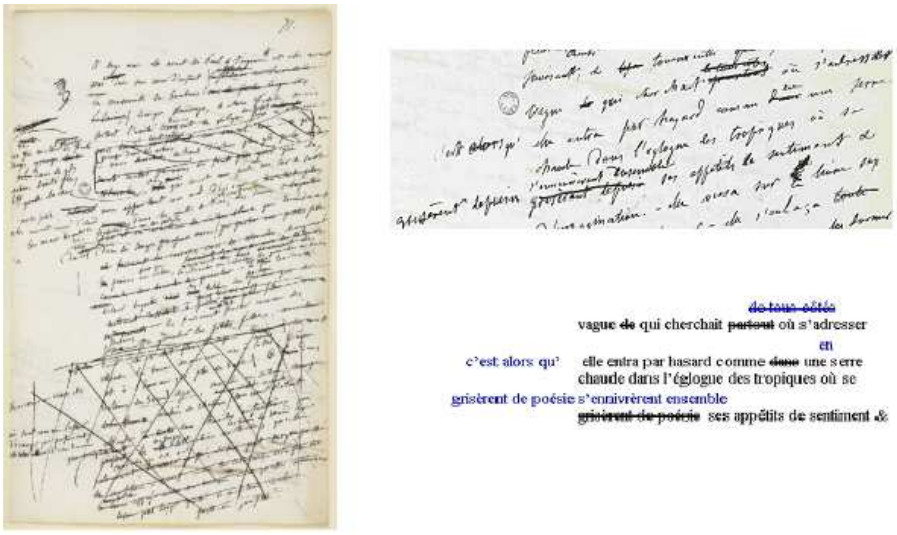

Fig. 1. (left) a complex Flauberts's manuscript. (top right) a manuscript fragment and associated diplomatic transcription (bottom right)

the results of such a work, it means an ordered set of manuscripts constituting the genesis of the text and their associated transcriptions, and allows to glance through this set of heterogeneous data. An electronic version of such a critical edition must provide the same functionnalities than a traditionnal one, but furthermore have to allow to deal easily with the textual representation and the graphical representation of the manuscripts and create links between each other , using the capabilities of structured languages and hyperlinks.

4 Related works

In spite of the development of multimedia technologies and the possibilities provided by structured languages and hypertext, few electronic versions of genetic publishing have been released up to now. This is probably due to the difficulty of such a task and the lack of professionnal tools dedicated to the manipulation of ancient or modern manuscripts. In the following we overview the existing genetic editions available in electronic format (CD-ROM or website), and we discuss the capabilities and limitations of these editions. In a second part, we overview the latest projects related to the digitization of manuscripts.

4.1 Critical Publishing

Among the numerous digitized literary works available in text or image mode on Gallica , the server of the French National Library, two thematical editions

related to the genesis of Emile Zola's work "Le rêve" and Marcel Proust's work "Le temps retrouvé" are proposed. These electronic publishing allow to visualize images of the author's handwritten notes (in black and white TIFF format or Adobe PDF files) and associated textual transcriptions in HTML. These releases contain a lot of explicative notes about the history of these works, but don't provide any capabilities of annotation or edition. They are more pedagogic editions than genetic ones. The critical edition of Flaubert's work "Education sentimentale" proposed by Tony Williams[2] is very interesting even though only one chapter of the work is adressed. In fact it is the only genetic publishing of a work of Flaubert available in electronic format and it shows the complexity of Flaubert's writing process. In spite of the wish of the designers to develop a website dedicated to a large public, its use is quite difficult for non experts of literary work. The interface is not ergonomic and no advanced image visualization and processing tools are provided. The commercial genetic edition of André Gide's work "Les caves du Vatican" available on CD-ROM [3] is certainely the most achieved critical publishing available in electronic format. It allows to visualize Gide's manuscripts and associated transcriptions. Tools for image manipulation are provided and multiple acces to the text are possible using different thematic tables (characters, places, keywords, ...) and a search engine, allowing different reading of the work, according to user's needs . As we can notice, few critical editions are available in electronic format, and generally concern material with less genetic complexity than Flaubert's manuscripts or don't deal with the full text.

4.2 Manuscripts Digitization Projects

Considering the lack of tools adapted to work on manuscripts and to the edition of electronic documents from handwritten sources, some projects have tried these last years to define the needs of the users of digital libraries and to propose different environments devoted to the study of handwritten material. Most of them have led to the development of workstation prototypes. We briefly present the projects which are similar to the Bovary project and discuss the proposed technical solutions. The problems related in PHILECTRE [3] are very similar to ours. The aim of this project was to explore the techniques needed by researchers in literary sciences, especially geneticians and medievists. In the context of this project, a workstation prototype dedicated to the edition of critical material was developed. This workstation integrates document analysis modules and interactive tools in order to provide an help for the transcription of manuscripts and to perform an automatic coupling between this structured textual representation and the manuscripts' images. However the proposed algorithm for document analysis doesn't work well with complex documents like Flaubert's drafts. Furthermore this propotype has never been used for the edition of a full literary work. Similar aspects have been adressed in BAMBI [4] in the case of ancient

[2] www.hull.ac.uk.hitm

manuscripts. The aim of this project was to provide tools facilitating the transcription of medieval texts. However the problems encountered when dealing with such documents are very different to the ones related to the work on complex modern manuscripts, in particular in the case of document analysis. The DEBORA project [5] tried to define the needs related to the use of digital libraries and electronic books. The aim of this project was to develop tools for the digitization and the access to a selection of book of the 16th century. The results of this work have allowed to provide an environment dedicated to collaborative studies on such digitized collections. This workstation is not designed for authorial manuscripts but some important technical aspects have been adressed during this project, and among them the problem of images format and compression. Once again we can notice that few projects have adressed the problem of electronic manipulation of manuscripts and they have proposed specific solutions adapted only to limited categories of manuscripts (generally medieval text).

5 Related Technical Aspects

5.1 Image Compression

One of the first problem we have to adress concerns the storage and the format of manuscripts images. In fact as manuscripts are studied for their graphical aspect, the facsimile has to be as conform as possible to the original, it means that the images must have a high quality. However a high resolution leads to files of great size (several mega-bytes), and so implies long time of loading in the case of an online edition. So it is necessary to compress the images in order to reduce the files to a size less than 1 Mo. The choice of an adapted compression algorithm is a real problem. With the GIF and JPEG formats commonly used on the Internet, the degradation on the images is too important. New algorithms like DjVu seem to be more adapted to the compression of document images, and are more and more used in digital libraries. They provide good compression ratio while preserving good image resolution, even on handwritten documents. An improved algorithm using the same principle has been developed in DEBORA for printed documents.

5.2 Indexation and Information

Retrieval The advantage of an electronic edition is to provide multiple access to the text and to allow the user to find specific information. To provide such capabilities it is necessary to proceed to an indexation of the manuscripts images. But image indexation is a difficult task. A possible solution is to provide structured transcriptions of the textual content of manuscripts. So the choice of a structured language adapted to the encoding of critical edition is necessary.

5.3 Transcription Production

The transcription of the manuscripts requires a collaborative work with the researchers and the specialists of Flaubert. At the begining of this project we

don't have any transcriptions of the manuscripts. Generally this task is performed manually using simple text editors. In order to facilitate this task and to produce a structured textual representation adapted to users' requirements, it is necessary to provide an editing environment integrating document analysis methods and interactive tools. As projects like PHILECTRE and BAMBI have shown, document analysis can bring an help for the transcription and indexation of manuscripts. Document analysis consists in extracting the structure of a document. In the case of manuscripts, it consists in extracting text lines and graphical elements like erasures. In spite of the progress made in the analysis of machine printed documents, the analysis of handwritten documents is still a challenging task. In the following section we overview the existing methods for the analysis of handwritten sources, then we discuss the limitations of these methods and propose a new approach to deal with complex manuscripts.

6 Handwritten Documents analysis

6.1 General problem of document analysis

Document analysis is a crucial step in document processing, and consists in extracting the geometrical layout of a given document from its low level representation (image). The aim is to construct a higher level representation: the structure of the document. This task involves to locate and separate the different homogeneous regions or objects of the document, and determine the spatial relations between these objects. In the case of textual documents it implies to extract the textual structure of the documents in term of sections, paragraphs, text lines, words. The structure we aim to extract is hierarchical. Such a structure can be modelized by a tree. The elements of a document constitute an object hierarchy. In fact a page of handwritten text is composed of paragraphs, which are composed of text lines, and text lines are formed by words,...
Two strategies are possible to extract the geometrical structure. The "bottom-up" strategy iteratively groups objects using local informations, starting from primitives of the document well segmented such as connected components, in order to reconstruct objects of higher level. A "bottom-up" strategy tries to reconstruct the tree representing the structure of the document, starting from the leaf of the tree (bottom) to the root (up). The "top-down" strategy on the contrary starts from the the root representing the entire document, and recursively tries to developp each branch of the tree. A third possible strategy consists in combining a bottom-up and a top-down analysis. Numerous methods using one of these strategies have been proposed for the analysis of machine printed documents. Among the most popular we can cite Kise's method [13] based on area Voronoi diagram, O'Gorman's Docstrum method[14] based on neighbor clustering and Nagy's X-Y cut [15] based on the analysis of projection profiles. These methods provide good results on printed documents, but are not directly adapted to handwritten documents, because they generally take only into account global features of the page, and are thus dedicated to well structured documents. Unlike printed documents, handwritten documents have a local struture which presents

an important variability: fluctuating or skewed text lines, overlapping words, unaligned paragraphs,... To cope with this local variability, the methods proposed in the litterature for the segmentation of handwritten documents are generally "bottom-up" and are based on local analysis.

6.2 State of the art

The few methods proposed in the literature focus in the particular case of text line extraction and are all based on a bottom-up strategy. [9] proposes an approach based on perceptual grouping of connected components of black pixels. First, the connected components of the document image are extracted. The elongated components which have a reliable direction are choosen as starting point candidats of possible alignments. Then text lines are iteratively constructed by grouping neighboring connected components according to some perceptual criteria inspired by the Gestalt theory such as similarity, continuity and proximity. Hence local neighborhood constraints are combined with global quality measures. As conflicts can appear during the grouping process, the method integrates a refinement procedure combining a global and a local analysis of the conflicts. This procedure consists in applying a set of rules which consider the configuration of the alignments and their quality. A text line extraction module using this method has been integrated in the reading and editing environment for scholarly research on literary works developped during the PHILECTRE project [10]. According to the author the proposed method cannot be applied on degrated or poorly structured documents, such as modern authorial manuscripts.
A method based on a shortest spanning tree search is presented in [7]. The principle of the method consists in building a graph of main strokes of the document image and to search for the shortest spanning tree of this graph. The main idea of this method is to assume that for each text line it is possible to find a minimum distance path that extends from one end to the other end of the line, that is to say a shortest spanning tree that spans the main stroke of the considered text line. This method assumes that the distance between the words in a text line, is less than the distance between to adjacent text lines. This method has been applied to Arabic text.
In [8] an iterative hypothesis-validation strategy based on Hough transform is proposed. The skew orientation of handwritten text lines is obtained by applying the Hough transform on the gravity center of each connected components of the document image. This allows to generate several text line hypothesis. Then a validation procedure in the image domain is performed to eliminate erroneous alignments among connected components using contextual information such as proximity and direction continuity criteria. According to the authors this method is able to detect text line in handwritten documents which may contain lines oriented in sevaral directions, erasures and annotations between main lines.
An algorithm based on the analysis of horizontal run projections and connected components grouping and splitting procedures is presented in [6]. First the image is partionned into vertical strips and then an analysis of the run projections on each strip is applied. This method allows to deal with fluctuating or skewed

text lines and to preserve the punctuation.
[12] proposes a method for line detection and segmentation in historical church registers. This method is based on local minima detection of connected components. It is applied on a chain code representation of the connected components. The idea is to gradually construct line segments until an unique text line is formed. This algorithm is able to segment text lines closed to each other, touching text lines and fluctuating text lines.
The main problem of these methods is that they generally take local decisions during the grouping process, and they sometimes fail to find the "best" segmentation when dealing with complex documents like modern manuscripts. Furthermore these methods don't use prior knowledge or don't express it explicitly, making an adaptation to different classes of documents difficult. To avoid these drawbacks, we propose to adapt the techniques used in the domain of structured document recognition, to formalize prior syntactical knowledge [16], and we propose a segmentation strategy based on the principles of Dynamic Programming.

7 A dynamic programming approach for the segmentation of handwritten documents

We proposed a new approach based on dynamic programming principles for the segmentation of complex or poorly structured offline handwritten documents, such as cultural heritage manuscripts. The main idea of our method is to use prior knowledge formalized using layout rules and an adapted search strategy which take contextual information into account. The segmentation strategy is viewed as a bottom-up grouping process directed by the search strategy. We use a state tree based formalism to modelize the bottom-up grouping process. Each state of the tree represents a partial segmentation of the document. The aim of the search strategy is to lead the grouping process in order to find the best partition of the document into physical objects.

7.1 Modelization of prior knowledge

The prior knowledge concerning the structure of the document is modelized using a grammar, that is a set of layout rules. The terms we consider in this grammar are hierarchical. They corresponds to the physical objects of the document. A paragraph is composed of text lines, text lines are composed of connected components,... The rule of the grammar highlight the grouping of objects into an object hierarchically higher. For example, text lines are produced and developped by grouping connected components. The forms of the rules are as following:

- A $\Rightarrow$ B
- C $\Rightarrow$ D E or C $\Rightarrow$ E D

Where B and E are components respectively of A and C, and D is a terms of the same hierarchical level than C.

For example, the following rule corresponds to adding a connected components to an existing text line:

- textLine ⇒ textLine + connectedComponent (1)

When applied this rule provides a new instance of the considered text line. The symbol "+" is an adjacency spatial operator.

A cost is associated to each rule of the grammar. This cost corresponds to the probability of the rule to be applied. It order to take into account some contextual information during the segmentation, the terms of the grammar are defined by a feature vector. The probability of a rule depends on the features of the terms involved by the rule. For each rule we assume that the density probability function can be estimated by a normal distribution. Given a rule r_i and a feature vector X describing the terms involved, the probability of the rule is given by:

$$P_{r_i}(X) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-m_i)^T \Sigma_i^{-1}(X-m_i)}$$

Where m_i is the mean feature vector and Σ_i, the covariance matrix for the rule r_i. These parameters can be estimated on a training set of documents. d is the feature vector dimension.

For example, if we consider rule (1), the cost depends on the features of the considering text line and the features of the connected components to be added to. The distance between the connected component and the text line, according to the average distance between components in the text line, or the skew of the connected component according to the average skew of the components in the text line, are possible features to be taken into account.

7.2 Search strategy

We have explain how to modelize the prior knowledge about the structure of the document, we have now to determine how to parse the document. The question is how to parse the data (document objects) and which rule to apply? The aim is to find the best segmentation, that is the best parse tree of the document considering the grammar, and the data. This problem can be resolved using path finding techniques. This formalism is equivalent to search for the lowest cost path from root to a goal node in a state tree. That necessits to define what is the initial state of the problem to be resolved and the final state. We assume that the root of the state tree corresponds to the initial data of the problem, that is the set of data to parse and a set of rules. A goal node represents a possible solution of the problem, that is a possible segmentation of the document and intermediate nodes of the state tree, a partial segmentation. A transition between two states corresponds to the application of a rewriting rule. The transition cost is the probability of the applied rule. The probability of a path is obtained by

multiplying the probabilities of the rules applied along this path. What we want to find is the most likely path leading to a goal. This is an optimal pathfinding problem. To solve it, the search strategy we use, is a Branch-and-Bound procedure. This method assumes to find the optimal path. At each step of the search, the best path developped so far is extended first. All the valid extensions (rewriting rule application) of the path are generated and memorized. This necessits to know exactly which rule can be applied and on which data this rule can be applied. However when parsing bidimensionnal structure, the next data to be analyzed isn't know. Theorically we must explore all the solutions by applying the rules on all the data not yet analyzed. To avoid the combinatorial explosion, it is necessary to prune some solutions. For this purpose we use a neighboring graph which allows at each step to select the data to be considered. This neighboring graph represents the neighbourhood relations between the objects of the document.

7.3 Segmentation example

To illustrate the application of the approach we consider the following example. Given a manuscript image fragment containing two skewed text lines (figure 1), we want to isolate one. For this problem, the grammar we consider is quite simple, and consists of one rule and its contrary:

- R_1: textLine $\Rightarrow$ textLine connectedComponent
- $\bar{R}_1$: textLine $\not\Rightarrow$ textLine connectedComponent

The rule R_1 represents the adding of the current connected component to the text line, and $\bar{R}_1$ is the negation of this rule.

The figure 2 represents the neighbor graph which is used to select the data during the parsing. This graph is updated each time a rule is applied. A goal is found when there is no more data to be analyzed.

Finally the figure 3 highlights the succession instanciation of the rule R_1 and the resulting extracting text line.

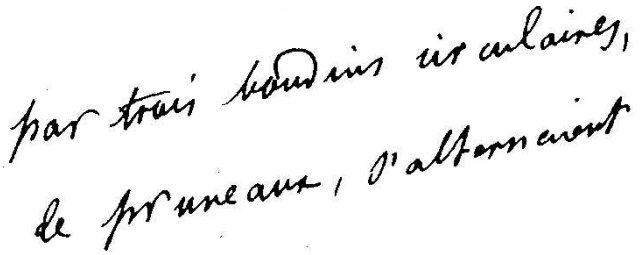

Fig. 2. two text lines of Flaubert's manuscript

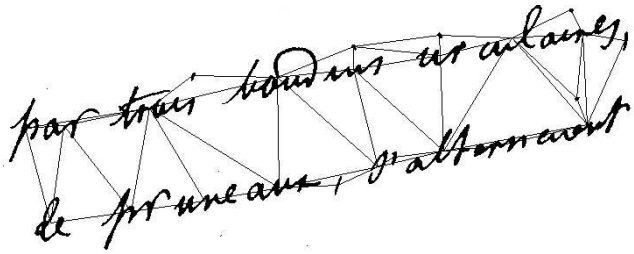

Fig. 3. Neighbor graph

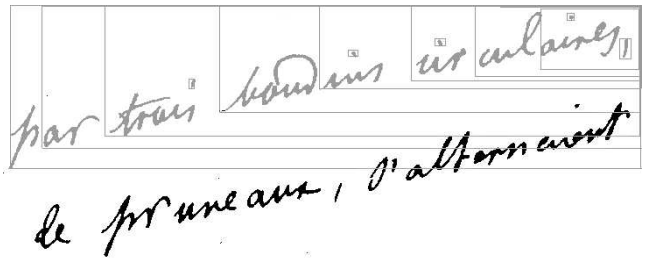

Fig. 4. successive instanciation of rule R_1

8 Conclusion and future works

We have presented the Bovary Project, a cultural heritage manuscript digitization project, and discussed the related requirements and technical aspects. As we have seen, document analysis can provide help for the transcription of modern manuscripts, and allows a text-image coupling between these structured textual representation, which is required for computerized processing, and the image representation which gives the graphical aspect of the manuscript. However no document analysis method is now able to deal with such complex and weakly structured documents. We have proposed a new approach for the segmentation of handwritten documents. This approach combines local and global analysis, by integrating prior knowledge about the structure to be segmented and local contextual features. In order to cope with ambiguities, the proposed formalism is probabilistic. The parsing algorithm makes use of the principles of path finding of the graph theory. First results on simple test problems show the feasability of the method. Future works will consist in improving the method by considering much more contextual features and using learning techniques to determine the probabilities of the rewriting rules.

9 Acknowledgments

Our work is sponsored by the "Conseil Régional de Haute-Normandie"

References

1. Andre, J., Fekete, J.D., Richy, H.: Mixed text/image processing of old documents. Congrs GuTenberg (1995) 75–85
2. Goulet, A.: Genetic publishing on CD-ROM of André Gide's book "Les caves du Vatican". Gallimard Eds.
3. Likforman-Sulem, L., Robert, L., Lecolinet, E., Lebrave, J-L., Cerquiglini, B.: Edition hypertextuelle et consultation de manuscrits: le projet Philectre. Revue Hypertexte et Hypermdia, Vol. 1, No. 2-3-4, Herms Paris (1997) 299–310
4. Bozzi, A., Sapuppo, A.: Computer-aided preservation and transcription of ancient manuscripts. Ercim News, Vol. 19, (1999) 27–28
5. Belisle, C., Hembise, C.: Etat de l'art sur les pratiques et sur les usagers des bibliothques virtuekkes. Projet DEBORA, release No 2.1 (1999)
6. Bruzzone, E., Coffetti, M.C.: An algorithm for extracting cursive text lines. ICDAR (1999) 749–752
7. AbuHaiba, I.S.I., Holt, M.J.J., Datta, S.: Line Extraction and Stroke Ordering of Text Pages. ICDAR (1995) 390–393
8. Likforman-Sulem, L., Hanimyan, A., Faure, C.: A Hough based algorithm for extracting text lines in handwritten documents. ICDAR (1995) 774–777
9. Likforman-Sulem, L., Faure, C.: Extracting text lines in handwritten documents by perceptual grouping. Advances in handwriting and drawing : a multidisciplinary approach, C. Faure, P. Keuss, G. Lorette and A. Winter Eds, Europia, Paris, (1994) 117–135
10. Lecolinet, E, Role, F., Robert, L., Likforman, L.: An Integrated Reading and Editing Environment for Scholarly Research on Literary Works and their Handwritten Sources. Proceedings of the Third ACM Conference on Digital Libraries, Witten, I., Akscyn, R. and Shipman, F.M. Eds (1998) 144–151
11. Coffetti, M.C.: Text Line Extraction from Documents Images. PhD Thesis (1996)
12. Feldbach, M., Tnnies, K.D.: Line Detection and Segmentation in Historical Church Registers. ICDAR (2001) 743–747
13. Kise, K., Sato, A., Iwata, M.: Segmentation of page images using the area voronoi diagram. Computer Vision and Image Understanding. **70** (1998) 370–382
14. O'Gorman, L.: The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. **15** (1993) 1162–1173
15. Nagy, G., Seth, S., Viswanathan, M. A Prototype Document Image Analysis System for Technical Journals. Computer. 25(7) (1992) 10–22
16. Coüasnon, B.: DMOS: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems. ICDAR, International Conference on Document Analysis and Recognition, Seattle, USA, (2001)