Computer assistance for Digital Libraries: Contributions to Middle-ages and Authors' Manuscripts exploitation and enrichment

V. EGLIN, F.LEBOURGEOIS, S. BRES, H. EMPTOZ, Y. LEYDIER, I. MOALLA, F.DRIRA

LIRIS INSA de Lyon, Villeurbanne cedex – France - veronique.eglin@insa-lyon.fr

Abstract

In this paper, we are interested in digitized middle-ages and 18th century authors' manuscripts analysis for the realization of suitable assistance tools dedicated to humanists and historians. The purpose is to help them in their intuitive and empirical work of information retrieval (word retrieval, identification of writers, writing styles classification…) and to enrich ancient manuscripts with additional descriptions which are based on writing styles and shapes analysis. In this project, we intend to create efficient software that analyses images contents and automatically extracts necessary information for writing indexation. This work can be done by adapting our expertise in document image processing to the domain of middle-ages and authors' manuscript. The development of such dedicated solutions is a complex task which reaches today technological limits due to the difficulty to automatically process a growing mass of digitized images of handwriting documents from different origins.

1. Introduction and context

In this project, we are interested in digitized manuscripts analysis so as to provide suitable tools to help historians and experts in their comprehension of different periods of the history, of different cultural inheritance and traditions which are strongly correlated to the origins of the documents. Our intervention consists in proposing adapted ergonomic tools which are based on the exploitation and enrichment of the documents images. We propose here to present a software plate-form to analyse digitized pages for handwriting information retrieval, for page indexation, for writing style identification, and in a further way for writings authentication and dating (in forecast for an entire book genesis). The finality of this scientific work is to propose a suitable assistance by adapting our expertise in document image processing to the domain of middle-ages (from Europe and Africa) and 18th century authors' manuscripts and by developing new methods and new techniques with a well-adapted characterization of different classes of handwriting shapes. The metadata which have been extracted from those pages, tried to take into consideration the different aspects of the user's interest, the history of the manuscripts, the codicology, the paleography, the pages composition. The originality of this work is to propose a synthesis of different pieces of works that have been made on handwriting documents in our laboratory and which can be gathered here to contribute to the development and the improvement of digital libraries. Numerous research projects on digital libraries are dealing with the future tools of libraries for query, retrieval, analysis, management, accessibility, usage, archiving and preservation of information, [8]. This work has been supported by different digitization and valorisation projects (ACI Masses de données: *MADONNE* Project, [1], CNRS *Patterns and Colors* Project, [10], [11], UNESCO Mali Project, *Montesquieu's Secretaries Authentication* Project[1], [7], [23]).

1.1 Image based digital libraries

A growing part of Digital Libraries (especially the libraries which are dedicated to handwriting documents) are accessible in image mode because old manuscripts cannot be automatically recognized by specialized tools (like OCR and transcribers). For example, manuscripts indexing needs great expertise that a computer vision system cannot achieved properly. But image analysis can retrieve useful

[1] ACI MADONNE - ACI Masse de données of the pluridisciplinary thematic network, RTP-Doc, 2003-2006. « Patterns and Colors » Project, interdisciplinary program « Information Society», 09/2003-02/2005. « Montesquieu's secretaries authentication Project », with C. Volpilhac-Auger, ENS Lettres, Lyon, 2005.

information which can help documents indexing. In this paper, we will only consider documents in their bitmap representation. It is important to consider bitmap format, because most of users prefer an access to documents in their original layout rather in their ASCII format. The image layout contains the overall page appearance, the background texture (even if it is noisy and degraded), the authentic and original ink and ornaments colors, drawings contents, annotations, strokes lines, and all typical marks that can identify the manuscripts or recognize the authors.

Among those manuscript collections from different historical periods, it has been necessary to process handwritings which have suffered from intensive use and natural time consuming. Consequently, all those degradations have led to resulting poor visual qualities of documents, with multi-writer annotations or corrections, with background spots and coins, with different delocalized folds and asperities (see figures 1 to 4), and with ink sips through the pages and colour degradation due to long periods of storage for medieval collections. The assistance tools which have been developed for those collections have been conceived with a great care to those degradations and specificities.

1.2 A rich, rare and diversified Cultural inheritance

Historical documents are the mirrors of civilizations and cultures: they are the fundaments of our patrimony. It is essential to preserve them and to find the adapted tools to access them. This work is dealing with different kinds of manuscripts from different geographical places (essentially from Africa and Europe) and periods (from the 10th to the 19th century).

For medieval manuscripts, we have worked with experts from the IRHT[2], the French Institute on Texts History. This Institute performs fundamental research on ancient manuscripts anterior to the 14th century. We have chosen to extract from those documents specialized metadata that can be automatically retrieved by an image analysis system. The system has been developed so as to help the collaboration with the archivists and historians and make them familiar with pattern recognition. It also provides the opportunity to understand user's needs. Examples of Latin and Arabic middle-ages documents are given in figure 1.

The most important role of the manuscripts is to support research in the domain of History and science. African manuscripts have for example something to do with African identity which has been denied in the past because of the lack of writings. In addition to reflecting African culture and civilization which are different from those of the rest of the world, they constitute the proof that Africa is not only the continent of oral civilisation.

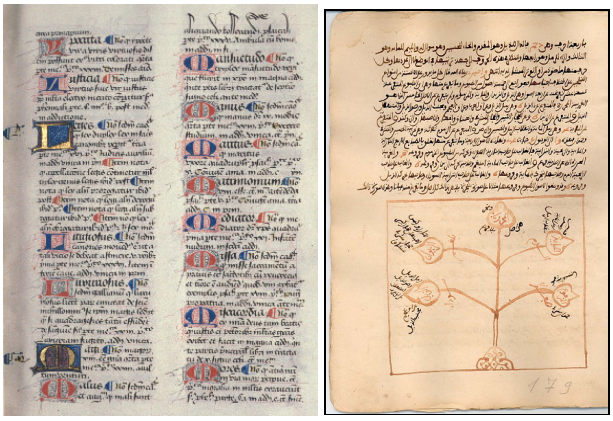


Figure 1. Medieval manuscripts showing illuminated paintings, initials, decorations, and decorated characters. Middle-Ages Arabic manuscript from Mali.

For contemporary authors' manuscripts, we have worked on a selection of documents which have been extracted from gigantic patrimonial collections which gathered numerous kinds of writings of 18th and 19th centuries. For this period, we are interested in famous French authors' manuscripts which are at the origin of rare collections and stored in libraries recently associated to innovating digitization projects. A typical example is given by the Montesquieu's collection (see figure 4). His manuscripts are characterized by a great diversity of writers: they were all written by more than twenty secretaries having different visual characteristics of writings, [23].

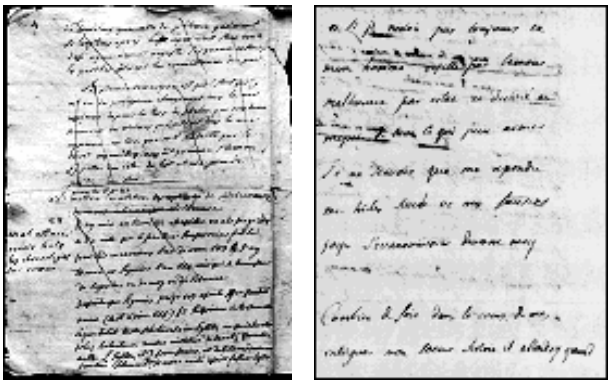


Figure 2. Drafts of 18th century authors manuscripts: Montesquieu's autograph ("De l'Esprit des Lois", 1789) and a secretary (1780).

[2] IRHT: http://www.irht.cnrs.fr

2. Meta-data and specificities of ancient manuscripts

Most of documents from cultural heritage have specific metadata, but their indexing requires an expensive and long intervention of experts. The manual extraction of the metadata from all images takes enormous work and time to accomplish. The development of understanding systems to process automatically all type of digitized documents is also difficult because each application need very specific developments on demand, [11]. The solution consists in collaborating with libraries and end-users so as to define precisely the information to retrieve for images indexing and image content description. It is important to notice that the complexity and the variability of those metadata (relative to the manuscripts origins) make difficult to recognize them with a single model. In that context, a computer assisted extraction of metadata is an important issue for the development of digital libraries.

2.1 The Middle-ages manuscripts

For searchers, the direct accessible visual unit in an open book is the page, or the double page. The text is initially structured in a double page symmetrically by the ruling and the justification. An isolated page is necessary asymmetric, see figure 3.

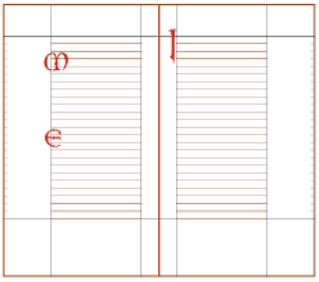


Figure 3. Ruling scheme for folios « La Trinité de Vendôme », 1129-1132 *(IRHT)*.

The physical layout is the simplest information needed for finding the book structure. The metadata which are interesting for medieval manuscripts are the text baselines which guide the copyist in writing, the number of the text lines, the number of columns and text justification. The page layout is very interesting to understand the page content, the reader interest, and the effective uses. The page layout can be complex and can contain different ornamental and illuminated objects (gold ornaments, miniatures with pictures or scenes painted on the page, initial with a large initial capital letter usually painted in a contrasting colour or illuminated in gold) which have all a pregnant signification toward the historical period. Handwriting shapes are in constant evolution during middle-ages periods: copyists gestures are regularly changing, his arm and hand position, his material, the used models (characters and ornaments shapes) and the kind of text that he must produce. All those factors are varying and impose to the copyist new constraints which influence the characters drawing from the Antiquity to the Renaissance periods. In that context, palaeographic works are aiming at gathering manuscripts into visual classes. This cataloguing is showing the graphical evolution of writing styles and can be used for writing dating. Middle ages documents are characterized by pregnant colours. It becomes possible to separate different text regions with their colours in an alternation of red, blue, green and black. The study of colour spectrum is also important for research; however, it is achieved by physical and chemical analysis. Nevertheless, a costless characterization of colours can be achieved by image analysis if the camera is well calibrated and colour references stored. A synthesis of the most frequent metadata is illustrated in figure 4.

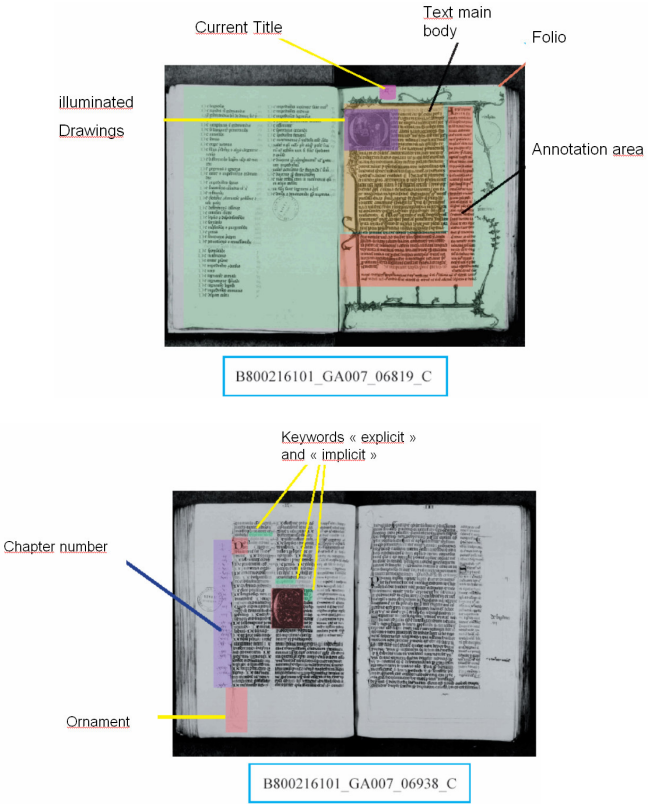


Figure 4. Examples of specific metadata for Middle ages Latin manuscripts

For Arabic documents, we can notice the following meta-data: Illuminated objects and ornaments, physical layout, main text body, illustrations, table location, punctuation, titles, circular or zigzag writings… see figure 5. With this great diversity, it becomes necessary to develop specific techniques and well adapted tools according to the specificities of each kind of encountered documents class.

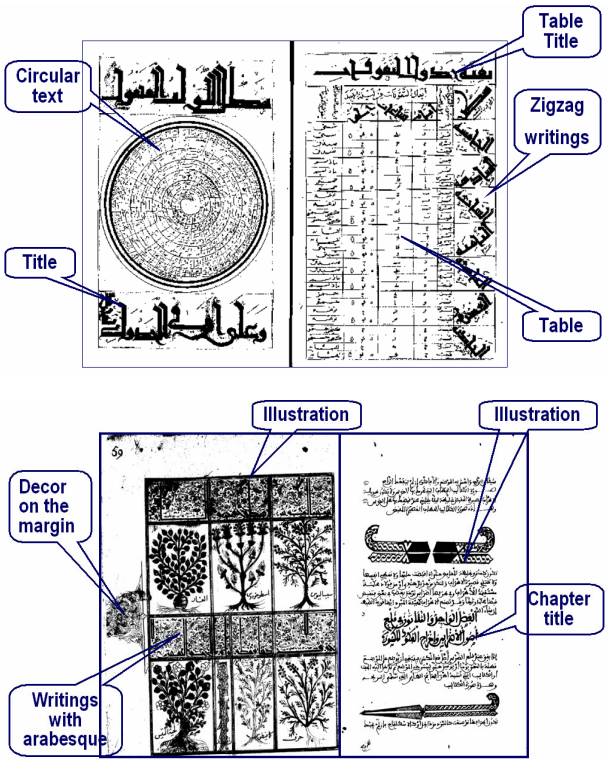


Figure 5. Examples of specific metadata for Arabic manuscripts.

By considering the great diversity and complexity of meta data extraction, w are widely used the spectral domain of handwritten images for frequencies bands decompositions (Hermite transforms and Gabor bank filtering for selective text information extraction) but also the differential analysis (Gradient based features), mathematical morphology and different classification schemes.

2.2 The 18th century authors' collections

Written contemporary authors' documents contain a wide variety of metadata like text contents, documents layouts, typography, logical structure, author authentication, written annotations, writer discrimination (for multi-writer document), but also non-textual information like signs, drawings, sometimes but rarely ornaments.

In the context of authors' manuscripts, we must notice some typical meta-data which are generic to handwriting drafts: the text non linearity, the presence of multi oriented text areas, the existence of frequent marginal annotations and irregular body paragraphs. The main point concerning those documents is linked to their unpredictable page layout which can not be modeled by any formal representation and which can be sometimes very complex, see figure 6.

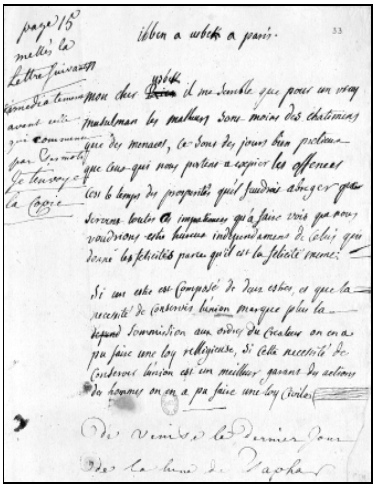


Figure 6. Autograph page of Montesquieu corpus: presence of multi-writer annotations (1750).

The second specificities deal with the inherent irregularities of shapes, with variable interlines spaces and frequent words contacts. The kind of writing device, the ink type and the manner they are employed can transform the global aspect of a curve or a character. Different tools that are based on spectral domain of images or on perceptive approaches can find here a real sense (fractal measures, Gabor and Hermite transforms, wavelets coding…). For those reasons, the system must take into consideration both texture properties of handwriting and local variations all along shapes and graphemes contours.

2.3 Degraded and noisy ancient manuscripts

Many digital images of documents and more generally ancient manuscripts are degraded by the presence of strong artefacts in the background, [2], see figure 7. This phenomenon can affect the readability of the text and, in the case of writing style classification, it compromises a relevant handwriting characterization. Background artefacts can arise from many different kinds of degradations such as scan optical blur and noise, spots, underwriting or overwriting, time wearing, intensive use or bad preservation conditions, humidity,

marks resulting from the ink that goes through the paper generally called show-through or bleed-through or interference strokes effects, strokes of pen and underlines are features that, in general, are unwanted.

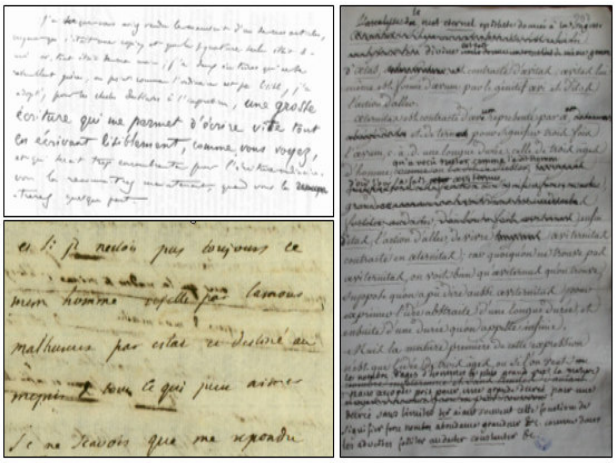


Figure 7. Examples of typical noisy background (multi spots, ink of the reverse side, granularity of the support), [23].

For example, removing the underlines is important as the text can be efficiently segmented and recognized. These kinds of degradations are simulated by new layers at different grey level that are superposed to the original document image.

We focus here on different damages due to seeping of ink from the reverse side, ink degradation (attenuation of the ink marks which compromises a correct text reading) or palimpsests (an earlier text has been erased and the vellum or parchment reused by another writer). In that last case, we need a digital enhancement technique to recover the erased text.

3. Our contributions

3.1 "Segmentation free" approaches: a new way for degraded handwritings analysis

Traditional regular handwriting segmentation approaches have shown their inefficiency in ancient heterogeneous or degraded corpus. What is more, we have noticed that because of the variability of medieval manuscripts layouts, it is necessary to privilege bottom-up (data-driven) methods which do not require prior knowledge. In that context, we can notice that a bi-level page segmentation generally damages the handwritings by merging words or text lines together or by enhancing some background spots, see figure 8. Traditional regular handwriting segmentation approaches become inefficient in ancient degraded corpus (connected components analysis, directional filtering like RLSA, lines and columns projections profiles, [12]). In that context, we can notice that a thresholding step generally degrades the handwritten regions by merging words and text lines together. So as to avoid those difficulties, we have chosen a segmentation free approach which leads to a selective mapping of the page image and which is guided by textual informative areas. We have chosen to characterize handwritten text by considering visual emergent properties in the way of human visual expertise, like Nosary in [17].

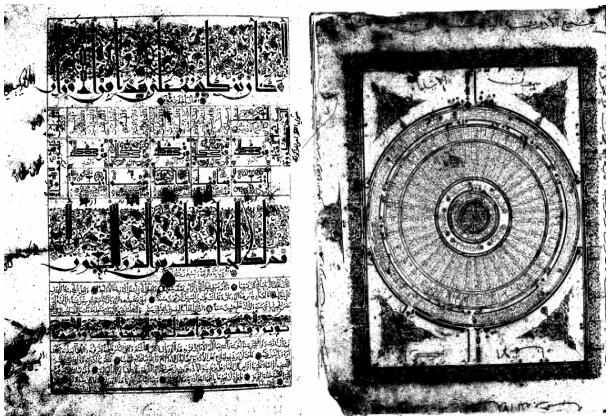


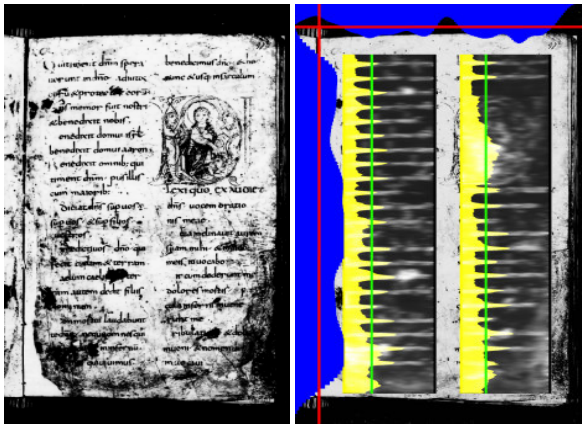
Figure 8. Degraded binary Arabic documents

The originality of our overall point of view comes from the consideration of texture properties of handwriting (for example with the Hermite based analysis) and local oriented variations all along patterns contours (which are revealed by the Gabor processing and gradient features extraction). Some similar approaches have been proposed by Kuckuck in [9] and developed later by Said in [20] with multi-spectral text images decomposition. In this paper, we will propose different segmentation free approaches which will enhance the writing contrasts, suppress residual background noises, extract words and lines segments, classify handwritings and retrieve similarities in writing shapes.

We take the layout analysis to illustrate the "segmentation-free" philosophy. We have already shown that layout segmentation by using thresholding techniques and connected components analysis are not suited for Middle Ages Latin and Arabic manuscripts [13]. Because medieval manuscripts cannot be threshold properly, we analyze directly colors or gray levels of images using differential or frequencies based approaches, texture analysis, scale-space theory,

mathematical morphology or perceptive models. For example the segmentation of text line by using scale space theory [15] is far more efficient for complex images with different qualities than other classical methods. Same approaches can be used to find the main text body and columns (see figure 9)

9.1



9.2

Figure 9. 9.1. Text line extraction using scale space theory. 9.**2.** Columns extraction and main body localisation using scale space theory

3.2 Ancient documents denoising

3.2.1 Hermite based denoising

The main interest of a denoising step in the context of historical documents processing is to suppress information coming from the background which will modify the following analysis. Using denoised image allows focusing on the handwriting by itself. This method, in its principle, is very close to method using wavelets decomposition for restoration like the one presented in [21]. In this part, we present a method for images of patrimonial documents cleaning. It is based on a polynomial transform, the Hermite transform [19], [16], which models the set of channels describing the HVS. In general, a polynomial transform decomposes locally a signal into a set of orthogonal polynomials with respect to the window used for localizing the signal. Hermite transforms are exploited here to decompose the initial signal into different parts depending on their frequencies characteristics (high or low). Most of the time, the noise or degradations we can see on ancient documents have low frequencies characteristics, while the writing by itself is composed of higher frequencies. If we take into account even higher frequencies, we can extract the contours of the writings, with a level depending on the original contrast. Thus, if we assume that high frequencies with sufficient high level represent the contour of the writings we want to keep, and low frequencies regions (in background regions) mostly contain degradations, it is of a great interest to separate them. Keeping the first one and suppress the second one will then restore visual pages appearance. This is exactly what we want to achieve with the Hermite transform. The formal definition of Hermite transforms are synthesized in [19]. Hermite properties allow a much smoother reconstruction of the original signal or image after filtering, without block effects or discontinuities we observe in wavelet based reconstructions.

Figure 10 presents an example of discrete Hermite transform up to degree 2.

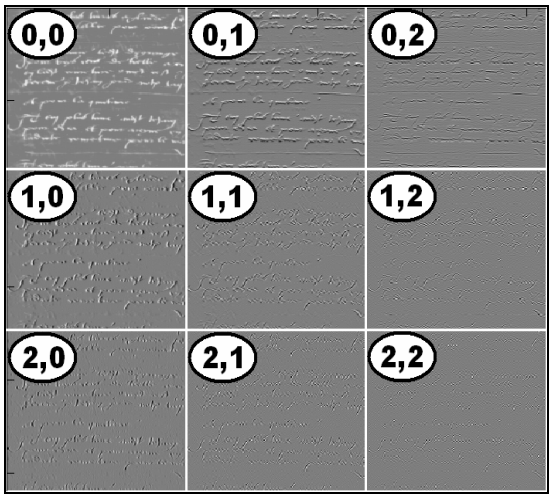


Figure 10. 2D Hermite filters using 7x7 windows and up to degree n=2 for the rows and the columns

The quadrant (0,0) is equivalent to a Gaussian filtered image. The other quadrants correspond to higher frequencies. In this case, the image is analysed using 7x7 windows, and consequently, the complete

decomposition contains 7x7 quadrants. The analysed frequencies are thus relatively high frequencies of the original image. Middle grey values correspond to zeros, blacker values are negative and whiter values are positive. As we explain earlier, the lowest frequencies contain information on the background and high frequencies contains information on the writings we want to keep, if their levels are sufficient (above a certain threshold) because we suppose them having higher contrast than noise and higher frequencies (thin lines, contours …). The Hermite based denoising process uses this decomposition. In a first step, we localize the writing areas using the energy contained in quadrants (1,0) and (0,1) (see figure 11).

Figure 11. Example of document denoising. **11.1.** Original image. **11.2**. Energy mask used to localize writings. **11.3.** Denoised document .**11.4.** Detail of the denoised document . **11.5.** Another example. **11.6**. Its denoised version.

This information is very close to a gradient energy. The second step uses the normalized energy map *M* (values between 0 and 1) as a mask to filter all the quadrant of the decomposition. The normalized energy map gives, for each position, a probability to contain writing. An example of normalized energy mask is given on image 11.2. We reconstruct the image using the threshold Hermite quadrants. We obtain an image with a cleaner background. Figure 11.3 presents the example of image after denoising by Hermite reconstruction (details on figure 11.4. An other example, coming from a synthetic degradation of image 11.1 is presented on figure 11.5 and 11.6 presents its denoised version. In this case, simulations of ink spots were added.

3.2.2 Recursive PCA based cleansing approach

In this section, we propose to apply "ink bleed-through" removal on the degraded document images. To do that, we propose to use a segmentation approach. In fact, the main idea behind our algorithm is to classify the pixels of the page into three classes: (1) background, (2) original text, or (3) interfering text. This last class must be removed from the original page and replaced by the background colour (the average of the detected background pixels for example). Nevertheless, a single clustering step is not sufficient to correctly extract the text of the front side, see figure 12.

Figure 12. Results of the 3-means classification algorithm. **12.1.** An extract of a degraded document image. **12.2.** Image class n°1, **12.3.** Image class n°2. **12.4.** Image class n°3. **12.5.** ancient manuscript with "bleed-through" interference provided by the archive of "Chatillon-Chalaronne" . **12.6.** Restored image.

Thus, we propose to apply a recursive segmentation approach on the reduced and decorrelated data. Our proposition relies on two types of analysis: the Principal Component Analysis and the k-means algorithm applied recursively on each generated data image. In that context, we can quote Tonazzini's work in [22] who works on independent component analysis for documents restoration. To simplify the analysis and reduce its computational complexity, we consider the case of a two-class problem: original text or not. The proposed method is built then via recursively dividing the test image into two subsets of classes. Experiments were carried out on many samples of these images to evaluate the performance of our approach. Figure 12.6 illustrates an example of the restored image resulting from the application of our method on the degraded document image 12.5.

3.2.3 Adaptative colour image segmentation

Colours are important metadata which must be retrieved to localise coloured text or illuminated marks. We have developed for the CNRS project "*Patterns and Colours"* an adaptive segmentation algorithm suited for colour document images analysis [13].

Document image processing is generally better achieved by adaptative segmentation methods. It has been previously shown that adaptative thresholding algorithms, such as Niblack's or Sauvola's, perform better than non-adaptative ones like Otsu's or Fisher's. It is explained by the specific defects found in document images, such as stains, illumination variation and colour fading which need to be processed by a highly adaptative algorithm. The principle of our method consists to classify colours into a sliding window so that the segmentation is neighbour-dependent and allows the classifier to slightly adapt the colour clusters to any new local colour information. The dependence between successive classifications is justified by the fact that most of the information is similar in overlapping windows. We choose to serialize an unsupervised *k*-means algorithm that is applied sequentially by using a sliding window over the image. The serialization of the *k*-means has three properties:

- An important reduction of the amount of iterations to stabilize the centres of clusters.
- A higher adaptability of the segmentation as the centre of each cluster can move swiftly.
- Every cluster does *not* have to be represented in each window.

From the user point of view, the serialized k-means algorithm has several drawbacks: It requires the number K of colors to retrieve and color samples for each class during the initialization of the algorithm. But this method has been tested successfully on different type of images like digitised colour manuscripts, video images and multiple natural and non-natural images having heavy defects and showing illumination variation and transparency. The proposed algorithm is generic enough to be used on a large variety of applications such as colour segmentation, image thresholding, segmentation of dithered images, ink bleed-through" removal, see figure 13.

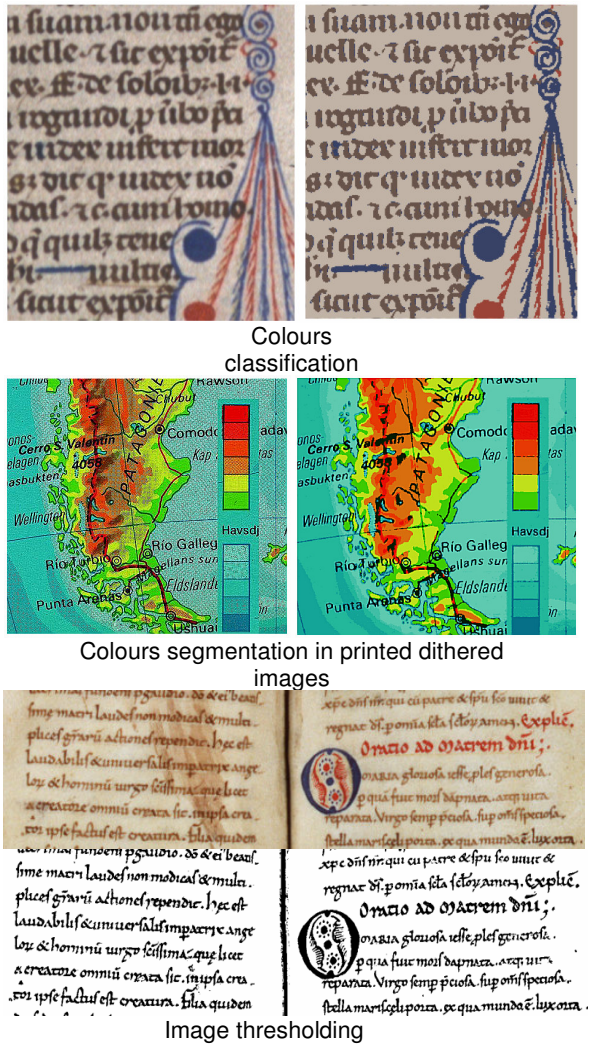


Figure 13. Different applications of the serialized k-means

3.3 Texture based approach for authors' writing classification

The texture based characterization has been developed for authors' manuscripts authentication and has been tested on the 18th century Montesquieu's corpus. Within this global approach we intent to show

that it is possible to analyze handwritten drawings without any a priori graphemes segmentation.

The particularity of the studied corpus lies on a great handwritings diversity which naturally brings up the following question for writers identification: ”From which hand has been produced this handwriting sample?”. Effort will not be apply on the transcription of texts which is made by literary experts, but we intend to enhance the value of those manuscripts by studying and exploiting handwriting patterns in their image representation.

In this part, we develop an original method for handwriting classification in visual separable families. Writer identification is the task of determining for a questioned document as to which individual, with known handwriting, it belongs to, [5], [6]. The first step of the process consists in defining a similarity measure to compare two handwritings. The second step consists in taking a decision: is their any stability intra-class and does it exist any visible differences between the tested handwriting and the known handwritings of the corpus. For that purpose, the similarity measures must be robust enough to minimize wrong acceptations and wrong rejections.

Generally, in most writer identification approaches, we try to produce a set of exhaustive graphical characteristics which can efficiently describe all handwritings. It is difficult to pretend to be exhaustive in such a description: that is the reason why we have chosen to use a single feature *the orientation* which expresses both global and local handwritings particularities in a context of limited corpus, [4], [7]. In this analysis, we are interested in evaluating the ratios which exist between typical handwriting local orientations, and which imply that two distinct handwritings, even identically skewed, will not be considered as similar because all other detected orientations will show significant differences. This method lies on the evaluation of a compact signature for each handwriting. The signature is obtained by the estimation of Gabor filters coefficients which reveal the presence of salient orientations. This process is applied on selected samples which must be the most homogeneous as possible. The selection of the sample has been automated for all digitized pages of the corpus with the computation of an entropy value which is directly correlated to the visual impression of “complexity” of the writing. In practice, an initial handwriting image must contain no less than 5 text lines to be interesting, with a bounded entropy. Then, Gabor analysis lies neither on homogeneous handwriting image samples and not on empty areas nor on noisy strokes lines regions.

The frequencies decomposition is based on the detection of most regular directions obtained by the application of the autocorrelation function on a sample of the entire initial image, see figure 14, [3].

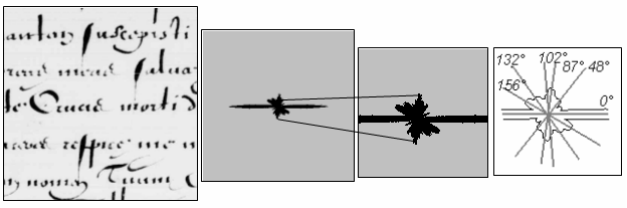


Figure 14. Directional rose and zooms in significant directions of the rose petals

For a given image, Gabor filters are computed in all significant directions of the handwriting. In this work, the scale of Gabor filters bank is constant in order to produce readable results (a good compromise between a too blurred response and a not significant filtering). The directional Gabor filters produce directional maps that reveal oriented patterns (graphemes). For each θ direction, these graphemes are then quantified by a density measure that reveals the contribution of the θ direction in the handwriting. In the example of figure 16, the more represented directions are: 2°, 55°, 90° and 145°. The successive AND-logical operations that are realized between the four maps guaranty that the initial selection of significant directions is relevant, see figure 15. Each direction is then weighted and ordered in a list: the handwriting *signature*.

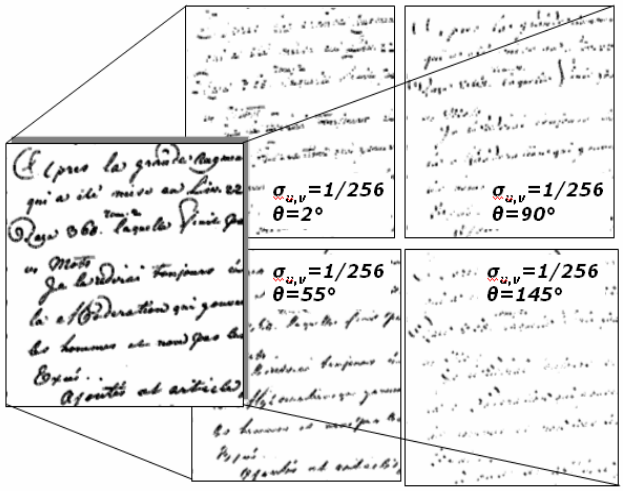


Figure 15. Handwriting decomposition for only four directional maps.

Figure 16 shows numerical signatures of a handwriting blocks that have been obtained by quantifying Gabor densities in all significant directions. In the x-axis we have the angular θ values and in the y-

axis the corresponding Gabor quantification. The local maxima of the curves for θ directions signify that the corresponding θ-oriented shapes of the handwriting samples are significantly represented. The local minima of the curves (the valleys) show that other directions are less important in the handwriting.

The more the curve is horizontal, the more the handwriting is curved, with a balanced distribution of angular direction in the handwriting. On the other hand, when angular Gabor densities are strongly contrasted with high maxima, the handwriting presents eye-catching shape properties with generally skewed and thin handwritten lines. The comparison between two signatures uses a warping function that allows possible fusion and fraction operations between two signatures. With this function, it is possible to compare two signatures that characterize two image samples having different sizes with, for example, big or small writings, see figure 16.

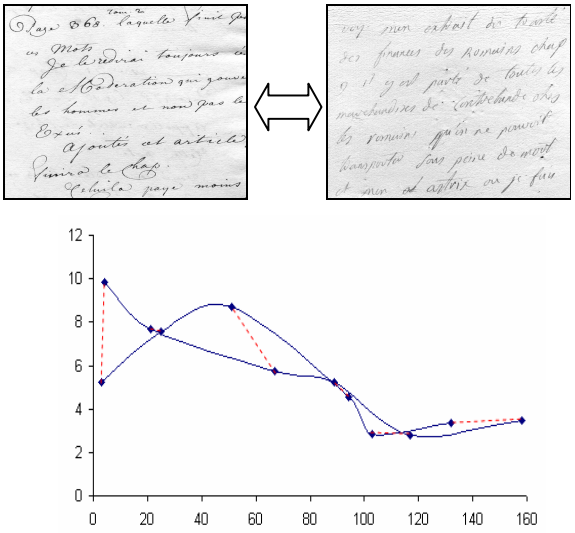


Figure 16. Matching curve between two signatures: application to the comparison of two signatures.

The warping distance is considered as the within-writer and the between-writer distance. The tolerance threshold to consider that two handwriting samples belongs to the same writer has been chosen for a maximal distance value and a maximal standard deviation. The distance D, the deviation σ and the Gabor quantification differences are all three necessary to express the resulting similarity between two handwritings. This approach is directly applied to writers' identification on an extended Montesquieu's corpus including Montesquieu's authentic drafts and anonymous 18th century authors' handwritings, i.e 500 handwriting samples of 48 writers. The warping distance is computed between two handwriting samples having similar entropy values. Figure 17 shows examples of within-writer stability evaluation.

In all those samples, we can notice that the signatures present similar tendencies with possible top or bottom curves translations which reveal handwriting thickness. The figure 18 shows a simplified set of relevant visual handwriting classes (or families) for a part of the test corpus. Each class is represented here by a single handwriting reference that is characterized by an individual typical signature.

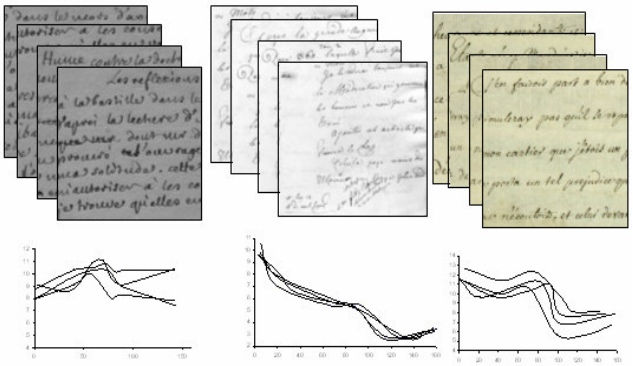


Figure 17. Within-writer variability. For an Economist of 1789 in the Bastille (Library of the municipality of Lyon, France); For a Montesquieu's copyist; For Montesquieu himself, *Histoire Véritable,* 1750.

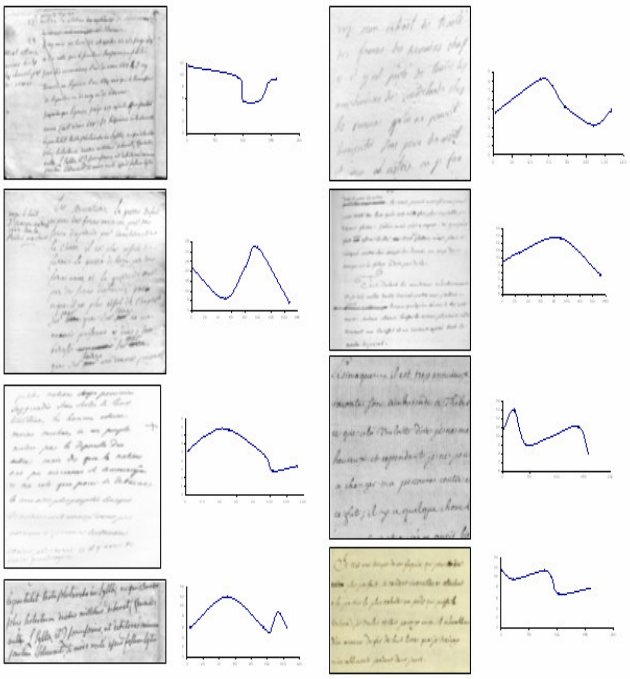


Figure 18. Individual handwriting signatures for 8 selected writers of the corpus.

The classification decision lies on a set of individual verifications between two samples: the query sample and each individual reference samples of the database. For each query handwriting page, we compute the signature on a reduced region that verifies the initial conditions of entropy and density. The relevance of the analysis is systematically evaluated by an intuitive visual *ground truth* and by a priori knowledge of the writers' styles that historian experts have taught us. The output of this process is a list of images that are ordered according to their similarity with the query sample. Within this methodology, we statistically obtain 91% of correct classification with the correct class as first response. The remaining 9% corresponds to query samples which are not homogeneous or which contain too many irregularities (essentially on draft pages). By enlarging the problem to a greater corpus, we can decide to generate automatically a new class when the warping distance between the tested handwriting and the reference models exceeds the predetermined threshold.

3.4 The recognition of medieval writing style for Palaeography

The paleography is discipline which collects handwritten texts corpus and knowledge accumulated on these documents. The goals of the palaeographic science are mainly the study of the correct decoding of the old writings and the study of the history of the transmission of the ancient texts. The most difficult problem to solve consist to study the writing style, independently from the author personal writing style, which can help to date and/or to transcribe ancient manuscripts. We began to work in collaboration with paleographers on a methodological and applicable contribution to the automatic analysis of writing styles of old manuscripts for the service of the research in history of texts. The usage of computer vision in this work is important for palaeographer because it may confirm or invalidate their work and bring more objectives conclusions. The recognition of the handwriting style which represents the historical period and/or the geographical localization of manuscripts ask different questions:

- How to define a reliable "*style similarity*" between complex writings?
- Which features to use in order to characterize a writing style and only the writing style independently from the writer, the text content and the image quality?

We are interested more in ancient Latin and Arabic manuscripts of the Middle-Ages which precedes the Renaissance period before the emerging of the printing. For palaeographers, the change from a writing to an other was not made in a radical way but by a slow and progressive evolution, which explains that it is difficult to identify categorically a given writing. For example we observe texts written in Caroline style which contain elements of the Gothic writing. The class of *Protogothic* writing is an intermediate writing style between the Caroline writing and the Gothic writing. Thus, the palaeographer should be able to quantify exactly the part of mixture of the writings families, see figure 19.

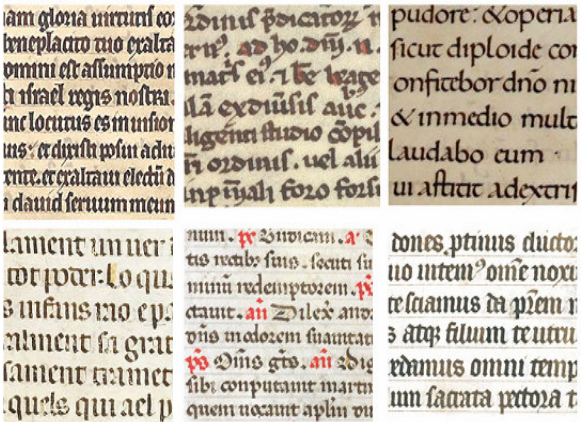


Figure 19. Different Latin Writing styles in Middle-ages manuscripts

We do not try to replicate the work of paleographers, because they generally use very particular features on specific letters ('r', 's', 'e', 'a') which must be taken inside words because their graphics change according to the writer when they are situated at the beginning or at the end of words. The reproduction of paleographer work means that we have to segment the document layout and especially all characters inside words which is a very difficult task for medieval manuscripts. We prefer to analyze statistically the whole image of a manuscript and measure globally all patterns. This approach should guarantee the independence from the text content, the writer's personal style, the used language and letters frequencies. If the image sample size is meaningful, all the letters are represented and in particular the characteristic letters used by palaeographers. Moreover, a global analysis allows the inclusion of some ornaments without affecting a statistical analysis because the text occupies a sufficient area.

The global approach advantages are very precious for the analysis of a great variety of documents having different qualities and origins. We develop a computer vision system for paleography studies which does not try to recognize the writing style but just returns sample

images having the most similar writing styles. In a first step, we want to develop a Content Based Image Retrieval system (CBIR), which uses a "Style Similarity" for image comparison and ranking. The user provides an image sample of the manuscript to the computer vision system which returns the more similar manuscripts images in terms of writing style and their ranking see figure 20. We have determined several criteria that image features must verify to define a "style similarity":

- **Robustness**: The image features must be calculated without image segmentation and must be robust to images noises and documents degradation.
- **Writer invariance**: the measures should be independent of the writer personal style2
- **Statistically robust**: image features must be invariant to the size of the text samples.
- **Independence:** Style must be independent from the language used, text contents and frequency of letters.
- **Document layout free**: Style must be independent from the document layout (text spacing, strokes densities...). Document layout can be used by paleographer to identify the writing style. For example the Gothic Textualis make regular text lines and almost use fix spacing between text lines and words.
- **The change to geometric transform**: Writing must be invariant to classical geometric transform like the modification of the scale factor, image ratio and image skew. But in practice, to process images from different origin, we need to normalize the images to make them comparable.

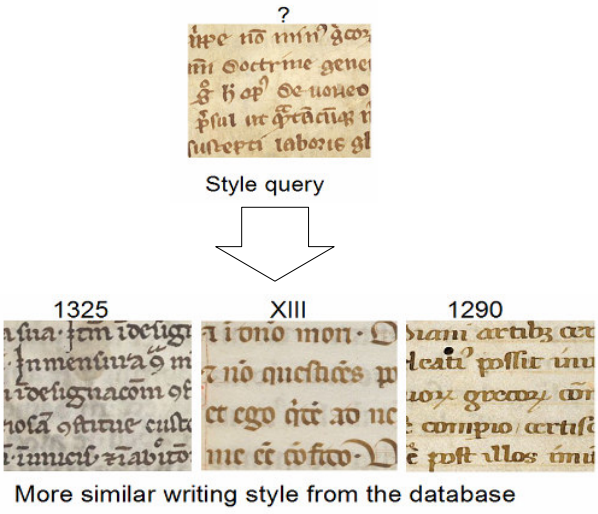


Figure 20. Examples of style query and most similar responses from the database.

Several features can be used to characterize the writing style like Fractal analysis, auto-similarity measure, and texture characterization by Gabor, Wavelets. We have concentrated our effort on cooccurrence approach which measures the relationship between intensity values. The cooccurrence can be evaluated from the SGLD (*Spatial Gray-Level Dependence*) which is a join probability to observe the same intensity value between two different pixels according to their spatial relation.

$$SGLD(\rho,\theta,i,j)=Mes[I(x,y)=i,I(x+\rho\times\cos\theta,y+\rho\times\cos\theta)=j]$$

In order to measure only the writing style, the SGLD is computed locally with very weak displacement to guarantee that we do not take into account the repetition of characters and text lines. We have chosen to study the Spatial Gray Level Dependency locally with a small radius corresponding to the half of the tex height. The SGLD bring a large amount of information, which characterize many writing styles, see figure 21.

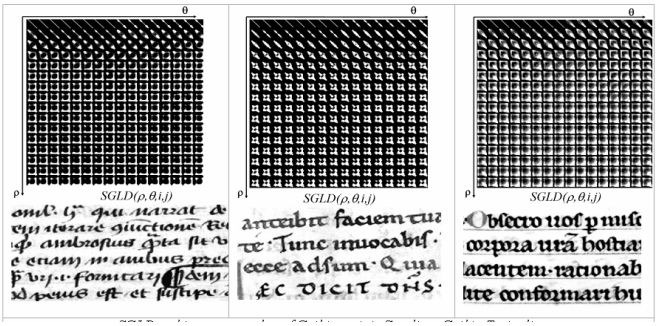


Figure 21. SGLD and images samples of Gothic script, Caroline, Gothic Textualis

The SGLD is identical on different text areas of the same document and is robust to noise and does not require any image segmentation nor layout analysis. The cooccurrence preserves the same information about shapes after the main geometric transformations. But this information is not preserved anymore by the same matrices following ρ and θ. To guarantee that we compare the same information, it is necessary that the images have the same orientation, scale and ratio. We reduce the features size by using Haralick descriptors. Then we study a feature selection and combination by applying a Linear Discriminant Analysis (LDA). The LDA give several possible combinations of features that discriminate correctly the most important writing style. From these combinations of features, we are going to define a "style similarity" measure and form a large database of images samples of writings with a

paleographic description to develop a reliable image retrieval system for medieval writings styles.

3.5 Words spotting by word decomposition and shape description

It is important to notice that recognition systems for unconstrained handwritten manuscripts are cannot be used for the automatic reading of old manuscripts. But we can use computer based technology to develop tools to assist the operator and reduce expensive manual description by using *Words spotting* in old scripts for text indexing. In the past years many levels of indexation have been developed to allow a fast retrieval of digitized documents. Among all the ways of indexing a document, textual indexation allows the finest queries on the documents' content. Usually, the plain text transcription of a digitized document is obtained by applying OCR (Optical Character Recognition) software on it. What if the OCR fails? Especially on handwriting ancient documents, it is not possible to access to the text content without human expert transcription. In this part, we introduce an alternative method to access to handwriting text trough the image: *the Word Spotting.*

The word spotting is a technique that localizes users' selected words in a text written or spoken in a natural language without any syntactic constraint. It is a generic approach that can be applied to any document written in any language using any alphabet, pictograph or ideogram. As a result, the system proposes a sorted list of hits that the user can prune manually. The word spotting is based on a similarity or a distance between two images, the reference image defined by the user and the target images representing the rest of the page or all the pages from a multi-page document.

Correlation methods applied directly on the grey levels can be used for image similarity comparison but this classical approach is very sensitive to geometrical transformations and is time consuming. Moreover, correlation cannot be easily adapted to the spatial variations of the handwriting. The solution consists in representing the informative parts of the images with feature vectors that can be compared with other feature vectors from another image. Therefore the problem consists in defining the *Zones of Interest* (ZOI) centred on the informative parts of the image and the most *discriminative* features.

The choice of the features is critical for the overall performance of the system. The grey levels (i.e. the luminance of each pixel) are not representative of the local image structure and cannot be used as a feature alone. The results of the experimentations using only grey levels confirm this assertion. On the other hand, differential features are more suited for the nature of images of documents because it is a good representation of the local image structure and the complex shapes of the handwritings. Moreover, some differential features are invariant to classical geometric transformation, and robust to scale modification, lighting variation and artificial or natural noises. The robustness of the differential features against local noises and JPEG artefacts is provided by a preliminary Gaussian filtering. We tested a large number of differential features (such as Gaussian curvature, hessian eigenvalues, flowline curvature…) on our validation database and selected the feature with the best P–R curve, using a naive matching algorithm. The feature that outperforms the others is the gradient angle. For natural images, the gradient magnitude has significant values almost everywhere, but for document images, the gradient magnitude has non null values only around strokes. The gradient vanishes in the background where colours are generally bright, or at the center of large black lines from the writing, see figure 22.

Therefore we put a threshold on the gradient magnitude before computing the angular distance, see figure 23. In order to make a matching algorithm robust to spatial variations, each pixel of the template has to be compared with all the pixels of a neighbourhood in the image. By selecting relevant Zones of Interest, we can reduce drastically the running time. As only the informative zones are matched, it can also improve the accuracy [14].

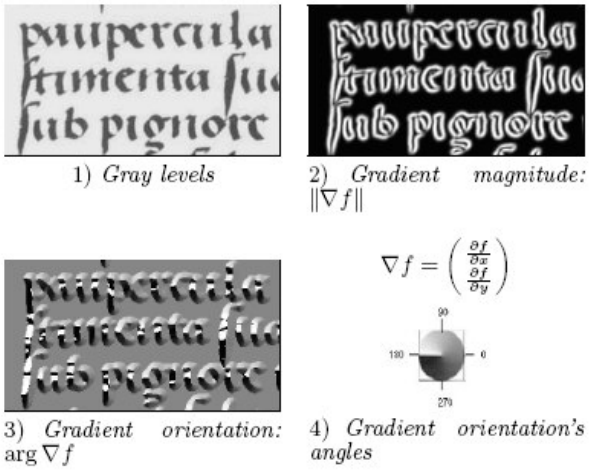


Figure 22. The gradient features.

$$d_\varepsilon(a,b) = \begin{cases} \min\left(|\arg(a) - \arg(b)|, 128 - |\arg(a) - \arg(b)|\right), \\ \quad \text{if } \|a\| > \varepsilon \text{ and } \|b\| > \varepsilon \\ 255, \text{else} \end{cases}$$

In middle-ages Latin manuscripts, the letters are mainly composed of large vertical strokes linked by thin horizontal strokes. With the passing time, those thin strokes tend to dim and the digitization process paired with a JPEG compression make a lot of thin strokes disappear. In terms of information theory, most of the information is situated on these vertical strokes that we extract. The graphemes we obtain are then used as guides to the matching. Let us illustrate this matter with the example of the letter "u". The distance between the vertical parts of two occurrences of "u" is slightly different, making naive matching algorithms fail. Therefore it is more relevant to compare the pixels around the vertical parts than to compare the whole shapes, see figure 23.

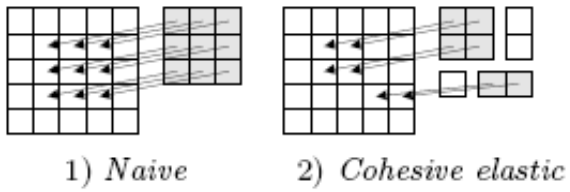


Figure 23. Naïve versus cohesive matching. In white, the image and in grey the template.

We use mathematical morphology on the gray levels of the images to compute guides. We perform an opening with a vertical structuring element. The bounding boxes of the guides are enlarged to obtain the Zones of Interest. The template is split into pieces through the ZOIs and the distance and orientation between the centers of those zones are stored. The ZOIs are sorted from left to right. The links between them are loose so that the template can be deformed to match better to the most deviant occurrences. The guides of the image are marks used to begin the search of a new occurrence. The first zone of interest of the template is matched sequentially with each ZOI of the image.

For the experiments, we selected 24 two-columned pages (approximately 2000 words) from the Latin manuscript Escalopier 22 (see Figure 24) from the Amiens library. We searched for words that we chose based on two criteria: semantic relevance for a realistic sampling of what an end-user would search, and frequency for statistics. As the text is in Latin, we cut some declensions and use roots as keywords. The keywords chosen were "benedic-", "deus", "domin-", "fili", "gloria", "Maria", "patri", "terra" and "virg-". Out of 195 occurrences of the nine keywords, we expected only 153 to be retrieved (78.5%). Finally, the algorithm retrieved 171 good word occurrences (87.7%) i.e. 9.2 points more than expected. The Precision-Recall (P–R) curve is given in Figure 24. All the keywords were retrieved at almost equivalent recall and precision.

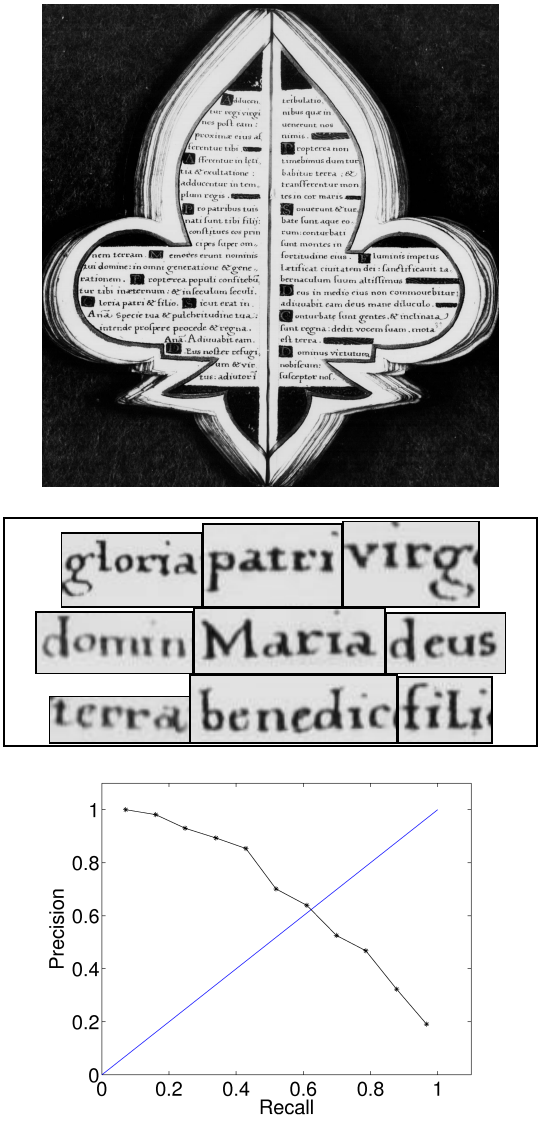


Figure 24. Latin manuscript Escalopier 22 from the Amiens library. Selection of retrieved image words and the corresponding P-R curve. [14]

We also tested our method on Rath's database GW20, which contains 20 pages of George Washington's manuscripts, [18] (see Figure 25). These are cursive modern manuscripts with approximatively 20 lines per page and about 4800 words. The manuscript was labelled and we sorted the words from the most redundant to the lowest number of occurrences.

We then selected the 15 most significant words in terms of semantics ("1755", "orders", "captain",

"December", "letters", "fort", "instructions", "company", "October", "regiment", "Virginia", "Winchester", "Cumberland", "recruits" and "sergeant"). Out of 298 occurences, we found 221 good hits, i.e. a 74.16% recall. The P–R curve is given in Figure 25. Such a curve was indeed expected.

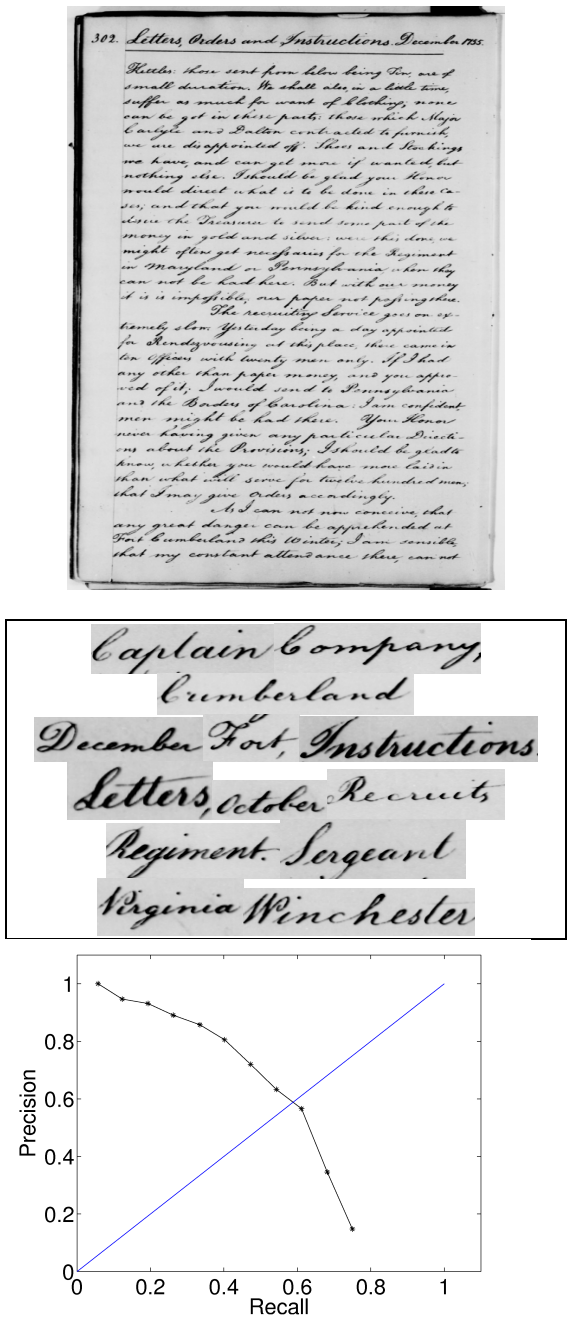


Figure 25. A sample of the George Washington manuscript and the words used for queries and the corresponding P-R curve.

The GW20 images contain text of different heights and the thickness of the line varies greatly. Most of the words can be clustered into two classes: a big and bold words class and a thin words class. If we select a template in one class, the occurences of this word belonging to the other class are not likely to be retrieved. In this test, our algorithm did what it was meant to do: it retrieved words written in the same typographical style with high geometrical variations. As the documents we had to process were not segmentable, we did not include this type of processing in our method and were not able to include a normalisation step as Rath et al. did. However, our results are fair considering that the GW20 is very different from the images our method was meant to deal with.

3. Conclusion

This paper is a response to the great challenge to access ancient enriched digitized manuscripts, to exploit them and to retrieve information tanks to computed meta-data. The main hypothesis which has been used here is the consideration of the *graphical* modality of digitized handwriting documents that must be opposed to textual ASCII formats. In this project, we do not need to transcript the texts nor to recognize words contents, but we propose image processing tools to resolve the problems of image enhancement, denoising & restoration, the medieval writers' classification & authors hands categorization, and finally the word image retrieval. In this research field, our research staffs already have a validated expertise of handwriting shapes characterization for the writers' categorization and word spotting. We aimed here at presenting different complementary software works which have been studied in our laboratory during the past four years on diversified medieval and authors' manuscripts collections.

In that context, this work is a response to scientific problems of historical handwritten corpus digitization. This paper is the first feasibility study which is necessary to build a complete indexation and classification system for degraded patrimonial handwriting documents. In all cases, we have privileged segmentation free approaches to avoid image segmentation. The results of handwriting page denoising and restoration, of handwriting classification and word retrieval are very promising. We are currently working on the integration of all those different contributions in a complete plateform so as to provide a enriched access to manuscripts collections.

4. Reference

[1] ACI MADONNE Project: http://l3iexp.univ-lr.fr/madonne/

[2] H. S. BAIRD, State of the Art of Document Image Degradation Modelling, *IAPR 2000 Workshop on Document Analysis Systems, Brazil*, December 2000.

[3] BRES, S., *Contributions à la quantification des critères de transparence et d'anisotropie par une approche globale.* PhD Thesis, 1994.

[4] BULACU, M., SCHOMAKER, L., Writer style from oriented edge fragments, in CAIP Computer Analysis of Images and Patterns, pp. 460-469, 2003.

[5] CATALIN, I.T., ZHANG, B., SRIHARI, S.N., Discriminatory power of Handwritten words for writer recognition, in International Conference on pattern Recognition, pp.638-643, 2004.

[6] CHA, S.H., SRIHARI, S. Multiple Feature Integration for Writer Verification. Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition, IWFHR VII, p 333-342, 2000.

[7] EGLIN, V., VOLPILHAC-AUGER, C., Caractérisation multi-échelle des tracés manuscrits en vue de la catégorisation de scripteurs, in Proc. Of *CIFED*, 2004, pp. 106-114.

[8] FRANKE, K., KOPPEN, M., *A computer-based system to support forensic studies on handwritten documents,* Int. Jour. Doc. Anal. Reco., 3:218-231, 2001.

[9] KUCKUCK, W., *Writer recognition by spectra analysis*, in Proc. Int. Conf. In Security through Science Engineering, 1980, pp.1-3.

[10] LEBOURGEOIS, F., TRINH, E., EMPTOZ, H., *Compression and accessibility with the images of digitized documents – Application to the Debora project, Numerical Document*, Flight 7n°3-4, 2003 p103-127.

[11] LEBOURGEOIS, F., TRINH, E., ALLIER, B., EGLIN, V., EMPTOZ, H., *Document Image Analysis solutions for Digital libraries*, In Proc. Of DIAL, pp. 20-32, 2004.

[12] LEEDHAM G., VARMA,S., PATNKAR, A., GOVINDARAJU,V., *Separating text and background in degraded document images- a comparison of global thresholding techniques for multi-stage thresholding.* In Proceedings of the 8th inter. Workshop on frontiers in handwriting recognition, 2002, pp. 244-249.

[13] Y. LEYDIER, F. LEBOURGEOIS, H. EMPTOZ, Serialized k-means for adaptative color image segmentation – application to document images and others, *DAS 2004*, LNCS 3163, Italy, September 2004, 252-263.

[14] Y. LEYDIER, F. LEBOURGEOIS, H. EMPTOZ, Omnilingual Segmentation-free Word Spotting for Ancient Manuscripts Indexation. In Proc. Of ICDAR'05, Séoul, 2005.

[15] T. LINDEBERG, Scale Space Theory in Computer Vision. Kluwer, Boston, 1994.

[16] MARTENS J.-B., The Hermite transform – Theory, IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, no. 9, pp. 1595-1606, 1990.

[17] NOSARY, A., PAQUET, T., HEUTTE, L., *Reconnaissance de textes manuscrits par adaptation au scripteur*, CIFED'2002, pp.365-374.

[18] T. RATH, S. KANE, A. LEHMAN, E. PARTRIDGE, AND R. MANMATHA. Indexing for a digital library of George Washington's manuscripts: A study of word matching techniques. Technical report mm–36, Center for Intelligent Information Retrieval, University of Massachusetts, 2002.

[19] RIVERO-MORENO C.J., BRES S., *Conditions of similarity between Hermite and Gabor filters as models of the human visual system,* In Petkov, N.& Westenberg, M.A. (eds.): Computer Anal. of Images & Patterns, Lectures Notes Computer Science, vol. 2756, pp.762-769, 2003.

[20] SAID, H.E.S, PEAKE, G.S., TAN,T.N., BAKER,K.D. Writer identification from non-uniformly skewed handwriting Images, British Machine Vision Conference, pp. 478-489, 1998.

[21] TAN, C.L., CAO, R., SHEN,P., Restoration of archival documents using a wavelet technique, In Pattern Analysis and Machine Intelligence, 4(10), pp. 1399, 1404, 2002.

[22] TONAZZINI A., BEDINI L., SALERNO E., Independant component analysis for document restoration, Inter. Jour. on Doc. Analysis and Recognition, 7: 17-21, 2004.

[23] VOLPILHAC-AUGER, C., EGLIN, V., La problématique des ouvrage manuscrit ancien : vers une authentification des écritures des secrétaires de Montesquieu, *Journée sur la valorisation des documents et numérisation des collections,* Ecole Normale Supérieure de Lyon, le 7 mars 2002.