

Towards historical document indexing: extraction of drop cap letters

Mickael Coustaty · Rudolf Pareti · Nicole Vincent · Jean-Marc Ogier

Received: 20 February 2010 / Revised: 22 October 2010 / Accepted: 9 February 2011 / Published online: 25 February 2011
© Springer-Verlag 2011

Abstract This paper deals with the difficult problem of indexing ancient graphic images. It tackles the particular case of indexing drop caps (also called Lettrines) and specifically, considers the problem of letter extraction from this complex graphic images. Based on an analysis of the features of the images to be indexed, an original strategy is proposed. This approach relies on filtering the relevant information, on the basis of Meyer decomposition. Then, in order to accommodate the variability of representation of the information, a Zipf's law modeling enables detection of the regions belonging to the letter, what allows it to be segmented. The overall process is evaluated using a relevant set of images, which shows the relevance of the approach.

Keywords Letter extraction · Historical documents

1 Introduction

There is a growing interest in digitally preserving and providing access to the historical document collections residing in libraries, museums, and archives. Such archives of old documents are a unique public asset, forming the collective and evolving memories of our societies. Indeed, ancient

documents have a historical value not only for their physical appearance but also for their contents. Examples include unique manuscripts written by well-known scientists, artists or writers; letters, trade forms, or official documents that help to reconstruct historical sequences for a given place or time; and artistic items such as stamps, illustrations, covers.

The challenge that is currently being addressed throughout Europe is the conversion of such heritages into digital libraries that enable them to be preserved, but also to make them available worldwide using web-based portals (like the impact project¹ for instance). Through the medium of better-designed digital libraries, citizens of the future should be able to gain access to a myriad forms of knowledge from anywhere, at any time, and in an efficient and user-friendly fashion. A number of initiatives exists focusing on the creation of large digital libraries that are globally accessible. Google is now engaged in a project to create a global virtual library. A number of European libraries has started a similar joint project [26]. DELOS is the European Network of Excellence for digital libraries [25]. The construction of such libraries has an additional important challenge: the analysis of documents and the extraction of knowledge. This goal involves projects to design and develop semantics-based systems to acquire, organize, share, and use the knowledge embedded in the documents. The field of data mining, combined with document analysis, offers a robust methodological basis for performing tasks such as descriptive modeling (clustering and segmentation), classification, discovering patterns and rules, and retrieval by content applied to document sources and databases. Old documents may be originals (paper, parchment etc.) or in image form (already scanned, possibly using now-outdated technologies). The key requirement is to be able to process these unique manuscripts,

M. Coustaty (✉) · J.-M. Ogier
Imedoc Team - L3i Laboratory, Avenue Michel Crepeau,
17042, La Rochelle, France
e-mail: mcoustat@univ-lr.fr

J.-M. Ogier
e-mail: jmogier@univ-lr.fr

R. Pareti · N. Vincent
SIP Team, LIPADE Laboratory,
45, rue des Saints-Peres, 75270 Paris Cedex 06, France
e-mail: nicole.vincent@mi.parisdescartes.fr

¹ <http://www.impact-project.eu/>.

whether they are presented as free-flowing text (treatises, novels, ...) or structured at various levels of physico-logical structure correspondence (letters, census lists, trade forms, ...). Degradation may be caused by a lifetime of use, and access must also be provided to user annotations and corrections, stamps and unique artwork. Each class of document requires a different approach throughout the conversion process and lends itself to different levels of information extraction and description. In summary, the work comprises the analysis of knowledge in historical documents to compile metadata that are then used to access digital libraries.

In the knowledge society, the objective is not simply to digitize documents but to create semantically enriched digital libraries of such digitized documents. Enrichment of a document means the addition of semantic annotations to digital images of the scanned documents. Such metadata are intended to describe, classify, and index documents by their content. It would then enable easy access to this cultural and scientific heritage from any place and at any time.

The main research goal of the French project "NaviDo-Mass" (NAVIGATION into DOcument Masses)² is thus to work in a collaborative framework on the Analysis of Old Documents. This goal consists of developing pattern recognition and image analysis techniques that allow the extraction of knowledge from documents and its conversion into digital libraries containing the scanned pages enriched with semantic information.

This paper deals with one specific point of the NaviDo-Mass project, concerning the indexing of the graphic portions of historical documents. Even if we present our method on a specific corpus, one part of the strategy remains generic and could be applied to any other graphic images in an indexing scheme of historical documents. In this paper, we will use three terms to define the graphic objects of our corpus: *lettrine*, *drop cap*, and *ornamental letter*. The literature contains examples of all three names being used for the same objects. The proposed approach is applied to lettrines, which are present in all books of the fifteenth and sixteenth century.³ Analyzing these images is thus significant for indexing books of the beginning of the printing period. The lettrines, which were originally used in a decorative goal, are actually used, on one hand, to distinguish the works, the printers and to detect the beginning of a chapter or a paragraph. On the other hand, historians aimed at using lettrines to navigate into database from a lettrine to another, to make query by example or to analyze their specifications. These research



Fig. 1 Various lettrines illustrating the great variability of representation of the information

activities are made in order to analyze the font, the color, or the alphabet that characterize the printer.

Images of lettrines are made up of three principal elements: the letter, the pattern, and the background (see Fig. 1). An important step in the indexing of lettrines consists of segmenting the letter and the elements from its background, in order to characterize them by using a relevant signature.

One approach may be to define different styles among the lettrines according to the background appearance. In fact, what would be a good criterion to achieve this clustering? Historians criteria are not those that ease the computations. Color is not significant as, according to the lettrine, there can be no leading colors. It is possible to make clusters as done in some studies [7] but what is the best number of clusters?

This approach would induce ad hoc method for each lettrine family, and on the whole, the recognition rate of letter is not improved. Indeed, the greatest difficulty of our work is the degradation of the images and ideal patterns.

The focus of this paper is a consideration of the problem of the extraction of the letter from a lettrine (dropcap), which appears to be quite a simple problem but is in fact quite complicated, owing to the great variability of the representation of the information: colors, shapes, connectivity to various illustrations etc. (see Fig. 1 for an illustration of various lettrines).

This paper presents in detail the several stages of our method:

1. Simplification of the images using layer decomposition techniques
2. Extraction of shapes from one of these layers
3. Selection of the shape that corresponds to the character.

This work is inspired by [8, 18, 19] who used a wold decomposition on the one hand, and a Zipf's law on the other hand, to extract the elements of drop caps. Our paper is organized as follows: first of all, we illustrate features of lettrines, in order to highlight the difficulty in extracting a letter. Based on these features, we present our strategy for the extraction of the letter, based on Meyer decomposition (layer segmentation), Zipf modeling, and letter extraction.

² This work is supported by the French National Research Agency project NAVIDOMASS referenced under ANR-06-MCDA-012. For more details and resources see <http://navidomass.univ-lr.fr/>.

³ Historians of our projects plan to digitize 15000 books from this period. Actually, 479 books are available online at <http://www.bvh.univ-tours.fr/>. These books represent more than 168000 pages digitized.

2 Image description and features

Before considering our strategy, it is important to describe the images to be processed and to understand their principal shared features, so that we may apply a generic method.

2.1 The support

The majority of our images were scanned at 300 dpi. The support for scanning is the original paper of the book, which may have become degraded over time. The paper is composed of three elements: the fiber, the filler, and the gluing (the filler and the gluing are added to give better properties to paper [20]). Fibers, mainly composed of organic materials such as linen, cotton, or wood, are a cause of deterioration of the paper in ancient documents. Among the problems due to aging of these fibers, we may cite yellowing or thinning of the pages (which may cause a transparency of the paper and problems of over-print scanning) and the weakening of the pages. This weakening is often accompanied by mechanical effects such as tearing or creasing of the pages. The main consequences for images resulting from these degradations of the support are (1) a very high level of noise, (2) a possible over-printing of the verso page, and (3) some possible soiled areas.

2.2 Printing technique

The second major feature of the documents that we are processing arises from the techniques used in early printing. At that time, stamps were hand-carved from wood to be inked and pressed on paper. These buffers, specific to each printer, are used today to authenticate documents and to characterize them. The buffers allowed documents to be created in black and white but did not allow levels of gray. In order to create shades and shadows on documents, and to suggest grays, printers used artistic strategies, replacing the shades of gray by parallel lines. For example, in the images in Fig. 1, the shadows of the arms and legs of various characters are represented by parallel lines. This technique is used for different decoration elements present in the book, for example at the end of chapters or in the beginning of paragraphs.

2.3 Description of the lettrines

Among various types of graphic images of old documents contained in [9], we are particularly interested in *Lettrines*. They correspond to images widely used in books over time. A lettrine is an ornamental letter that begins a chapter or a paragraph (see Fig. 2) and can be viewed as a binary image composed of strokes. Some studies [4, 12, 18, 23] have attempted to characterize this kind of graphic image. Discussions with our historian partners in the NaviDoMass project enable us

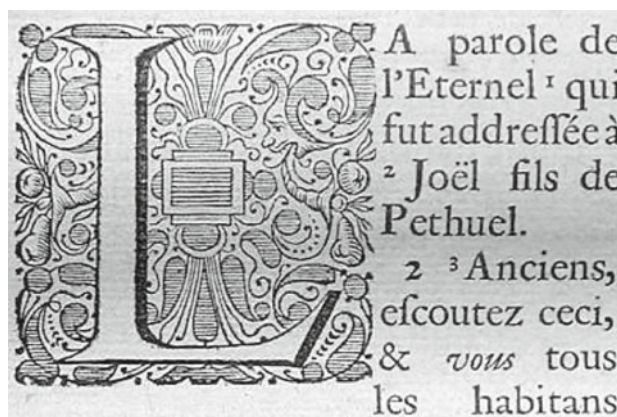


Fig. 2 Example of lettrines in context

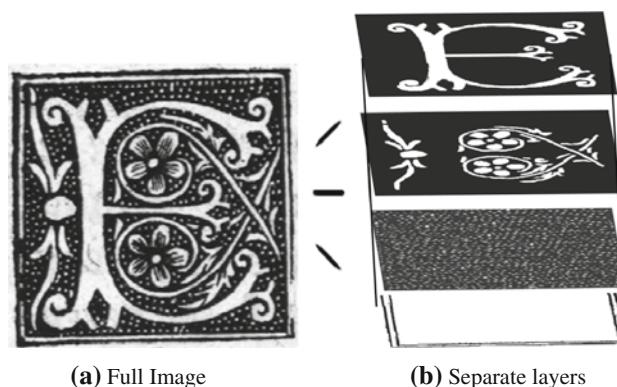


Fig. 3 Layers in a lettrine. From bottom to top, we can observe the frame layer, the background layer, the pattern layer and the letter

to understand how historians decompose lettrines. According to these specialists, lettrines can be viewed as an overlaying of two layers: the background and the letter.

- The background may be black, white, dashed, or streaked, and is entirely composed of ornamental objects. It may be decorative or figurative. The ornamental objects may be considered themselves as set of layers ; at last, a global frame can be present around the image taking different styles. These different elements can be seen in Fig. 3 ;
- The letter is one of the most important pieces of a lettrine. It is homogeneous in color (often black or white) and from a specific font. Most of the time, a simple extraction of connected component is not enough because the images are noisy, their boundaries are not clean, and they are often split into many pieces.

Finally, these layers may be contained within a frame (see Fig. 3) that defines the boundaries of the typographical block. It can be composed of zero, one, or two strokes.

From frequencial and spatial points of view, these layers can be separated. Background is composed of a mix of uniform and textured areas, while the letter corresponds to a uniform white or black area in the middle of the lettrine.

2.4 Important features for the definition of a relevant strategy

Considering features described above, one can establish the basis of a relevant strategy. First of all, considering the high probability of degradation of the useful information, a pre-processing technique should be applied to filter the noise out of images. Pre-processing techniques could also be used for segmentation purposes, in order to separate the layers identified by the experts in this domain, i.e. historians. In this regard, examination of a sizeable set of lettrine images highlights the fact that the ornamental background part is characterized by quite regular frequencies (texture analysis), while the letter and all the other information correspond to very low-level spatial frequencies and are characterized by quite large homogeneous areas.

Another important feature that must be considered in letter extraction is related to the fact that the letter can be of different colors : black or white. Consequently, the method to be developed must adapt itself automatically to this constraint.

A first method for extracting letters from drop caps was described in [21]. This method is based on the study of shape descriptors. Instead of trying to find the best discriminating descriptor, the authors suggest to use a combination of several descriptors depending on the type of symbol to be extracted. They automatically assign to each descriptor a recognition map, based not only on the recognition rate but also on the distributions of the errors, to define a descriptor measure for each cluster of symbols. They applied the method to drop caps and to letter extraction, but no quantitative evaluation was provided to quantify the quality of their extraction.

As presented in the previous section, lettrines are composed of different layers of information. Each layer can be defined by its specific spatio-frequencial features. It will be relevant to use a method that takes into account the decomposition in layers.

Wavelets decomposition is a one-step process that can be used for splitting low frequencies from high frequencies of image, at different scales [14]. In this context, between each scale, a smoothing function is applied, while a low-pass filter and a high-pass filter extract different frequencies. Since the letter corresponds to large homogeneous area with low gray level variation, we can recover it from low-frequency layer. However, this method could not be used to extract letter. Indeed, boundaries of letters do not appear in

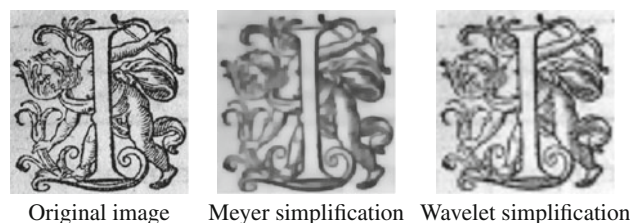


Fig. 4 Illustration of two kinds of image simplification effects on boundaries. From *left to right*: original image ; image obtained using Meyer decomposition; image obtained using Haar wavelet

low frequencies, and the smoothing function erases thinner boundaries (see fig. 4 to see an example of degraded image).

This discussion leads us to introduce our own strategy, based on Meyer decomposition, which enables to filter out the noise and to use spatial frequencies of lettrine images to segment them into separate layers: a shape layer, a texture layer, and a noise layer. Then, based on the resulting layer decomposition, the use of a Zipf's law on the gray levels of the shape layer will allow us to detect large homogeneous areas, which correspond to the letter. Extraction of the letter then becomes feasible.

3 The gray scale Meyer decomposition

Decomposing an image into meaningful components appears to be one of the major goals of recent developments in image processing. This kind of model arises from a well known but ill-posed problem in image processing: The aim is to recover an ideal image u , from a degraded observation f , as follows: $f = Au + v$ (where A is a linear operator representing blur and v is the noise, often additive). A classical approach consists of introducing a regularization term leading the system to admit a unique solution. This minimization problem leads to a number of applications already successfully implemented, such as image restoration, deconvolution, deblurring, zooming, inpainting, classification, colorization, segmentation, and optical flow regularization.

The first goal was image restoration and denoising, but following the ideas of Yves Meyer [16], in the total variation minimization framework of L. Rudin, S. Osher, and E. Fatemi [13], the decomposition of images into geometrical and oscillatory (i.e texture) components appears to be a useful and very interesting approach for our image analysis problem. We want to obtain the primary structure of images in order to properly segment the drop cap, independently of its textured portions, and to avoid acquisition problems such as noise.

The images of drop caps are very complex, very rich images in terms of information, and need to be simplified. These

images are mainly made up of lines, which are unsuitable for the usual texture methods. We, therefore, use an approach developed by Dubois and Lugiez [10, 11], which relies on a series of projection to separate the original image into several layers of information.

The first one, $u \in BV$,⁴ containing the structure of the image, a second one, $v \in G$,⁵ the texture, and the third one, $w \in E$,⁶ the noise. Meyer decomposition is an iterative process, which extracts details of image. Signal of image is projected in particular spaces (J, J^* and B^*) to only keep interesting parts. A residual part $\left(\frac{1}{2\alpha} \|f - u - v - w\|_{L^2}\right)$ is present in the minimization of the functional to get all the elements that are not caught by the other projections. At the opposite of wavelet transformation approaches, boundaries are kept into geometric part and letter can then be better extracted by using a connected component approach. For better comprehension of different spaces, see [1–3].

This decomposition model is based on a minimization of this discretized functional F :

$$\inf_{(u,v,w) \in X^3} F(u, v, w) \tag{1}$$

with

$$F(u, v, w) = J(u) + J^*\left(\frac{v}{\mu}\right) + B^*\left(\frac{w}{\lambda}\right) + \frac{1}{2\alpha} \|f - u - v - w\|_{L^2}$$

where J is the total variation related to the extraction of the geometrical component, B is a norm defined on the Besov space, $J^*\left(\frac{v}{\mu}\right)$, $B^*\left(\frac{w}{\lambda}\right)$ are the respective Legendre–Fenchel

⁴ $BV(\Omega)$ is the subspace of functions $u \in L^1(\Omega)$ such that the following quantity, called the total variation of u , is finite:

$$J(u) = \sup \left\{ \int_{\Omega} u(x) \div (\xi(x)) dx \right\}$$

for any $\xi \in C_c^1(\Omega, \mathbb{R}^2)$, $\|\xi\|_{L^\infty(\Omega)} \leq 1$

⁵ G is the subspace introduced by Meyer for oscillating patterns. G denotes the Banach space composed of the distributions f which can be written $f = \partial_1 g_1 + \partial_2 g_2 = \text{div}(g)$ with g_1 and g_2 in $L^\infty(\Omega)$. With $\|f\|_{L^\infty} = \sup_{t \in [a,b]} |f(t)|$, on G , the following norm is considered:

$$\|v\|_G = \inf \left\{ \|g\|_{L^\infty(\Omega, \mathbb{R}^2)} / v = \text{div}(g), \right. \\ \left. g = (g_1, g_2), |g(x)| = \sqrt{|g_1|^2 + |g_2|^2}(x) \right\}$$

⁶ It enables to model oscillating patterns. Let $\dot{B}_{1,1}^1$ be the usual homogeneous Besov space then the dual space of $\dot{B}_{1,1}^1$ is the Banach space $E = \dot{B}_{-1,\infty}^\infty$

Algorithm :

At step n :

1. u and v have been previously computed, we estimate:

$$\tilde{w} = P_{\delta B_E}(f - u - v)$$

2. then we compute:

$$\tilde{v} = P_{\mu B_G}(f - u - \tilde{w})$$

3. and we finally obtain:

$$\tilde{u} = f - u - \tilde{v} - \tilde{w} - P_{\lambda B_G}(f - \tilde{v} - \tilde{w})$$

This operation is repeated until :

$$\max(|\tilde{u} - u|, |\tilde{v} - v|, |\tilde{w} - w|) \leq \varepsilon$$

Fig. 5 Grayscale image decomposition algorithm

transforms⁷ of J and B [3] for the extraction of texture and noise components. Parameter α controls the L^2 – norm of the residual $f - u - v - w$, and X is the discrete Euclidean space $\mathbb{R}^{N \times N}$ for images of size $N \times N$.

To minimize this functional, Chambolle’s projection algorithm is used [2]. The Chambolle’s projection P on space λB_G ⁸ of f is denoted $P_{\lambda B_G}(f)$ and is computed by an iterative algorithm. This algorithm starts with $P^0 = 0$ and for each pixel (i, j) and at each step $n + 1$ we have the following:

$$P_{i,j}^{n+1} = \frac{P_{i,j}^n + \tau \left(\Delta \text{div} (P^n) - \frac{f}{\lambda} \right)_{i,j}}{1 + \tau \left| \Delta \text{div} (P^n) - \frac{f}{\lambda} \right|_{i,j}} \tag{2}$$

In [6], a sufficient condition ensuring the convergence of this algorithm is given as follows: $\tau \leq \frac{1}{8}$. To solve (1), the authors propose the algorithm of Fig. 5. The residual part of the functional is thus included in the noise layer.

In [2], the authors replace $P_{\delta B_E}(f - u - v)$ by $f - u - v - W_{ST}(f - u - v, \delta)$ where $W_{ST}(f - u - v, \delta)$ stands for the wavelet soft-thresholding of $f - u - v$ with threshold δ defined by:

$$S_\delta(d_i^j) = \begin{cases} d_i^j - \delta \text{sign}(d_i^j) & \text{if } |d_i^j| > \delta \\ 0 & \text{if } |d_i^j| \leq \delta \end{cases} \tag{3}$$

where d_i^j is the wavelet coefficient, j the resolution and $i \in \{x, y, xy\}$.

For our experimentations, we used the default parameters of the Meyer decomposition [2, 10]. Figure 7 shows the application of the gray scale decomposition model to an image.

⁷ The Legendre–Fenchel transform of F is given by $F^*(v) = \sup_u \langle (u, v)_{L^2} - F(u) \rangle$, where $\langle \cdot, \cdot \rangle_{L^2}$ stands for the L^2 inner product [22]

⁸ $\lambda B_G = \{f \in G / \|f\|_G \leq \lambda\}$.

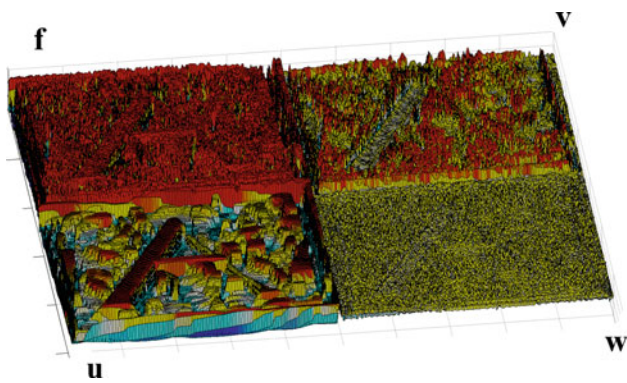


Fig. 6 3-D representation of image (7)a decomposition. From *top to bottom* and *left to right*: original noisy image (f), regular or geometrical part (u), Texture component (v) and noise component (w)

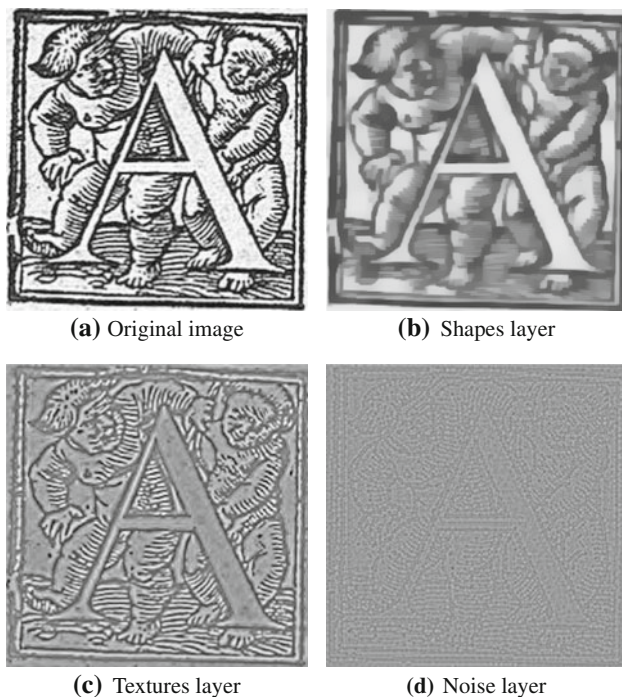


Fig. 7 An example of drop cap decomposition using Meyer decomposition

3.1 Layers in details

The three layers extracted using the implementation in [10] of the Meyer decomposition (Aujol and Chambolle algorithm) can also be seen as follows:

- The Regularized Layer corresponds to the area of the image that has low-level fluctuations of gray level. This layer enables us to highlight geometry, which corresponds to shapes in the image. In the rest of this paper, we will call this layer the “*Shape Layer*”.
- The Oscillating Layer, which corresponds to the oscillating element of the image. In our case, this layer highlights

the texture in drop caps, and in the rest of this paper, we will call this layer the “*Texture Layer*”.

- The highly Oscillating Layer, which corresponds to noise in the image. In fact, this layer recovers everything that does not belong to the first two layers. This layer, therefore, incorporates noise, background text, and the problem of aging paper or document. Our goal is to recognize images in the old document images while being robust toward noise variations. For this reason, we do not refer to this layer in the next sections of this paper.

An example of decomposition is given in Fig. 7.

Specific Treatment Each layer will be seen as an image composed of uniform elements (the first layer consists only of shapes, and the second consists only of textures). In the case of the regularized layer, we use a Zipf’s law to model the distribution of patterns.

4 Zipf’s Law

4.1 Introduction

As mentioned, in the Meyer-based segmentation part, the shape layer issuing from this stage has to be considered for extraction of the letter. For this purpose, since component extraction is not adapted, we propose to use a complementary Zipf’s law based segmentation stage in order to extract large and homogeneous regions which correspond to the letter.

In order to become robust with respect to noisy conditions and detail changes, we chose a statistical approach based on the frequency of the patterns. The variety of colors involved in drop caps is considerable and never known. Besides, the letter is more than often in a single tone within an image. We then looked for a model to approximate the distribution of patterns present in the drop caps. This model is Zipf’s law.

In this section, we will review Zipf’s law and its application to images, in particular to drop caps; in fact, a mixture of laws is observed. We then compare the results according to the nature of the image.

4.2 Zipf’s Law

Zipf’s law [24] is an empirical law formulated fifty years ago, which relies on a power law. The law states that in phenomena described by a set of topologically organized symbols, the distribution of the occurrence numbers of n -tuples named patterns is organized in such a way that the frequencies of the patterns M_1, M_2, \dots, M_n , denoted N_1, N_2, \dots, N_n , are related to the rank of these symbols when sorted with respect to their frequency occurrence. The following relation holds:

$$N_\sigma(i) = k * i^a$$

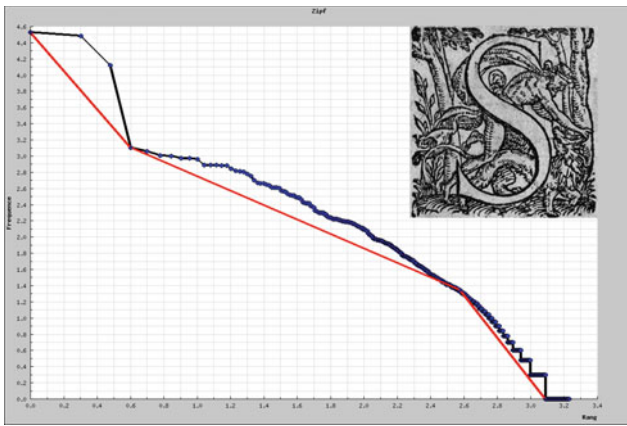


Fig. 8 Example of a Drop Cap and its Zipf plot, showing the various straight zones extracted

$N_{\sigma}(i)$ represents the occurrence number of the pattern with rank i , and k and a are constants. This power law is characterized by the value of the exponent a ; k is more closely linked to the length of the symbol sequence studied. The relationship is not linear but a simple transform leads to a linear relationship between the logarithm of N and the logarithm of the rank. The value of exponent a can then easily be estimated from the leading coefficient of the regression line approximating the experimental points of the 2D graph $(\log(i), \log(N_{\sigma}(i)))$ with $i = 1$ to n . Below, the graph is called the Zipf graph. An example is given in Fig. 8. As one can see in Fig. 8, Zipf's law graph can be approximated by a set of three linear functions. This point has been observed for all our experiments as we will see in the next part. If one wants to approximate these linear functions, one way to achieve the approximation is to use the least-squares method. Since the points are not regularly spaced, the points of the graph are rescaled along the horizontal axis before approximation.

4.3 Image decomposition: layer extraction

In this section, we note some problems that may arise with images. In the case of unidimensional data, the only possible tuples are limited to successive symbols. When images are concerned, they must be replaced by masks respecting the topology of the 2D space in which the data are embedded. We have chosen to use 3×3 masks as the most common neighborhood of a pixel in a 2D space.

The principle then remains the same: The number of occurrences of each pattern is computed. Nevertheless, since 256 symbols are used to code pixels, there could theoretically be 256^9 different patterns. This number is much larger than the number of pixels in an image. Indeed, if all patterns are represented only once, no reliable model can be deduced, and the statistics will lose their significance. For example, a 640×480 image contains only 304,964 patterns.

It then becomes necessary to restrict the number of perceived patterns in order to give sense to the model. The coding is decisive in the matter.

The aim is to find the most suitable method of encoding to define indexes capable of distinguishing each component of a drop cap.

Some studies have shown that a Zipf's law applies to the case of images with different encoding processes [5]. We are looking for a coding process that produces models capable of distinguishing the images we are studying. This qualifies as an effective coding process.

The drop caps we are studying have been scanned as gray level images where each pixel is encoded with 8 bits (256 different levels). The intensity is the information encoded.

As indicated by our previous remarks, the number of different symbols must be reduced. Two ways to achieve the reduction are possible: either the number of symbols used to characterize the pixel is reduced or the number of the pixels involved in the mask is reduced.

Here, our motivation is to preserve the pattern of a scene that relies more on the differences of gray levels than on their absolute values.

A simple approach would be to use only k gray levels to characterize the intensity level of the pixels. Most often, a small number of gray levels is sufficient to observe an image keeping the significant details. As a quantization into k equal classes would lead to unstable results, we have chosen to use a method for classifying the gray levels into k classes by way of a k -means algorithm [15]. We experimented with various values of k and decided to keep only three color levels. This is in accordance with the appearance of these historical images. In Fig. 8, we show an example of a Zipf curve obtained with the 3-means algorithm.

A closer examination of the curves shows they are not always linear throughout, *i.e.* Zipf's law does not hold for the whole pattern set. Nevertheless, a number of straight line segments can be observed. This means that several structures can be observed in the distribution of patterns. This is quite natural, since several things appear in a drop cap, at least the letter and the background, which may be very complex. A Zipf's law cannot properly approximate the distribution; however, a mixture of such laws may be considered. Depending on the coding process used, these zones (that is to say the corresponding zones in the image) may be interpreted. We observed that the left-hand portion of the graph was concerned with the regions in the image, whereas the right-hand portion gave information on the contours present in the image. We can extract from both some structural information about the regions and also the structure of the contours within the images. This can be called a Zipf decomposition of an image, as presented in (4), where I_0 corresponds to the original image and L_n to the n th layer with $n \in [1; 3]$. Here, the sum is an exclusive sum. We obtain a partition of

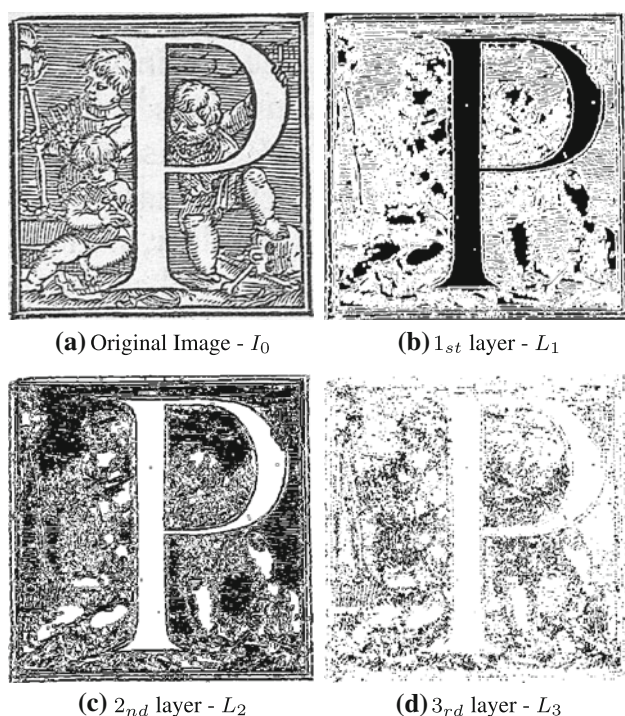


Fig. 9 An example of three layers extracted from images. Each black pixel corresponds to a pixel in the original image that belongs to the layer

the image.

$$Z(I_0) = L_1 + L_2 + L_3 \quad (4)$$

Since letters are often very regular, uniform zones, the patterns they contain are the most frequent and the zone in the curve, associated with the letter is the left-most portion; however, many zones are present. The various linear zones are automatically extracted as shown in Fig. 9, using a recursive process. The splitting point in a curve segment is defined as the farthest point from the straight line linking the two extreme points of the curve. Note that the image carries a mixture of several phenomena, which are highlighted in the expression. Several power laws are involved, from which several exponent values can be computed.

It is possible to recover from the image the pixels that are contributing to the various zones of the Zipf curve. We call them layers, and the image is then decomposed into several layers. The patterns involved in the first layer (associated with the LH portion of the Zipf curve) are texture patterns, whereas the patterns associated with the RH portion are contour patterns. An example is shown in Fig. 9. The image identified as the first layer consists of pixels obtained from the pattern of the first Zipf's law, second layer with pixels from the second power law, and the third layer with pixels from the third Zipf's law.

The method can characterize not only the image's overall appearance but also its structural composition. We note that the first layer comprises the letter and large areas derived from the background. The second layer is made using thick outlines and the third one, with thinner outlines. The first layer is the most interesting in terms of our present objective.

Our method is not driven by the gray level of the letter in the drop cap. In spite of the various possible colors of the letter, all the patterns belonging to the letter will be included in the first portion of the Zipf curve. This is one of the main reasons for deciding to use a model that relies on the frequency rather than on the overall aspect of the patterns.

5 Letter extraction

In the two previous sections, we have discussed methods that can address the difficulties encountered in the extraction of letters from lettrines. Nevertheless, neither of the transforms provides an obvious criterion for extracting the letter.

Indeed, we have two decomposition methods with two distinct objectives. One method considers the image as the superposition of three different signals (Meyer-based approach), while the other considers the image as the juxtaposition of different objects that carry different meanings (Zipf's law). Of course, neither of these points of view is correct, and we will benefit from the cooperation between the two methods. Applying the first decomposition allows to have a more pure signal before considering the partition of the image.

In case of the Meyer decomposition, it is difficult to determine whether the letter is obtained from the lightest or the darkest pixels, but textures are separated from shapes. From the first layer modeled with the most frequent Zipf's law, one can separate the connected components of lettrines but they do not fit well with the letter, their area is often very small and the largest is not always the letter, owing to a lack of regularity in the original image. In what follows we attempt to coordinate the two approaches.

When the first layer of Zipf decomposition is extracted from the regularized layer of the Meyer transformation, the connected components are more regular and have a larger area. An example of Zipf decomposition result obtained from the original image and from the regularized layer of the Meyer transformation can be observed in Fig. 10.

The largest connected component extracted begins to acquire some semantic meaning, for example, if the largest one covers a significant portion of a face or a letter. Nevertheless, the letter is not always the largest connected component of the first Zipf layer of the regularized layer in a Meyer transformation. In this case, criteria have to be established for distinguishing between the letter and elements of the background. For reasons of legibility, the letter is a large element,

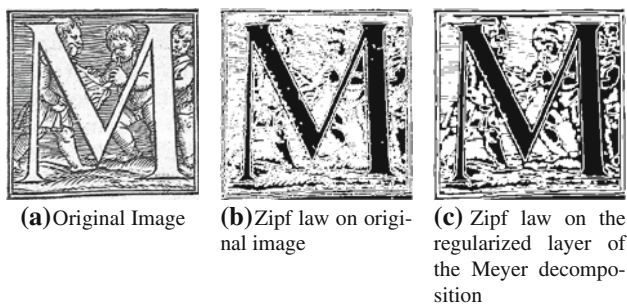


Fig. 10 An example of Zipf decomposition result obtained from the original image and from the regularized layer of the Meyer transformation. We can observe that the connected components extracted from the regularized layer of the Meyer transformation are more regular and they have a larger area. This allows to get a better visual understanding of image

it is centered in the drop cap, and it does not touch the border of the drop cap. We are searching among the largest centered connected components that do not belong to the boundaries, in order to extract the letter. Figure 11 presents some results progressing from the original image to the letter, displaying the image at each step (Meyer shapes layer and first Zipf layer obtained from the previous example).

6 Experiments and validation

Based on the appropriate sequencing of Meyer decomposition and Zipf's law analysis, we developed a system allowing to extract the letter from the lettrines.

The evaluation of such a system is a fundamental step because it guarantees its usability to the users and because it provides an objective view of the system. In the context of such a project, the implementation of an objective evaluation system is rather difficult, because of the variability of the users' requirements: Research historians, and net surfers, are likely to search for very different kinds of information.

In the context of the NaviDoMass project, and more specifically for the purpose of dropcap indexing, we decided to evaluate the quality of our system by considering the objective of "Letter Based Retrieval". This choice is motivated by the fact that many historians want to be able to retrieve drop caps in terms of this criterion. Accordingly, the evaluation of our system relies on the application of an OCR system at the conclusion of the letter segmentation. As a result, the classification rate is the main performance evaluation criterion for our system.

We conducted different series of tests in order to validate our choices. The first series of tests consist in highlighting the difficulties of letter extraction by comparing our results with a classical Otsu binarization process [17].

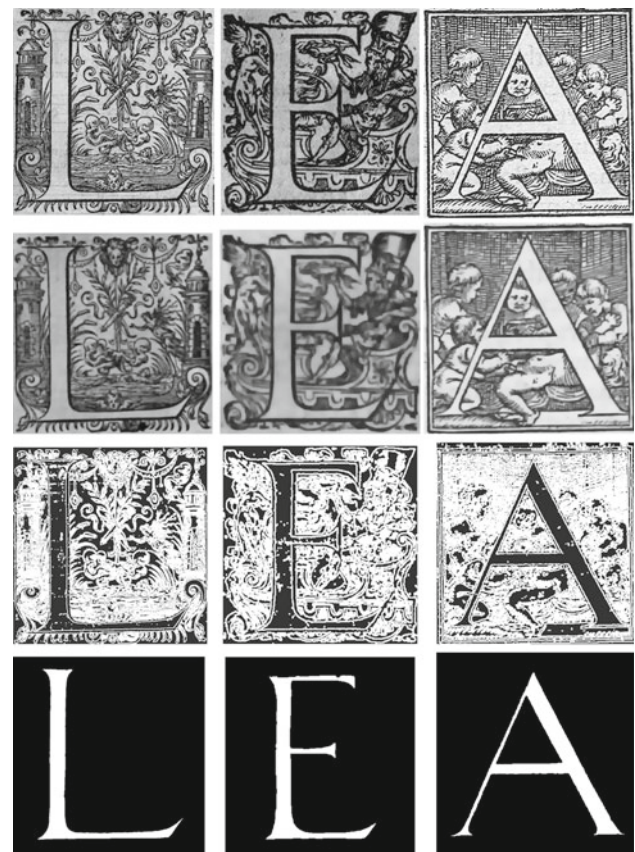


Fig. 11 Examples of letters automatically extracted from drop caps. The first row corresponds to the original images, the second one to the shape layer of Meyer decomposition, the third row to the most frequent patterns extracted from the original images, and the last row to the connected component selected

Both approaches are followed by an extraction and selection of connected components. Then, we compare results obtained by two OCR on a set of lettrines. For the evaluation, we used a commercial OCR system (FineReader), as well as an open-source system (Tesseract). We applied our approach to an image database containing 4,293 images: 1,293 of these images were used for the training set and 3,000 for the tests. The results are summarized in Table 1.

We observed that the recognition rates obtained by using the Zipf algorithms are two times better than those obtained using Otsu binarization. Figures 12 and 13 show some examples of letters extracted in order to illustrate our approach. We see in Fig. 12 that extraction of connected component requires to know a priori the color of the connected component to be extracted. In the case of Otsu's approach, black letter cannot be extracted. Thanks to Zipf's law abilities for extracting most frequent patterns, our approach permits to dynamically adapt the letter extraction. As a letter is a wide uniform area, it is highlighted regardless of its color.

The second major difference: Extraction of greater part of a connected component is presented in Fig. 13. In this

Table 1 Recognition rates for letters in letrines using two different approaches of letter segmentation and two kinds of OCR

	FineReader (%)	Tesseract (%)
Using Otsu criterion		
Classification Rate	20.76	22
Using original images with Zipf law		
Classification Rate	48.2	39.6

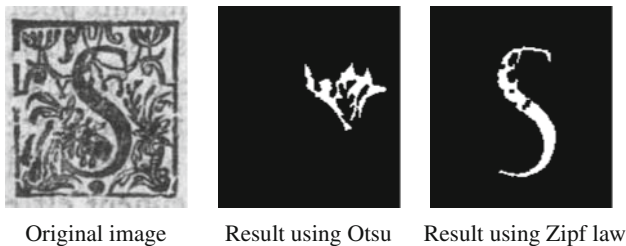


Fig. 12 Comparison between Otsu binarization and Zipf's law segmentation. The main difference relies on the fact that Otsu method binarize image and *black letters* cannot be extracted

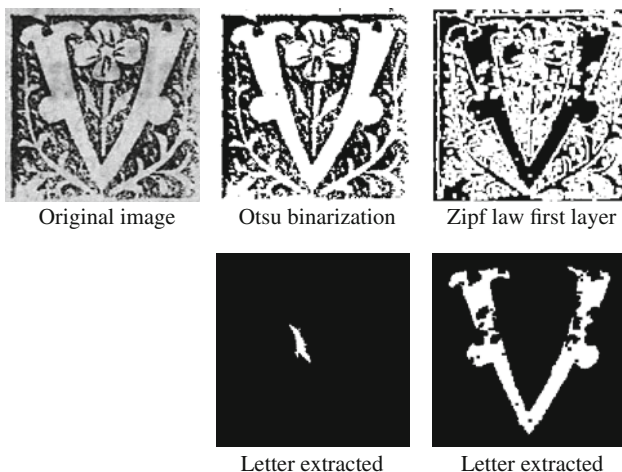


Fig. 13 Comparison between Otsu binarization and Zipf's law segmentation. Otsu binarization only associate each pixel with the background or foreground from its value. Zipf's law segmentation relies on pattern frequencies, so large area like letter can be extracted even if it is connected to background (link correspond to pattern with low frequency)

figure, the same image has been segmented using Otsu binarization and Zipf's law segmentation, and the letter extracted is presented. We can see in the original image that background is degraded, and letter is connected to the outline of letrine. With the Otsu binarization, letter stays connected to the outline, while Zipf's law segmentation leaves out pixels of the link.

From this first series of experimentations, we deduce that Zipf's law approach give better results for connected components extraction. We thus decided to use it in the second series of experimentation. These tests consisted in

comparing Meyer decomposition to a wavelet approach to assess the better adequacy of Meyer decomposition to an extraction of connected component. Table 2 presents results obtained with the database and OCR of the first experiment.

As we predicted, the recognition rates obtained by using the combination of Meyer and Zipf algorithms are better than those obtained using Haar wavelet and Zipf algorithm. Figure 14 shows the same letter extracted using both approaches. At each level of wavelet decomposition, the smoothing function slims down boundaries and open regions. In the example presented in first line, letter M is split into many parts and cannot be properly extracted. In the case of Meyer decomposition, regularized image is obtained after a projection in a space that preserves boundaries while smoothing uniform areas. Letter extracted is then complete and recognized by OCR. Moreover, with Meyer decomposition, the original image is denoised and the variation of gray levels is reduced. The letter is more perfectly extracted, and OCR is able to recognize it.

Finally, the cases for which our system fails correspond to very difficult images, where the letter is composed not of one but of many connected components. Separation into several connected components can be explained by two phenomena: the degradation of the paper over time and the stylistic effects. Some examples of poor extraction results are shown in Fig. 15. The first case highlights a stylistic effect used to represent the letter, and the two other cases show degraded letrines. To further develop this approach, we are considering groups of connected components that resemble the selected one, that are in its neighborhood, and that share similar features (for example the same principal orientation or the same width).

7 Conclusion

In this paper, we consider the problem of extracting letters from drop-cap images. Our strategy consists of decomposing the information into several layers, as a function of the identified features for each of these layers. Based on the frequency properties of the information, the Meyer-based decomposition allows a separation of the various layers, in which the information can be more easily analyzed. Then, based on the use of a Zipf modeling, which enables the extraction of areas

Table 2 Recognition rates for letters in letrines using two different approaches of image simplification and two kinds of OCR

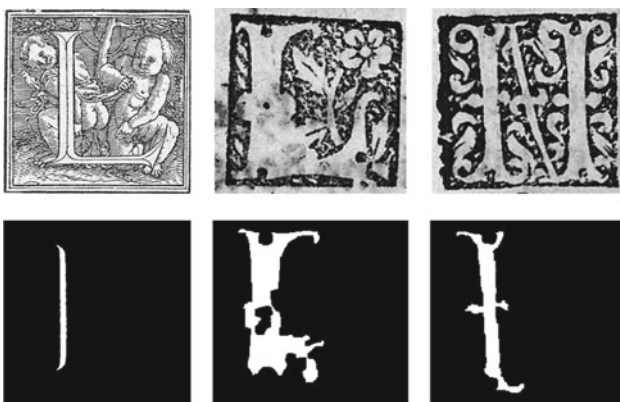
	FineReader (%)	Tesseract (%)
Using Haar wavelet with Zipf law		
Classification Rate	29.1	27.8
Using Meyer's shapes layer with zipf law		
Classification Rate	72.8	67.9



Extraction of Zipf first layer and letter from original image



Extraction of Zipf first layer and letter from Meyer shapes layer

Fig. 14 Examples of letter extractions. First column corresponds to the image regularized, the second column corresponds to image after applying the Zipf algorithm, and the third column corresponds to letter extracted. Each row presents the result for each approach tested, and the difference between two results explains differences in recognition rates**Fig. 15** An example of very difficult letter extraction. The first line corresponds to the original image and the last line to the letter extracted

potentially corresponding to the letter, a letter extraction can be performed.

Evaluation of the overall process highlights the relevance of the approach, by enabling the extraction of letters from a significant set of data. Further development of this work may require the use of other layers (textures and noise), in order to complete the indexing process. In addition, we will

consider the use of visual keywords for the various layers, in order to achieve a comprehensive and practical indexing process.

References

1. Aujol, J.F., Aubert, G., Feraud, L.B., Chambolle, A.: Image decomposition into a bounded variation component and an oscillating component. *J. Math. Imaging Vis.* **22**(1), 71–88 (2005)
2. Aujol, J.-F., Chambolle, A.: Dual norms and image decomposition models. *Int. J. Comput. Vis.* **63**(1), 85–104 (2005)
3. Aujol, J.-F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decomposition—modeling, algorithms, and parameter selection. *Int. J. Comput. Vis.* **67**(1), 111–136 (2006)
4. Bigun, J., Bhattacharjee, S.K., Michel, S.: Orientation radiograms for image retrieval: An alternative to segmentation. In: *International Conference on Pattern Recognition*, vol. 7276 (1996)
5. Caron, Y., Makris, P., Vincent, N.: Use of power law models in detecting region of interest. *Pattern Recogn.* **40**(9), 2521–2529 (2007)
6. Chambolle, A.: Total Variation Minimization and a Class of Binary MRF models. *EMMCVPR*, 3757 of *Lecture Notes in Computer Sciences*, pp. 136–152 (2005)
7. Chouaib, H., Cloppet, F., Vincent, N.: Graphical Drop Caps Indexing. In: *GREC 212–219* (2009)
8. Coustaty, M., Ogier, J.-M., Pateri, R., Vincent, N.: Drop caps decomposition for indexing a new letter extraction method. In: *10th International Conference on Document Analysis and Recognition*, pp. 476–480, Barcelona, Spain, IEEE Computer Society (2009)
9. Delalandre, M.: Retrieval of the ornaments from the Hand-Press period: an overview. In: *International Conference on Document Analysis and Recognition*, vol. 2, pp. 496–500, Barcelona, Spain (2009)
10. Dubois, S., Lugiez, M., Péteri, R., Ménard, M.: Adding a noise component to a color decomposition model for improving color texture extraction. In: *4th European Conference on Colour in Graphics, Imaging, and Vision*, pp. 394–398 (2008)
11. Hamidi, A.E., Menard, M., Lugiez, M., Ghannam, C.: Weighted and extended total variation for image restoration and decomposition. *Pattern Recogn.* **43**(4), 1564–1576 (2010)
12. Journet, N., Ramel, J.-Y., Mullot, R., Eglin, V.: Document image characterization using a multiresolution analysis of the texture: application to old documents. *IJDAR* **11**(1), 9–18 (2008)
13. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal. *Phys. D* **60**, 259–269 (1992)
14. Mallat, S.: *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, London (2008)
15. McQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (1967)
16. Meyer, Y.: Oscillating patterns in image processing and nonlinear evolution equations. *The fifteenth dean Jacqueline B. Lewis Memorial Lectures* (2001)

17. OTSU, N.: A threshold selection method from Gray-Level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
18. Pareti, R., Utama, S., Salmon, J., Ogier, J., Tabbone, S., Wendling, L., Adam, S., Vincent, N.: On defining signatures for the retrieval and the classification of graphical drop caps. In: *Second International Conference on Document Image Analysis for Libraries*, pp. 220–231. IEEE Computer Society (2006)
19. Pareti, R., Vincent, N.: Ancient initial letters indexing. In: *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pp. 756–759. IEEE Computer Society, Washington, DC (2006)
20. Remazeilles, C.: Etude des processus de degradation des manuscrits anciens ecrit a l'encre ferrogallique. PhD thesis, La Rochelle, (2001)
21. Salmon, J.P., Wendling, L., Tabbone, S.: Improving the recognition by integrating the combination of descriptors. *Int. J. Doc. Anal. Recogn.* **9**(1), 3–12 (2007)
22. Starck, J.L., Elad, M., Donoho, D.L.: Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Process.* **14**(10), 1570–1582 (2005)
23. Utama, S., Loonis, P., Delalandre, M., Ogier, J.: Segmentation and retrieval of ancient graphic documents. In *Graphics Recognition. Ten Years Review and Future Perspectives, LNCS*. pp. 88–98. Springer, Berlin (2006)
24. Zipf, G.: *Human Behavior and the Principle of Least Effort*. Hafner Pub. Co, New York (1949)
25. <http://www.delos.info/>
26. <http://www.dw-world.de/dw/article/0,1564,1566717,00.html>