Processing Handwritten Documents of Historical Archives

Thierry PAQUET

PSI – FRE –CNRS 2645
Faculty of Sciences, University of ROUEN
F-76821 Mont-Saint-Aignan Cedex, France
Tel: +33 2 35 14 68 75
E-mail: Thierry.Paquet@univ-rouen.fr

Abstract

In this paper, we investigate a wide spectrum of the scientific and technological developments that are currently under investigation at PSI in the aim to provide better access to Handwritten Documents of Historical Archives. Traditionaly, Handritten Document Processing is mainly focused on word or phrase recognition which of course remains a difficult and challenging task. However, with the advance of digitization technology, libraries and literacy researchers are moving towards the use of digital document images rather than traditional paper copies of the original documents. Through two important projects concerning Historical Handwritten Documents of the French literacy patrimony, we present two new techniques dedicated to handwritten document image analysis. They concern handwritten document layout analysis using Hidden Markov Random fields, and writer identification using a graphical Information Retrieval approach. While these approaches are currently evaluted on historical documents, they are general enough to be applied to any other kind of handwritten documents, and therefore also concern industrial applications. These projects have also motivated the proposition of a new electronic format for encoding and rendering handwritten information so as to facilitate the reading of manuscripts to a human reader. The actual proposition is an XML based format called GUSTAVE_ML.

INTRODUCTION

Within the domain of document image analysis, processing of handwritten fields has received much attention during the last 40 years. Only recently, some limited but important commercial applications have been successfully designed for reading handwritten messages [1]. They concern the reading of numerical fields in forms such as personal ID number, date of birth etc… Due to the complexity of the task, reading handwritten character strings in forms is generally limited to a small lexicon. In any case, most of the time, people are asked to fill in forms using capital letters or at least a discrete handwriting style. Processing unconstrained handwriting styles has proved to be more challenging because of the variability of the graphical representation of each character from one writer to another, and also because of the possible cursive styles that allow some ligature between characters at some unknown, non deterministic positions in the message. Reading a handwritten field in

this case is not only a problem of character recognition but also a problem of character segmentation. In the 90's this problem was solved partially allowing the emergence of new applications such as automatic bankchecks or address reading, or unconstrained handwritten fields reading in forms.

It is worth noticing that despite the emergence of some specific applications dedicated to handwriting recognition in particular contexts, progress is still to be made in order to enlarge the spectrum of the possible applications of handwriting processing. A common central point among many potential applications concerns the possibility to deal with less constrained handwritten documents e.g. dealing with full handwritten texts rather than restricted alpha-numerical fields. But obviously, such a perspective will remain unrealistic for many years if we only aim at producing full ASCII transcription of handwritten documents. Improvement of the current state of the art depends on many related fields such as pattern recognition, learning theory, image and natural language processing, etc... Because such a wide multidisciplinary domain is very difficult to cover at the same time, progress in the reading of unrestricted handwritten texts will slowly evolve to reach acceptable performance for industrial applications.

Despite this rather pessimistic perspective , there exists probably more realistic and less demanding applications for which the current state of the art could contribute significantly in the near future. Following the perspective already addressed by Henry Baird in ICDAR 2003 [2], in this paper we envisage some possible applications of Handwritten Document Processing to Historical Archives. While this type of documents does not fall into the traditional spectrum of industrial applications (mainly business oriented), the increasing amount of historical documents that are beeing currently digitized all around the world requires urgent developpement of dedicated tools for indexing and retrieving useful information to end-users.

We illustrate the subject based on two projects currently under development at PSI: the Bovary project and the person identification in historical documents project. These two particular projects contribute more widely to the developpement of new tools in the field of handwritten document processing. The first part of the paper is devoted to the presentation of the Bovary project and the ongoing developpements, including interactive tools for indexing handwritten documents and handwritten document layout analysis. In the second part of the paper we concentrate on the problem of writer identification in historical documents, a general problem that has many other applications.

THE BOVARY PROJECT

Recently the municipal library of ROUEN has begun a program to digitize its collections. For this purpose an efficient system for digitizing and displaying high resolution documents has been purchased. For technical reasons, a digital camera is used. It allows to capture images of various items such as documents, paintings, collectors items, coins etc… in very good conditions of lighting.

One of the first objectives 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, e.g. the successive drafts which highlight the writing and rewriting processes of the author. This digitization task is now over. 4727 manuscripts have been digitized in color and stored using uncompressed TIFF format. Each image file occupies 48 mega bytes and the corpus is stored on a total number of 188 CD Rom. For consultation puposes on the web, compressed format will be used to allow fast access.

The final objective of this program is to provide an hypertextual genetic edition allowing an interactive and free web access to this material. Such an electronic edition will be of great interest for researchers, teachers, students, and anyone who wants to see Flaubert's manuscripts, especially because there is presently no mean to access to the original manuscripts. Indeed, for preservation purposes, only a very limited number of experts is allowed to access to the corpus. Therefore, an electronic version of a full critical edition work accessible on the internet would be of great interest all around the world. This multidisciplinary project called "Bovary Project" involves people from different fields of interest: librarians, researchers in literary sciences and researchers in computer science.

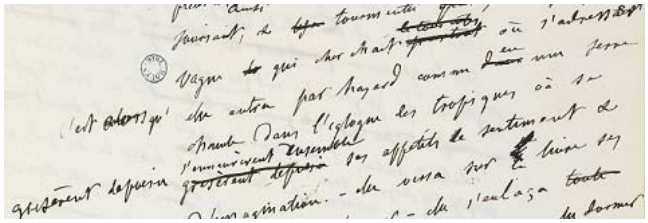
The set of 4727 manuscripts constitutes the genesis of "Madame Bovary", it means the pre-text material of this work. Flaubert's drafts have a complex structure. They contain several blocks of text arranged in a non linear way, and many editorial marks (erasures, word insertion,...) (see figures 2 and 3). Therefore, these manuscripts are very hard to decipher and interpret. Providing an electronic version of such a corpus is a challenging task which has to respect some requirements in order to meet the user's needs. It is worth noticing that the perspectives of this project encompass the limited scope of the specialists of Flauberts' literacy. Indeed, the size of the corpus together with the poor quality of Flauberts' drafts require the development of robust tools to help human transcriptors to provide useful and reusable anotations.

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 textual content. In fact the nature of a manuscript is dual. A manuscript could be considered as a pure graphical representation and as a pure textual representation. A manuscript is a text with graphical interest [3].

As modern manuscripts reflect the writing process of the author, they often have a complicated structure and may be difficult to decipher. Therefore, a genetic edition generally provides both the transcriptions and *facsimile* of the original manuscripts. A transcription allows an easier reading of the manuscript. 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 (deletion, insertion, substitution) sometimes located over one or several pages. Diplomatic transcriptions overcome this difficulty by rendering the physical aspect of the manuscript in the electronic format of the transcription (figure 1). Therefore, for the end-user, the reading of a diplomatic transcription is much more easier than reading a linear one. However, for the human transcriptor, the transcription task is much more difficult because he

has to enter not only the textual entries of the manuscript but also their presise relative positions.

(a) 

(b)

~~de tous côtés~~
vague ~~de~~ qui cherchait ~~partout~~ où s'adresser
en
c'est alors qu' elle entra par hasard comme ~~dans~~ une serre
chaude dans l'églogue des tropiques où se
grisèrent de poésie s'ennivrèrent ensemble
~~grisèrent de poésie~~ ses appétits de sentiment &

Figure 1: a manuscript fragment (a) and its diplomatic transcription (b)

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

Related works

In spite of the development of multimedia technologies and the capabilities provided by structured hypertext languages, few electronic versions of genetic publishing have been released up to now. This is probably due to the difficulty of such a task and to the lack of professionnal tools dedicated to the manipulation of ancient or modern manuscripts.

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 work "Le rêve"[1] and Marcel Proust work "Le temps retrouvé"[2] 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 the associated textual transcriptions in HTML. These releases contain a lot of explicative notes about the history of these works, but do not provide any capabilities to work on the manuscripts. They are more pedagogic editions than genetic ones.

[1] http://expositions.bnf.fr/zola/

[2] http://expositions.bnf.fr/proust/

The critical edition of Flaubert's work "Education sentimentale" proposed by Tony Williams[3] 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 expert of literacy work. The interface is not ergonomic and no tools for the study of manuscripts are provided.

The commercial genetic edition of André Gide's work "Les caves du Vatican" available on CD-ROM [4] 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 access 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 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 do not deal with the full text. For this reason, we think that the document image analysis community should contribute significantly to the development of new tools for processing handwritten document of historical archives. Such tools should include the following subjects: color image processing and compression, complex layout analysis in noisy image, handwriting recognition, coding and rendering diplomatic transcriptions in electronic format. In the following section we give a brief overview of the tools needed for processing handwritten document of historical archives.

Technical developments

Many technical aspects must be considered for improving and helping the construction of hypertextual genetic editions on the web. Some of the following propositions are currently implemented on the Bovary Project's website[4] that has been designed for demonstration purposes among the participants of the project.

Image compression of handwritten documents

The specificities of document images compared to still images are now well understood in the scientific community and efficient approaches are now available for manipulating these images. Among many other requirements, documents have to be scanned with high precision so as to be able to reproduce all the details of printed characters and graphical strokes it may contain. The required precision therefore has a direct consequence on the size of image files that must be transferred to acces a web page that contain document image (8Mb for one A4 page scanned at 300dpi in gray levels). Traditionaly, general purpose compression algorithms such as JPEG fail to compress efficiently document images without serious deterioration of the quality. Recently, dedicated compression methods have been proposed to improve the quality of compressed document images [5]. Although this method has been optimized mainly for printed documents, we found that it also compares favorably to JPEG

[3] http://www.hull.ac.uk/hitm/

[4] www.univ-rouen.fr/psi/BOVARY/

when operating on the Bovary corpus. For demonstrartion purposes, the Bovary's web site has been implemented using the two compression techniques, and we let the reader verifying by himself these considerations.

It is however difficult to generalyze these results to other handwritten corpora due to the many parameters that can influence the compression result (variability of the background, of colors, quality of writing etc…). Therefore more research efforts are required in order to generalize these compression performance to any other kind of handwritten documents.

Spatial encoding and rendering of electronic transcriptions

The automatic recognition of Flaubert's manuscripts has been rapidly excluded of the short term perspectives due to the great difficulty of the task. Indeed, manuscripts are sometimes very difficult to decipher, even by a human expert. For this reason, it has been decided to recrute a network of volunteers (experts, or just lovers of Flaubert's literacy) all around the world in order to produce diplomatic transcriptions. These volunteers are asked to transcribe around 20 pages of the manuscript and to produce a diplomatic rendering of each page, as illustrated in figure 1. These orientations have pointed out the need to develop new tools able to help the production of diplomatic transcriptions. They concern the following ones:

An interactive tool : EMMA

The definition of an interactive editor has been required to enter not only textual ASCII transcriptions (that any textual editor can do) but also spatial information concerning the relative position between textual entries so as to enable a diplomatic rendering. Furthermore, some other information concerning the graphical status of the textual entries are required such as erasure, intra-linear entries, etc… This editor was called EMMA for Annotations and Modern Manuscript Editor (namely *Editeur de Manuscrits Modernes et Annotations* in French).

An electronic format : GUSTAVE_ML

The definition of an interactive tool such as EMMA also requires an electronic format for encoding a diplomatic transcription. An XML based format has been defined for this purpose called GUSTAVE_ML. It allows the graphical characterization of textual entries as well as their spatial description within the page. The complexity of Flaubert's drafts indicates that this encoding will be compatible with most of the manuscripts that can be encountered in literacy archives. GUSTAVE_ML DTD is as depicted in figure 2.

A transformation tool

While GUSTAVE_ML is dedicated to the spatial encoding of textual entries, it is required that the transcriptions can finally be visible by transforming GUSTAVE_ML to a browsable format such as html. This is accomplished thanks to the use of a transformation tool that uses spatial tags to generate the layout of the transcription and uses the graphical tags so as to render graphical characteristics of the textual entries.

These various tools are currently under validation within the Bovary project team. It is intended that they could further be accessible on the web for the network of transcriptors, but also with the aim to promote the dissemination and use of these tools in application to other handwritten corpora. Finally, such tools should facilitate the indexation of handwritten documents, thus promoting the production of annotated data that the community of document image analysis should exploit thereafter for improving reading systems.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- GUSTAVE_ML DTD dedicated to the description of manuscripts layout -->

<!ELEMENT transcription (Bloc+)>
<!-- a transcription is made up of one “bloc” entry at least -->

<!ELEMENT Bloc (Ligne*,coordonnees)>
<!— a "Bloc" is made up of one or several “line” entries and their corresponding
coordinates -->

<!ATTLIST Bloc Num ID #REQUIRED Type (corps_de_texte | marge | bas_de_page | haut_de_page)
#REQUIRED Attribut (Normal | Biffé | Encadré) #REQUIRED>
<!-- a "Bloc" entry has for attributs one ID "Num", a required "Type" defined by  (body of
text | margin | bottom of the page | top of the page), and a required attribut among
(normal | Biffé | squared -->

<!ELEMENT Ligne (texte+)>
<!-- a "Ligne" entry contains at leat one "texte" element -->

<!ATTLIST Ligne Type (Ligne | Inter-ligne) #REQUIRED>
<!— a "Ligne" element has one mandatory attribut "Type" among (Line | Inter-line) -->

<!ELEMENT coordonnees (point,point*)>
<!— coordinates are defined by a set of points -->

<!ELEMENT texte (#PCDATA)
<!— the textual ASCII entry of a text bloc -->

<!ATTLIST texte Num CDATA #REQUIRED Type (Normal | Barré | Souligné | Illisible)
#REQUIRED>
<!— a "texte" element has for attibuts one mandatory id "Num" and one mandatory attribut
"Type" among (Normal | Erased | Underlined | Illegible) -->

<!ELEMENT point (#PCDATA)>
<!-- the "point" element contains the point coordinates (x,y) -->
```

Figure 2. GUSTAVE_ML DTD.

Layout analysis of documents with strong spatial variability

Document analysis is a crucial step in document processing, which consists in extracting the physical 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 locating and separating the different homogeneous regions or objects of the document, and determining 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. Actually 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 physical structure. The "bottom-up" strategy iteratively groups objects using local information, starting from local entities in the document such as connected components for example, in order to

reconstruct objects at a 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 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 ones, we can cite Kise's method [6] based on area Voronoi diagram, O'Gorman's Docstrum method [7] based on neighbour clustering and Nagy's X-Y cut [8] based on the analysis of projection profiles. These methods provide good results on printed documents, but are not directly adaptable to handwritten documents, because they generally take only into account global features on the page, and are thus dedicated to well structured documents. Unlike printed documents, handwritten documents have a local structure prone to an important variability: fluctuating or skewed text lines, overlapping words, unaligned paragraphs,... To cope with this local variability, the methods proposed in the literature for the segmentation of handwritten documents are generally "bottom-up" and are based on local analysis. 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 do not use prior knowledge nor do they express it explicitly, thus making the adaptation of the system to different classes of documents difficult. To avoid these problems, we currently investigate the use of stochastic approaches, namely Hidden Markov Random Fields.

Theoretical framework

Each image is considered to be produced using layout rules used by the author with spatial variability. While these rules cannot be formaly justified, it is however experimentaly verified by literacy experts that Flaubert’s manuscripts exhibit some typical layout rules as illustrated in figure 3.

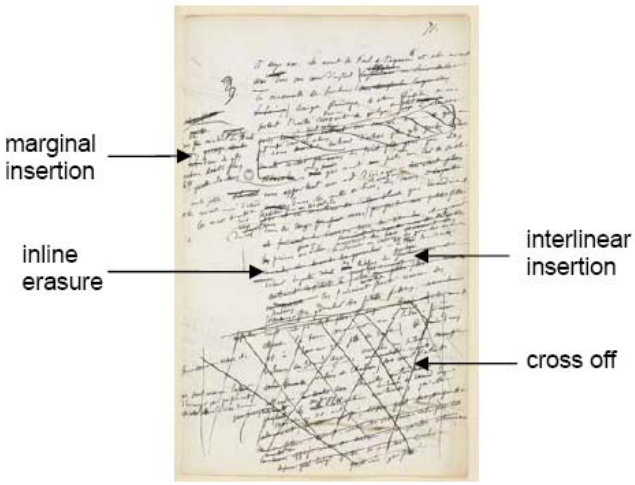


Figure 3. One exemple of Flaubert’s manuscript layout.

The image is associated with a rectangular grid G of size $n \times m$. Each site on the grid is defined by its coordinates over G and is denoted $g(i,j), 1 \le i \le n \; 1 \le j \le m$.

A neighboring system N can be defined over the grid G by the following properties:

1-) $\forall g \in G, \; g \notin N_S(g)$ 2-) $\forall \{g,h\} \subset G, \; h \in N_S(g) \Leftrightarrow g \in N_S(h)$

The definition of N over G is similar to the definition of the set of cliques C where the cliques c are the parts of G that verify one of the following properties:

1-) $\exists g \in G : c = \{g\}$ 2-) $\forall \{g,h\} \subset c \; , \; h \in N_S(g)$

The set of cliques C includes the singletons and the parts of G that contain elements that are neighbours of each other.

Most neighboring systems used are of order 1 or 2, as illustrated in figure 4 below with their corresponding cliques.

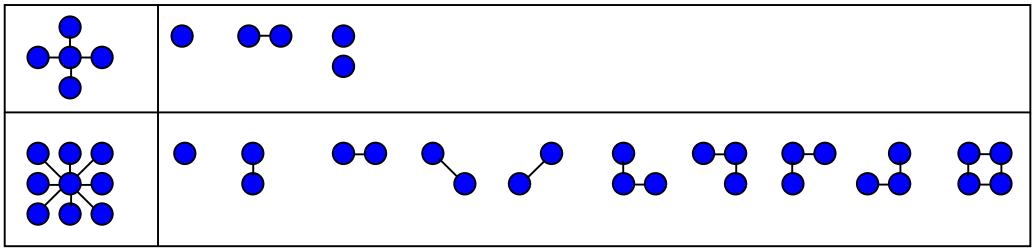


Figure 4. Neighboring systems of order 1 and 2 with their corresponding cliques.

A Random Field $X = (E, \mathrm{P}(E), P)$ is defined over the grid G by the following properties [9]:

- A random variable is associated to each site g, thus defining a set of random variables denoted by $\{S_g; g \in G\}$.

- Each random variable takes its value in $S = \{S^1, S^2, ..., S^N\}$ where S^k is one of the possible states e.g. one of the possible layout rules in our problem.

- The set $E = \prod_{g \in G} S_g$ defines the configuration space.

- A probability measure P is defined on $(E, \mathrm{P}(E))$ where $\mathrm{P}(E)$ is the set of all the parts of E.

For simplification, a random field can be considered to be the random vector $\{S_g; g \in G\}$.

A Markov random field is a random field if the following condition is satisfied:

$$\forall g \in G,\ \forall x \in E,\ P(X_g = x_g \mid X_t = x_t, t \in G, t \neq g) = P(X_g = x_g \mid X_t = x_t, t \in N_G(g))$$

Thus Markov Random Fields allow the modelization of large sets of random variables for which mutual interactions are the result of local interactions.

Now, following the stochastic framework of Hidden Markov Random Fields, the image gives access to a set of observations on each site of the grid G denoted by $O = \{o(i,j),\ 1 \leq i \leq n\ 1 \leq j \leq m\}$. Furthermore, considering that each state of the Markov Field is associated to a particular layout rule, the problem of layout extraction in the image can be formulated as finding the most probable configuation that can be associated to the image according to the model, i.e. finding:

$$\hat{X} = \underset{X \in E}{\operatorname{argmax}}\left(P(X,O)\right) = \underset{X \in E}{\operatorname{argmax}}\left(P(O \mid X)P(X)\right)$$

which results in the following formula when applying Markovian hypothesis and independence assumption of observations :

$$\hat{X} = \underset{X \in E}{\operatorname{argmax}}\left(\prod_g P(o_g \mid x_g) \prod_g P(x_g \mid x_h, h \in N_G(g))\right)$$

While in this expression the term $\prod_g P(o_g \mid x_g)$ can be computed using Gaussian mixtures to modelize the conditionnal probability densities of the observations, the calculation of the second term $\prod_g P(x_g \mid x_h, h \in N_G(g))$, which represents the contextual knowledge introduced by the model, appears to be intractable due to its non causal expression e.g. interdependance between neighboring states. To overcome this difficulty, one generally uses simulation methods such as Gibbs sampler or Metropolis algorithm [10]. Another possibility is to restrict the expression to a causal neighboring system. In any case however, finding the optimal segmentation solution requires a huge exploration of the configuration set E. This consideration is especially important because handwritten document images are particularly large. For this reason, we are currently using one intermediate suboptimal strategy which is based on region merging of the N best configurations proposed in [11].

Region merging of the N best configurations

Assume that two neighboring rectanglar regions O_1 and O_2 are associated to their respective state configuration $X_1(i,j)$ and $X_2(k,l)$ as depicted in figure 5. Then, the joint probability of region $O = O_1 \cup O_2$ and its associated state configuration $X(u,v)$ defined by:

$$X(u,v)=\begin{cases} X_1(i,j) \ if \ \ (u,v)=(i,j) \\ X_2(k,l) \ if \ \ (u,v)=(k,l) \end{cases}$$

can be derived as follows:

$$P(X,O)=P(X_1,O_1)P(X_2,O_2)I(X_1,X_2)$$

where the expression

$$I(X_1,X_2)=\prod_{g\in G_1, h\in G_2} P(x_g | x_h, h\in N_{G_1}(g))$$

denotes the interaction between the two state configurations. It is evaluated on the sites at the frontiers of the two regions as depicted in figure 5 by the horizontal cliques. As a consquence, a particular state configuration can be evaluated for the entire image by iteratively merging neigboring regions and evaluating the interaction term at the frontiers of the two regions.

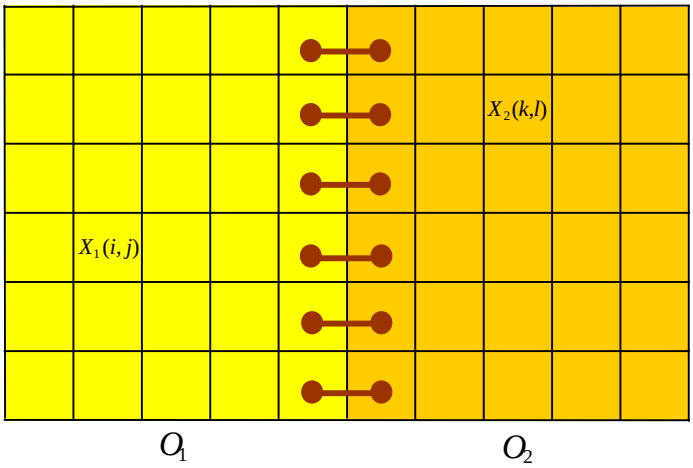


Figure 5. Two neighboring regions with their respective hidden state configuration, and frontiers sites, considering 1st order cliques.

This simple principle allows to evaluate the probability of a particular configuration associated to one image block. However, we are looking for the optimal configuration that can be associated to a document image. Ideally, this would require to compute all the possible configurations associated to each region when proceeding to a merge, and retain the optimal configuration at the end of the process. But this optimal strategy becomes rapidly intractable as soon as the size of the image exceeds a small size. For this reason we have called uppon a sub-optimal strategy that takes into account only the N best configurations when proceeding to a merge of two regions, as suggested in [11].

Various region merging strategies can be used. The aim is to start with all the single sites and then to merge regions 2 by 2 until the whole image is covered. We are currently using an

alternate strategy that consists in merging regions horizontally and vertically successively as illustrated in figure 6.

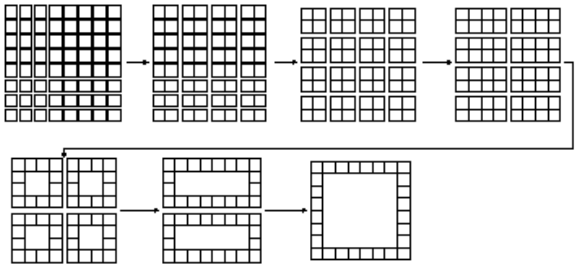


Figure 6. Alternate strategy for merging regions.

Preliminary results

We have used Markov Random Fields to proceed to the segmentation of Flaubert's manuscripts into their elementary parts, namely: text lines, erasures, ponctuation marks, inter-linear annotations, marginal annotations, just to mention the most important of them.

Observations on the grid are a 18 feature vector constituted by black pixel densities measured in the vicinity of each site (the current site and its 8 neighbours) and using 2 resolutions, thus producing 18 features.

Conditional probabilities of observations with respect to each hidden states are currently estimated using Gaussian mixtures estimated using EM algorithm [12] on a sample of labeled images, as well as conditional probabilities of neighboring states.

Figure 7 below gives some segmentation results obtained on two test images. These example images clearly show the potentiality of the method. Notably one can see how the method is able to correctly segment erasures that join two different lines. Such a result would be very difficult to obtain using other approaches based on the analysis of connected components for example. Despite these encouraging results, further developments are required in order to assess more significantly the methods. They may concern the definition of the feature set, and the influence of the various parameters of the approach.

Conclusion

As one can see, the Bovary project is a motivation for the exploration of many complementary problemes dealing with document image analysis as well as electronic formats. We have shown that the complexity of Flaubert's draft requires new approaches in the domain of handwritten Document Image Analysis. In this area the use of stochastic approaches has already demonstrated good results and robustness. This why such an orientation has been choosen. It is forseen that these oriantations will also have some interesting consequences on the actual state of the art, by improving the capabilities of layaout analysis of handwritten documents in various applications and not only historical documents.

In the next section we present another aspect of historical document processing dedicated to writer identification. Once again, the approach developed is not restricted to historical documents and may have a wider spectrum of application in the domain of person identification by their handwriting.

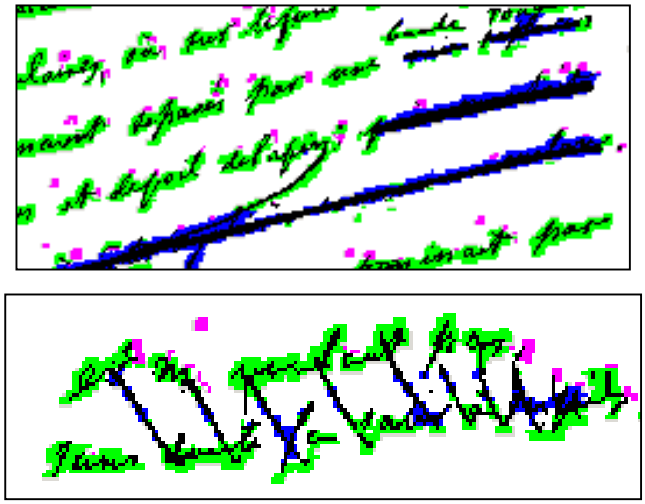


Figure 7. Two segmentation results obtained with the following color/label convention: White = Background, Green = textual component, blue = erasure, Pink = diacritic.

WRITER IDENTIFICATION IN HISTORICAL ARCHIVES

Research studies concerning handwriting analysis have mainly investigated the automatic recognition of handwritten words in various particular contexts using either temporal information (on-line recognition) or scanned images (off-line recognition) [13]. But, the advance of digitization technologies has provided interesting means for preserving historical documents while giving the possibility to view electronic *facsimile* of the originals to large worlwide communities of users (researchers, cultural tourits, …). Archives of historical handwritten documents must provide people with means to query, and browse, these documents. In the previous section of this paper, we have concentrated on browsing and indexing the textual content of handwritten documents. But handwritten documents can also be considered for their graphical contents. In this case, querying handwritten document databases can be carried out using graphical requests. One seeks for example to retrieve the documents of the database that contain certain calligraphy corresponding to one specific writer. Another possible application can deal with the detection of the various writings that may occur on a document, or the dating of the documents compared to the chronology of the work of the author. In the field of automatic handwriting analysis the task falls into the writer identification paradigm. We now give some of our investigations concerning writer identification in historical archives.

Related Works

Each writer can be characterized by his own handwriting, by the reproduction of details and unconscious practices. This is why in certain cases of expertise, handwriting samples have

the same value as that of fingerprints. The problem of writer identification arises frequently in the court of justice where one must come to a conclusion about the authenticity of a document (e.g. a will). It also arises in banks for signature verification [14], or in some institutes which analyze ancient manuscripts of authors, and are interested in the genetics of these texts, as for example the identification of the various writers who took part in the drafting of a manuscript.

As for any biometric-based identification applications (fingerprints, faces, voices, signatures…), forensic analysis of handwriting requires to query large databases of handwritten samples of known writers due to the large number of individuals to be considered. As a rule, one strives for a near 100 percent recall of the correct writer in a hit list of 100 writers, computed from a database of up to 10000 samples, which is the size of the search sets in current European forensic databases [15].

Due to the large number of classes, the identification cannot be considered as a simple classification task. Therefore, a two-stage strategy has been used to come to a conclusion concerning the authenticity of one person. The first stage is the writer identification task while the second one has been defined as the writer verification task :

- the writer identification task concerns the retrieval of handwritten samples from a database using the handwritten sample under study as a graphical query. It provides a subset of relevant candidate documents, on which complementary analysis will be achieved by the expert.

- the writer verification task, on its own, must come to a conclusion about two samples of handwriting and determines whether they are written by the same writer or not.

When dealing with large databases, the writer identification task can be viewed as a filtering step prior to the verification task. In this case, the verification task consists in matching the unknown writer with each of those in the selected subset produced by the verification stage. Therefore, the verification task can sometimes be adapted to each known reference writer based on the individual description of their handwriting. On the contrary, when the number of potential writers is too large, even unknown or infinite, an individual description of each handwriting cannot be used. In this case one can for instance derive a specific set of feature differences to model the overall within-writer and between-writer distances (intra and inter writer variability) on a set of examples [16].

Writer Identification

We present in the following subsections the various steps of our writer identification system. Figure 8 gives a brief overview of the data processing sequence. It uses three steps: a segmentation step whose main aim is to locate information that will be used to perform writer identification, then a binary feature step is used where the goal is to obtain a relevant representation for the retrieval process which represents the final step of the system. We now give full details of each processing step of our writer identification system.

Segmentation

Our method is able to cope with unconstrained handwriting thanks to this segmentation step. Up to now, very few studies have dealt with unconstrained handwriting. Yet, while perfect segmentation into characters is impossible and therefore implies sophisticated recognition procedures, the writer identification task is not so restrictive with respect to the segmentation. On the contrary, the variability in segmenting characters will bring to the method more information to characterize each writer.

First, the connected components of the document image are extracted and analyzed in order to eliminate some charts like erasures, overloaded or underlined zones which as one knows, do not characterize the handwriting, in principle. Then, the remaining connected components are segmented into graphemes. This denomination does not refer to any specific handwriting description and may be confusing. The graphemes are actually elementary patterns of the handwriting that are produced by a segmentation algorithm based on the analysis of the minima of the upper contour [17]. Figure 9 gives one example of the segmentation obtained on the french word “manuscrit”. The concatenation of two (respectively three) adjacent graphemes provides what we call bigrams (respectively trigrams) of graphemes.

Writer characterization

The writer identification task lies in the definition of a feature space common to all the handwritten documents. In a previous study, we have shown that graphemes can characterize each handwriting [17]. In this study we have extended this principle to the whole document database. Following the segmentation of the handwritten documents into graphemes, a set of binary features is defined thanks to a clustering procedure. Thus, the feature set is adapted to the database under study, and not defined in advance.

We briefly recall the main characteristics of our clustering procedure which is based on sequential clustering [18]. Unlike most of the clustering procedures such as k-means or self-organizing maps, the sequential clustering has some advantages that are suitable for our purpose. First it is a simple, fast and effective procedure to cluster a set of data points. Second, it does not require any fixed number of clusters: the total number of clusters can be unknown and is therefore dynamically determined. Finally, it does not resort to multiple iterations to converge towards the final location of cluster centroïds. Despite these advantages that make this clustering procedure well adapted for our problem, it is very sensitive to the order the data points to cluster are beeing visited. It requires therefore several clustering phases with random selection of the data points in order to be less sensitive to the initial conditions. Each of the clustering phases provides thus a variable number of clusters. The invariant clusters are defined as the groups of patterns that are always clustered together during each sequential clustering phase. These clusters constitute a set of binary features that will be used in the writer identification process. Figure 10 gives some clusters obtained on the database where gray level shows intra-cluster variability.

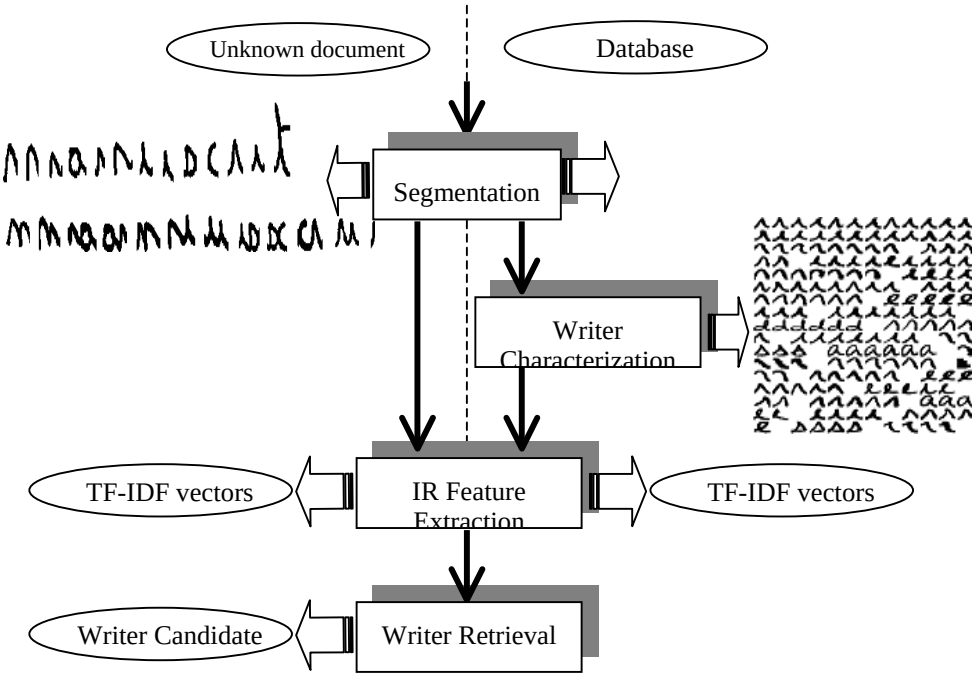


Figure 8. Writer identification organisation system

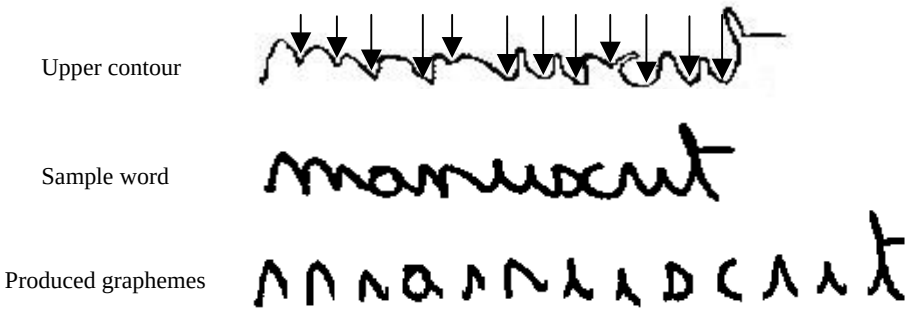


Figure 9. Potential segmentation points and final segmentation into graphemes.

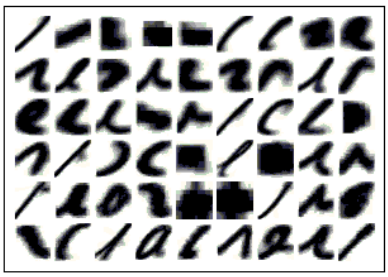


Figure 10. Some invariant clusters of level 1 obtained on the PSI_DataBase.

Information Retrieval based Feature Computation

In this study, we formulate the writer identification task within the framework of Information Retrieval. Information Retrieval is the process of finding relevant documents for a user need in a large database. The user need is expressed by a query. For this purpose the query and the documents of the database are described in the same feature space. The choice of the feature space is therefore of primary importance. As the documents must be described so as to cope with any kind of query, one cannot resort to any specific feature selection procedure that could reduce the dimensionality of the feature space. Therefore, one generally seeks to describe documents by preserving the whole set of extracted features. This leads to a description of documents in a high dimensional feature space.

The problem of writer identification can be defined as a process of finding graphical contents (set of graphemes extracted from the document to identify) in a large database of documents (set of reference documents). The retrieved documents will be ranked according to their similarity with the query. There are several types of Information Retrieval models [19]: the boolean model, the probabilistic model and the vector space model are the most popular models. Among them the Vector Space Model (VSM) proposed by Salton [20] still remains very effective [21], even though it is very simple and of a rather old design.

This model involves two different phases: a preliminary indexing phase is intended to describe each document with a high dimensional feature vector; the retrieval phase then makes it possible to evaluate the relevance of each document D_j of the database with respect to a specific query Q. According to the Vector Space Model, the relevance of each document is evaluated by the scalar product between the vector describing the query Q and the vector describing a document D_j of the database. We now present each of the two phases of the model.

Indexing phase

Assume a binary feature set has been chosen. Denote φ_i , $1 \le i \le m$ the ith binary feature. For IR purposes, each feature is all the more relevant to describe a document as it is relatively frequent in this document compared to any other document in the database. Using this principle, each document D_j as well as the query Q, can be described as follows:

$\vec{D}_j = (a_{o,j}, a_{1,j}, \ldots\ldots\ a_{m-1,j})^T$ and $\vec{Q} = (b_o, b_1, \ldots\ldots\ b_{m-1})^T$

where : $a_{i,j}$ and b_i are weights assigned to each feature φ_i, and are defined by:

$a_{i,j} = FF(\varphi_i, D_j)\ IDF(\varphi_i)$ and $b_{i,} = FF(\varphi_i, Q)\ IDF(\varphi_i)$

$FF(\varphi_i, D_j)$ is the *Feature Frequency* in document D_j. $IDF(\varphi_i)$ is the *Inverse Document Frequency* that is the inverse of the number of documents that contain this feature φ_i and is exactly defined by :

$$IDF(\varphi_i) = log\,(\frac{1+n}{1+DF(\varphi_i)})$$

where *n* denotes the total number of documents in the database and $DF(\varphi_i)$ is the *Document Frequency*, i.e. the number of documents that contain this feature.

Note that $IDF(\varphi_i) = 0$ when φ_i occurs in each document. Such features will therefore be given a null score and should indeed be eliminated from the feature set.

Retrieval phase

Each document as well as the query being described in the same high dimensional feature space, a similarity measure between a document and the query is required to provide an ordered list of pertinent documents. Many similarity measures have been proposed in the literature. Most of them are defined on binary feature vectors such as Dice, Jaccard, Okapi measures. When dealing with real valued feature vectors, a similarity measure can be defined by the normalized inner product of the two vectors e.g. by the cosine of the angle between the two vectors. Therefore, the similarity measure between a document D_j and the query Q is defined by:

$$\cos(Q, D_j) = \frac{\sum a_{i,j}\, b_j}{\sqrt{\sum_{\varphi_i} a_{i,j}^2 \sum_{\varphi_i} b_j^2}}$$

where the two terms in the denominator are the lengths of the document and of the query respectively. The retrieval process has thus a complexity of $O(TN)$, where T is the size of the feature vector and N the number of documents in the database.

Application to a Historical Document Database

Two databases have been used for this experiment. The first one (PSI_BASE) contains 88 writers that have been asked to copy one letter that contains 107 words. The scanned images have been divided into two parts: 2 thirds for the learning base and one third of each page for the test base.

The second base (ZOLA_BASE) contains 39 writers that have taken part to a correspondance with Emile ZOLA, a famous novelist of the last 19th century (1840-1902). These images have been scanned from a microfilm with a resolution of 300 dpi. They present a higher degree of difficulty than those of the PSI_BASE for various reasons: presence of noise, overlapping lines, slant, type of nib or quill used at the end of the 19th century. Finally this database contains completely free writing. The original microfilm contains nearly 700 documents. This database was first inspected and manually annotated in order to discard from the analysis irrelevant areas such as printed zones, marks, etc... Althought it contains a relatively large number of documents, they are far from being equally

distributed among writers. The number of words in a document can vary dramatically from one to another.

For these reasons, the ZOLA_BASE was designed with text blocks having a sufficient amount of information and at the same time with a sufficient number of writers. The result was thus a compromise of these two criteria. The learning base contains 39 documents each containing between 5 and 7 handwritten lines, while the test base contains text blocks with length between 3 and 5 lines. Figure 11 gives some samples of the ZOLA_BASE.

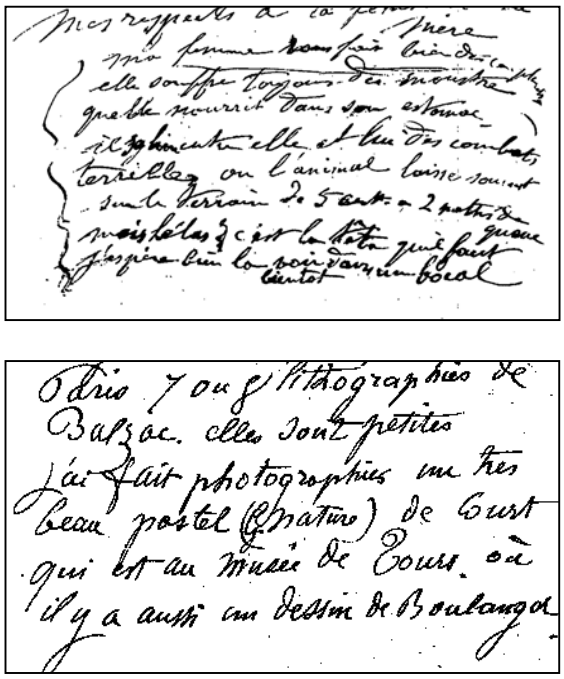


Figure 11. some samples of the ZOLA_BASE.

Due to the variability of the ZOLA_BASE it was necessary to modify our segmentation algorithm in order to operate on slanted connected components and without any knowledge of the reference lines. Therefore, the graphemes produced by this segmentation step can vary from those produced on the PSI_BASE. Figure 12 gives the segmentation results on one example.

As connected graphemes can be grouped together to produce either bi or trigram (a larger window could eventually be used), the writer identification has been carried out on these three levels. Indeed, if our previous study has shown that graphemes are good local features, it is however unclear whether concatenations of these features can better characterize a writing or not. Table 1 summerises the properties of the two databases on the three levels of analysis.

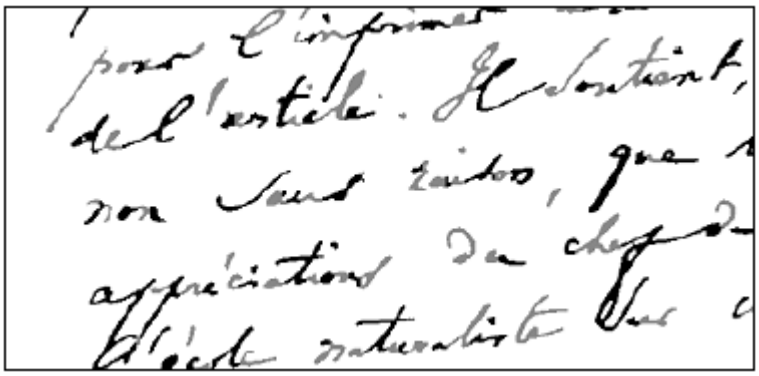


Figure 12. Graphem segmentation produced on the *ZOLA_BASE*

		Level 1	Level 2	Level 3
PSI_BASE	# graphemes	43178	25088	15953
	# binary features	7230	13876	12722
ZOLA_BASE	# graphemes	25907	15165	15165
	# binary features	3567	5585	8174

Table 1 : properties of the two databases.

Results

Figure 13 gives performance of the approach on the PSI_BASE. It shows that the correct writer is determined in 93% (83/88) of the cases using first level graphemes. The identification rate raises up to 95.45% (84/88) using bigrams as features while trigrams give only 80% (70/88) of correct identification. Let us recall that in our initial work [Ben 02] a correct identification rate of 97% was obtained on the first level graphemes but intensive pattern matching was required in this case. This first result shows that the vector space model of IR is pertinent for the task of writer identification when using local features. Furthermore bi-gram features may be even better features for the task. Two reasons can explain the lower performance obtained on tri-gram. The first one is due to the fact that tri-gram features beeing more numerous, each one of them is thus less frequent and therefore cannot be as representative of a particular writer as lower level features (bi-gram or graphemes). The second reason is that tri-gram features may be more dependent on the textual content. Therefore, while it may be a pertinent feature for the writer, its frequency may be so low (due to the low frequency of textual passage) that the size of our database does not allow to measure it.

Results obtained on the ZOLA_BASE are shown on figure 14 and are significantly lower compared to those obtained on the PSI_BASE. Particularities of this database due to the presence of noise and bad condition of digitization from micro-films can explain these results. Nevertheless, the method allows a correct identification of 93,3% (36/39) in the top 5 propositions. However bigram features are not as informative as for the PSI_BASE.

CONCLUSION

In this paper, we have covered a wide spectrum of the scientific and techological developments that are currently under investigation at PSI with the aim to provide better access to Handwritten Documents of Historical Archives. Traditionaly, Handwritten Document Processing is mainly focused on word or phrase recognition which of course remains a difficult and challenging task. However, with the advance of digitization technology (in oarticular with the use of high precision color cameras and scanners), library and literacy researchers are moving towards the use of digital document images rather than traditional paper copies of the original documents. Indeed, the quality of the digitized document images now available is a guaranty of the interest to enter into a massively perennial digitization process in the long term.

This situation requires the development of new tools allowing digital access to handwritten document images via the internet or on dedicated workstations and for various use cases. One of the most particularities of historical handwritten documents is their double interest e.g. graphical and textual, combined with a high degree of variability in the graphical representations. It is known that this variability is the major restriction for application of techniques already available for printed documents. This variability is a serious obstacle to handwritten document analysis, indexing, querying and browsing.

Through two important projects concerning Enriching Historical Handwritten Documents of the french literacy patrimony, we have presented two new techniques dedicated to handwritten document image analysis. They concern handwritten document layout analysis, and writer identification. While these approaches are currently evaluted on historical documents, they are general enough to be applied to any other kind of handwritten documents, and therefore also concern industrial applications.

These projects have also pointed out the requirements of new electronic formats for encoding and rendering handwritten information so as to facilitate the reading of manuscripts to a human reader. The actual proposition is an XML based format called GUSTAVE_ML, the development of which still requires a validation phase before diffusion.

Acknowledgement

The BOVARY project is sponsored by a research grant from the Regional Council of *Haute Normandie*, France, that has been attributed to Stéphane Nicolas, PhD student. The Person Identification project on historical documents is sponsored by CNRS, France, within the Information Society Program, and has been carried out by PhD Ameur Bensefia. The author is also particularly grateful to Mahassen Khemiri, Yousri Kessentini and Mohamed Sellami for their contributions to the development of software tools and BOVARY's web site. The author is also grateful to Professor Laurent Heutte for his contribution to these projects.

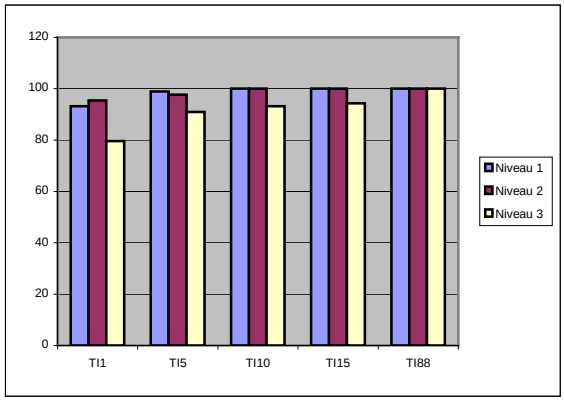


Figure 13. Writer Identification on the *PSI_BASE*

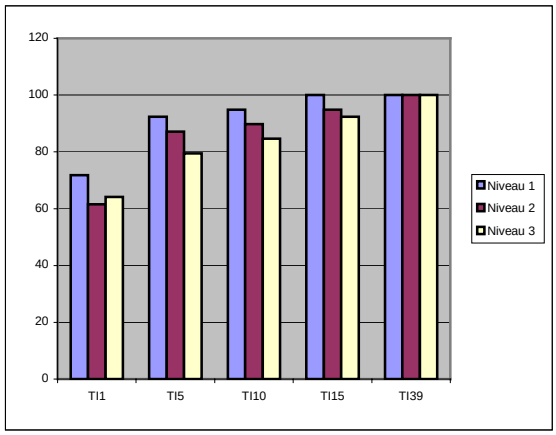


Figure 14. Writer identification on the *ZOLA_BASE*

BIBLIOGRAPHY

[1] H. Bunke, Recognition of Cursive Roman Handwriting, Past, Present and Future, International Conference on Document Analysis and Recognition, Endinburgh, 2003.

[2] H. Baird, Titre?, International Conference on Document Analysis and Recognition, Endinburgh, 2003.

[3] J. André, J.D. Fekete, H. Richy. Mixed text/image processing of old documents, In congrès GuTenberg, pages 75-85, 1995, France.

[4] A. Goulet, Genetic publishing on CD-ROM of André Gide's book "Les caves du Vatican". Gallimard Editor, France.

[5] L. Bottou, P. Haffner, P.G. Howard, High Quality Document Image Compression with DjVu, Journal of Electronic Imaging, Vol. 7, n° 3, pp. 410-425, SPIE, 1998.

[6] Kise, K., Sato, A., Iwata, M., "Segmentation of page images using the area voronoi diagram", Computer Vision and Image Understanding", Computer Vision and Image Understanding, Vol. 70, No. 3, pp. 370-382, June 1998.

[7] O'Gorman, L., "The document spectrum for page layout analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No 11, pp. 1162-1173, November 1993.

[8] Nagy, G., Seth, S., Viswanathan, M., "A Prototype Document Image Analysis System for Technical Journals", Computer, Vol. 25, No. 7, pp. 10-22, July 1992.

[9] Chellappa, R. et Jain, A., editors (1993). Markov Random Fields - Theory and application. Academic Press.

[10] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE transactions on Pattern Analysis and Machine Intelligence, Vol. 6, 1984.

[11] E. Geoffrois, Multi-dimensional Dynamic Programming for statistical image segmentation and recognition, International Conference on Image and Signal Processing, 2003.

[12] J. A. Bilmes, A gentle Tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models, Department of Electrical Engineering and Computer Science, Berkeley, 1998.

[13] R. Plamondon, S. N. Srihari, On-Line and Off-Line Handwriting Recognition : A Comprehensive Suvey, IEEE-PAMI, Vol. 22, N° 1, pp. 63-84, 2000.

[14] R. Plamondon, G. Lorette, “Automatic signature verification and writer identification – the tate of the art”; Pattern Recognition, vol. 22, n°2; pp 107-131, 1989.

[15] L. Schomaker, M. Bulacu, Automatic Writer identification usinf connected-component contours and edge-based features of upper-case western script, IEEE-PAMI, Vol. 26, N° 6, pp. 787-798, 2004.

[16] S. Srihari, S.H. Cha, H. Arora, S. Lee, “Individuality of Handwriting : A Validity Study”, Proc. ICDAR’01, Seattle (USA), pp 106-109, 2001.

[17] A. Nosary, L. Heutte, T. Paquet, Y. Lecourtier. “Defining Writer's Invariants to Adapt the Recognition Task”, Proc. ICDAR'99, Bengalore (India), pp 765-768, 1999.

[18] M. Friedman and A. Kandel, "Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches", chapter 3, pp. 55-98, Imperial College Press, 1999.

[19] F. Song, W. Bruce Croft, “A General Language Model for Information Retrieval”, Eighth International Conference on Information and Knowledge Management (ICIKM'99), 1999.

[20] G. Salton, A. Wong, “A vector Space Model for Automatic Indexing”, Information Retrieval and Language Processing, pp 613-620, 1975.

[21] D. Feng, W.C. Siu, H.J. Zhang, “Multimedia Information Retrieval and Management”. Springer Edition, 2003.