10º Encontro Português de Computação Gráfica
1-3 de Outubro 2001

Automatic Feature Extraction on Pages of Antique Books Through a Mathematical Morphology Based Methodology

Isabel Granado Pedro Pina Fernando Muge
CVRM / Centro de Geo-Sistemas, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa
{igranado,ppina,muge}@alfa.ist.utl.pt

Abstract
This paper presents a mathematical morphology based methodology to identify and extract several components on antique printed books in order to automatically build metadata. These components were previously classified into five different sets (drop capitals, stripes, figures, annotations and text matter) each one characterised by particular geometric features. Based on that assumption several novel algorithms appealing to morphological operators are proposed. The evaluation of the methodology is performed on pages of XVI century books.

Key-words
Digital antique books, mathematical morphology, geometric features, feature extraction

1. INTRODUCTION

Antique books constitute a heritage that must be preserved also with the intention of being available to users. In order to consult these works without damaging and degrading them a suitable media format should be constructed and be put at the disposition of interested people over a suitable system. Presently the tendency points to a digital media support eventually accessed through computer networks.

Although fundamental, the intention of creating digital versions of the documents should be enriched with the possibility of being accessed in a interactive form by users, allowing their consultation about, for instance, the presence or absence of certain specific graphical features or even given keywords. For achieving those objectives it is necessary to firstly identify the components that constitute the pages of the books. The possibility of using automatic procedures may be of great interest once the introduction of objective actions may allow the processing of large amounts of data. Digital image analysis techniques have been used widely with this purpose, mainly over the last decade, in order to automatically process digital versions of documents. Different work has been carried out on this difficult subject [Agam96][He96][Montolio93] that is highly dependent on the objectives and requirements proposed. It must be pointed out that there is not a general methodology that can perform the identification and recognition of the components that constitute the pages of antique books and that these approaches seldom appeal to the geometric nature of the structures present in book pages. The motivation for presenting a novel methodology that exploits the geometric features of the images appears this way. Among the several digital image analysis methods available, in order to deal with the geometric features of images, it emerges mathematical morphology. It is a theory used for image analysis created in the middle 1960's by Georges Matheron and Jean Serra from the *École des Mines de Paris*, with the purpose of quantifying structures according to their geometry. Its theoretical evolution over the years has provided to users a powerful tool not only to deal with the geometry of the structures but also with almost all other chapters in image analysis and in modelling and simulation procedures. To get acquainted with the level of generalisation of this theory and with the details related to the definitions readers are advised to consult the seminal work on the subject [Serra82] and the most recent and updated work [Soille99].

2. DIGITAL IMAGE ACQUISITION AND PAGE SEGMENTATION

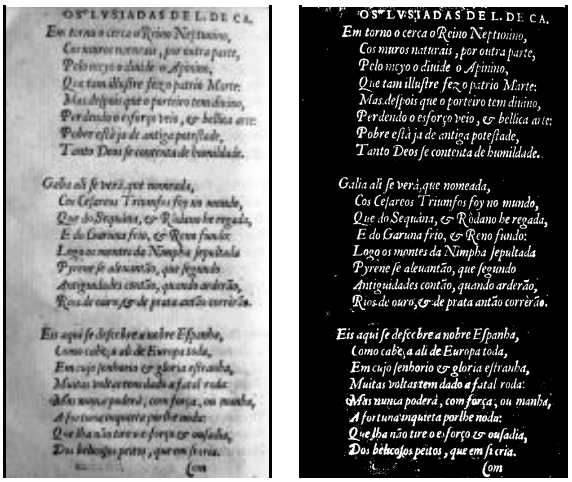
Books in general and antique books in particular must be handled with care to avoid damage. In what concerns digital image acquisition procedures, not only special illumination must be used (for instance, cold lights) but also any contact with glass or any pressure on the bookbinding is forbidden. Once is not allowed to flatten the pages on the digitisation process, geometric distortions appear for certain and, in addition, pages may be degraded due to humidity or to other natural causes. Since the quality of the images depends on the acquisition device, the acquisition task, *i.e.*, the creation of digital images of the pages of antique books in order to become understandable by computers should already

include some procedures, namely to perform geometric corrections and filtering operations. The digitising procedures for obtaining the digital images used in order to develop the current methodology were performed using the *DigiBook* scanner from *Xerox*. The images were acquired in grey level (8 bits) with a spatial resolution of about 600 dpi with the above mentioned pre-processing operations, namely geometrical corrections to attenuate the distortions and filtering to attenuate noise due to the document itself or due to the digitising procedures. Each digital image pixels refers to one complete page of the book varying its dimension from book to book (for instance, each page/image has commonly a typical dimension of about 2000 x 3000 pixels, which occupies 6 Mbytes).

In order to segment the grey level images of the pages, *i.e.*, to create binary images, several approaches were developed (adaptative thresholding, multi-resolution, mathematical morphology) and intensively tested. All of these methods provide advantages as well as some drawbacks in terms of computing time versus the quality of the segmented images. The adaptative thresholding and multiresolution approaches are quite fast but provide sometimes lesser quality binary images, being its description available in [Debora00].

The mathematical morphology approach has a higher computational cost but, generally, provides higher quality segmented pages. The concept guiding this methodology to segment the images refers to a topographic analogy, where the foreground (drop capitals, stripes, figures, annotations and text) can be considered as the "valleys" or "furrows" (darker regions) carved on a generally plane background (lighter regions). Its detection or segmentation consists, in brief, on filling them and subtracting the original image through a procedure inspired on a study developed for segmenting mail letters and postcards [Beucher96] and by introducing some concepts that exploit the information provided by the morphological gradient on the contours of the structures. The complete description of this approach can be consulted on [Mengucci00].

An example of a segmented page using the morphological approach is presented in figure 1. In the original grey level image (figure 1a) the interesting regions (text) are the darker ones where some noisy larger regions originated by humidity appear in dark grey over the background. In the segmented page (figure 1b) the regions that were the direct object of segmentation (in this example, the text) appear in white and the background in black. Although some puzzlement may exist at once due to the creation of a "negative" image, it was decided to maintain the "rule" that is often followed in binary digital image analysis: the regions of interest are represented with the value 1 (white) while the other ones are represented with the value 0 (black). The segmentation procedures do not produce directly binary images free of noise, as can be seen in figure 1b, where some white small regions over the background still remain, but that can be suppressed by adequate binary filtering [Mengucci00b].



(a) (b)

Figure 1 – Morphological page segmentation on a page of the 1st edition of "Os Lusíadas"(1572) by Luís de Camões: (a) initial image (grey level); (b) segmented image (binary).

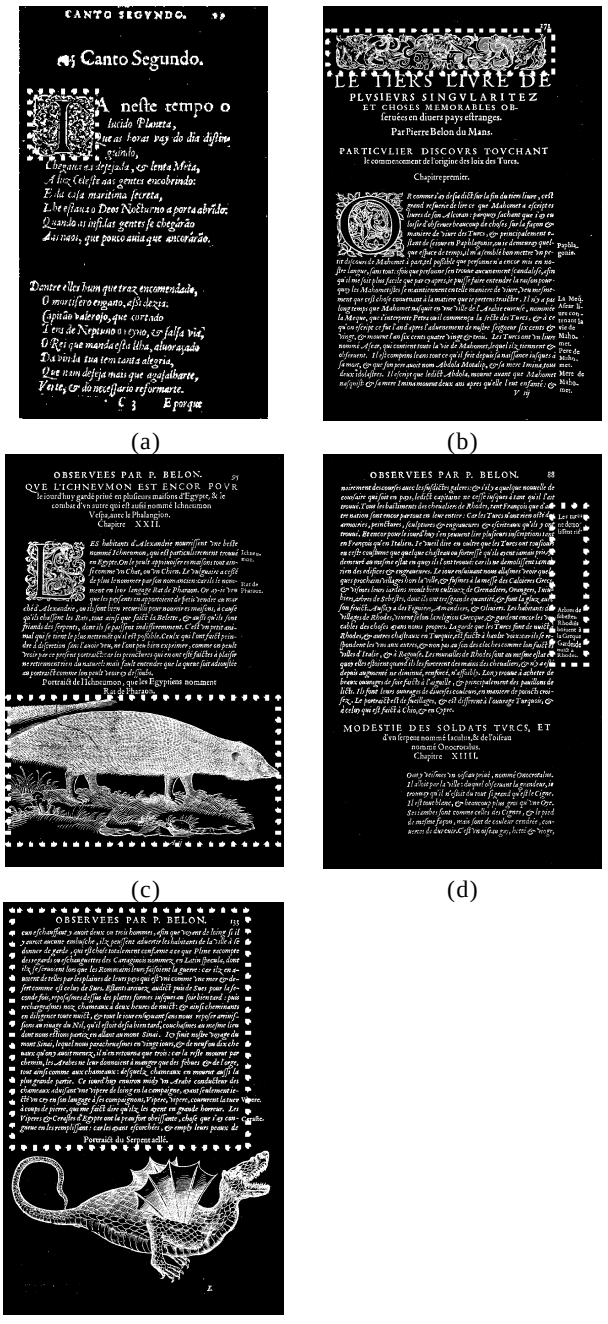
3. PROPOSED METHODOLOGY

A methodology to identify automatically the components that constitute the pages was developed. A page of a book of the XVI century contains several and varied components being previously chosen the following five sets of components, among a larger set, in order to be automatically identified (figure 2): drop capitals, stripes, figures, annotations and text matter. The illuminated letter or drop capital is a capital letter, often ornamental, located at the beginning of a paragraph or chapter (an example delimitated by a dashed square is shown in figure 2a). The strip is a non-decorative element with a long horizontal shape, which adorns the top of the pages, with the purpose of separating chapters and marking the beginning of paragraphs (figure 2b). The figure or illustration concerns all kind of images, engravings, maps or medallions (figure 2c). The annotation or marginal note is a brief explanatory note of text placed in the margins of the text matter (figure 2d). Finally, the text consists on the text matter of the book (figure 2e).

The methodology is hierarchically based once the different components are extracted sequentially (see the scheme in figure 3). The input is constituted by binary images of the pages, being the three levels of the tree processed as follows:

- Level 1 – Consists on the separation of the non-text components (constituted by drop capitals, stripes and figures) from the text components (constituted by annotations and text matter). Although the illuminated letter is a text character its ornamental features indicates that is more suitable to place it in the non-text set.

- Level 2 - Consists on one side (non-text set) on the separation of figures from non-figures (strip and illuminated letter) and on the other one (text set) on the separation of annotations from the text matter.
- Level 3 – At this level the non-figures are separated into stripes and drop capitals. The text matter and also the annotations can then be object of a detailed study, which is out of the scope of the current paper, by detecting the lines and words [Muge00] and by its recognition [CaldasPinto01].



(a) (b) (c) (d) (e)

Figure 2 – Examples of the components of a page delimitated by dashed rectangles/squares: (a) drop capital; (b) strip; (c) figure; (d) annotations and (e) text matter.

4. ALGORITHMS

This phase at level 1 consists on the automatic creation of two sets from the initial binary images: one containing the annotations and text matter (text set), the other one containing drop capitals, stripes and figures.

4.1 Separating text from non-text (level 1)

Based on the segmented pages, the objective in this phase is to separate non-text sets (figures, stripes, drop capitals) from text sets (annotation an text matter). The methodology developed uses, fundamentally, the main directions in which the components of the images evolve, *i.e.*, it is assumed that the characters that constitute the text are aligned along horizontal lines (orientation of 0º) and present a vertical disposition (orientation of 90º). Even for italic fonts this assumption can be considered valid once the "distortion" of these fonts is not accentuated.

Most of the figures are composed of several small lines, whose main directions may vary, not including in the majority of the situations, the ones concerning text characters. Thus, the key-idea is to reinforce the structures oriented on every direction except vertical and horizontal ones, in order to better resist to further filtering operations. On the case of the hexagonal digital grid used, besides its 3 main directions (0º, 60º and 120º) the intermediate directions (30º, 90º and 150º) were also considered.

The developed algorithm consists on the following main phases:

1. Reinforcement of the regions oriented in all directions except the ones of alignment of the text (0º and 90º): applications of directional closings with straight lines as structuring elements. The structures become more compact: holes are filled, concavities disappear or are attenuated and closer objects are connected (figure 4b). This is valid for all the objects present in the image, but at a lower degree to the text once its main orientations are not considered.

2. Suppression of the text set: application of directional openings on the direction normal to the main orientation of the text using a straight line as structuring element with a size equal to the maximum thickness of the lines. The lines are suppressed, remaining the regions bigger than the dimension of the structuring element used (figure 4c).

3. Reconstruction of non filtered regions: application of the geodesic dilation till idempotence. The exact shape of regions that are marked by, at least, a single point can be reconstructed. In this example, only the drop capital and the figure are marked by several sets of points. These structures are reconstructed (figure 4d) while the text, because is not marked by any point, is not recovered. The intersection between the reconstructed image (figure 4d) and the initial one (figure 4a) gives the desired result, an image containing only non-text sets (figure 4e). The text set is now easily obtained by a set difference between the

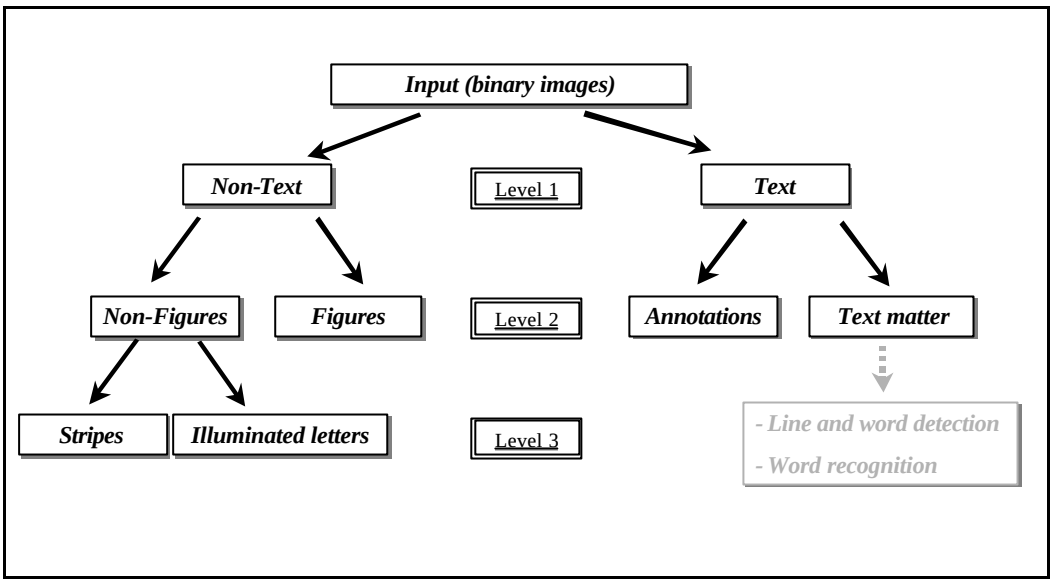


Figure 3 – Schematic methodological approach.

initial image (figure 4a) and the non-text image (figure 4e), whose final result is presented in figure 4f.

Some additional steps of this algorithm, not presented here, are also included in order to deal with some particular features of the images.

Although this algorithm is designed for a general application, it is necessary to adapt its parameter values to each type of book, due to its intrinsic features (typographical and printing features) as well as to its preservation conditions.

4.2 Processing the non-text sets

Within this branch of the hierarchy, the objective is to separate the images that contain only non-text sets, the figures from non-figures (level 2) and, in the following, in the resulting non-figures images, to separate stripes from drop capitals.

4.2.1 Separating figures from non-figures (level 2)

This step consists on creating one set of images containing only the figures and another set of images containing both stripes and drop capitals. The automatic creation of both sets is achieved through the steps of the following algorithm:

1. Creating "strong" objects: It consists on creating one single object per each entity by agglomerating or connecting closer regions (figures and non-figures are constituted by several connected components (figure 5a)). The application of a closing transform with an isotropic structuring element produces the desired result (figure 5b);
2. Creation of a marker of the figures: It is assumed that, at this stage, figures are the bigger structures present in the images. By application of an erosion with an isotropic structuring element (figure 5b), the structures not containing completely at least in some position the structuring element are suppressed. Thus, only some regions of the figures remain and are able to mark them (figure 5c).
3. Identification of figures: It is now simple to identify the figures, by reconstructing the markers obtained in the previous step (figure 5c) in the mask of the strong objects (figure 5b). The set intersection between the initial image (figure 5a) and the reconstructed one (figure 5d) gives the images containing only figures (figure 5e).
4. Identification of non-figures: The set difference between the initial and figures images produces the non-figures sets (figure 5f);

4.2.2 Separating stripes from drop capitals (level 3)

The automatic separation of stripes from drop capitals in images free of text and figures is obtained through the following algorithm stripes:

1. Creation of pseudo-convex hulls: It consists on creating quasi-convex hulls, *i.e.*, single objects for each entity by connecting closer regions through the application of a closing transform (figure 6b) and filling the remaining holes (figure 6c);
2. Creation of markers of the stripes: It exploits the shape of each entity, being the stripes long horizontal structures and the drop capitals more compact (approximately square). The application of a directional erosion along the horizontal direction leaves some regions of the stripes and suppresses all the points of the illuminated letters (figure 6d);
3. Identification of the stripes: Using the markers obtained in the previous step (figure 6d), the reconstruction of these sets with the mask of the pseudo-convex hulls (figure 6c) gives the pseudo-convex hulls of the stripes (figure 6e), whose final shape (figure 6f) is simply obtained by a set

intersection operation between these reconstructed sets and the original image (figure 6a);

4. Identification of the illuminated letters: Once the initial images at the beginning of this step only contain two types of structures, the illuminated letters (figure 6f) are obtained by a set difference between the initial image (figure 6a) and the image containing only the stripes (figure 6e).

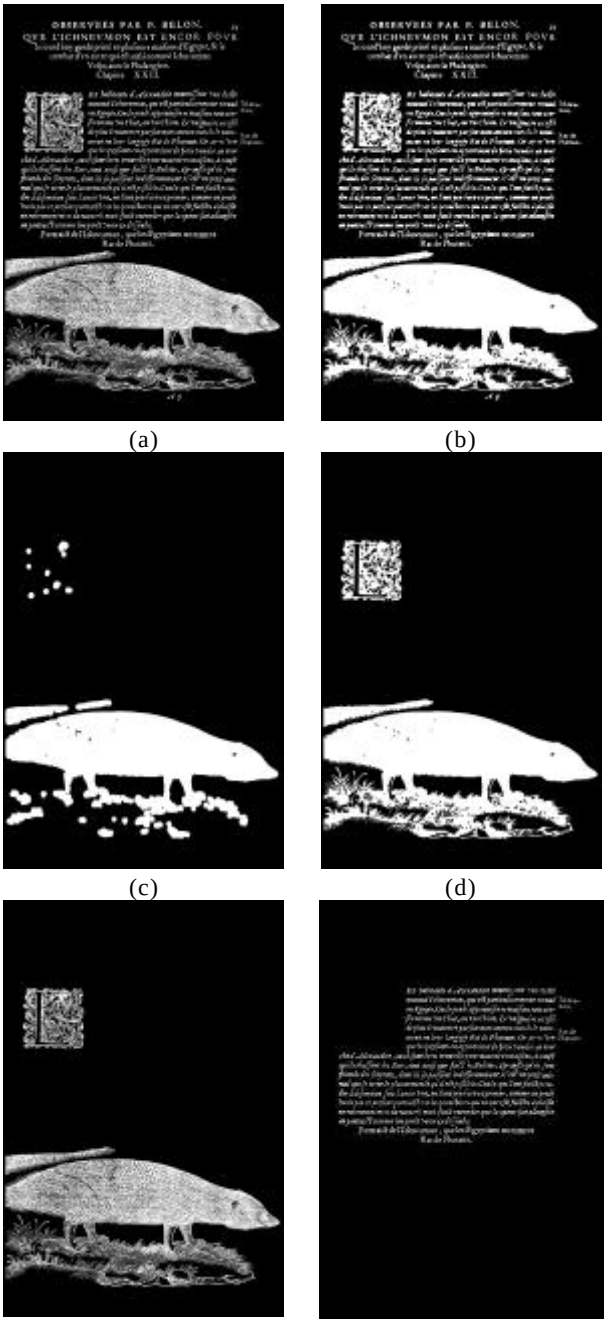


Figure 4 – Algorithm for separating text from non-text: (a) Initial image; (b) Directional closings; (c) Directional openings; (d) Reconstruction of (c) in (b); (e) Intersection of (d) with (a) (final result of non-text set); (f) subtraction of (e) from (a) (final result of text set).

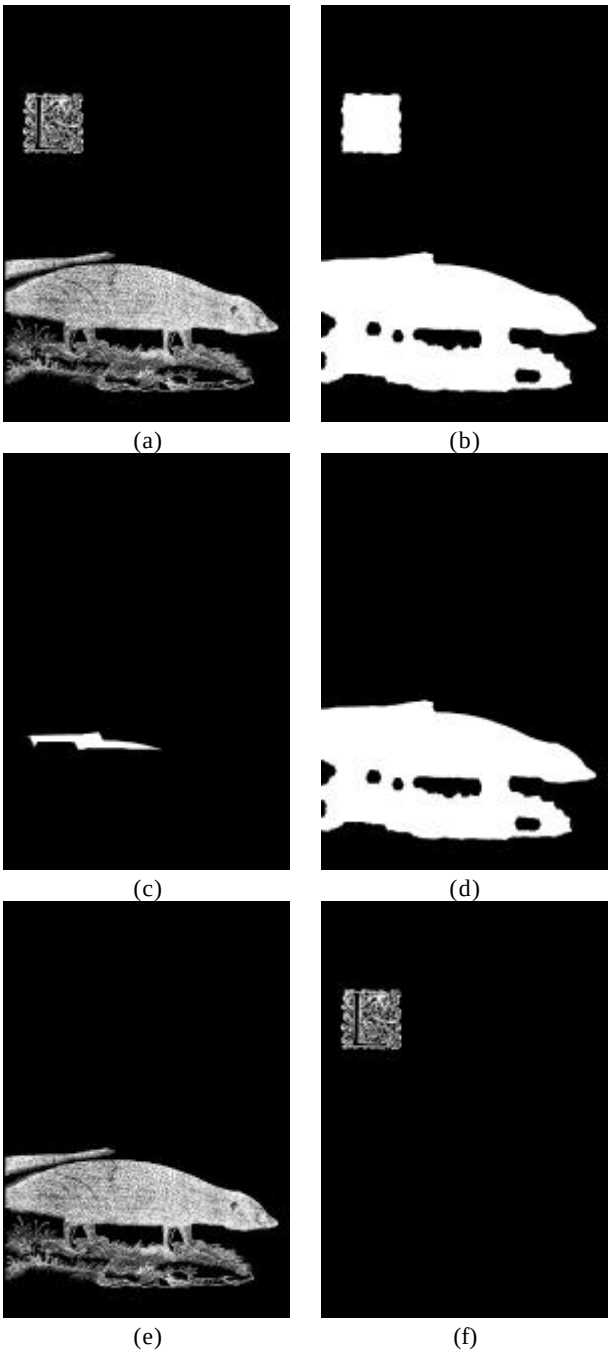


Figure 5 – Algorithm for separating figures from non-figures: (a) Initial image; (b) Closings; (c) Erosion; (d) Reconstruction of (c) in (b); (e) Intersection of (d) with (a)(figure final result); (f) Difference between (a) and (e) (non-figure final result).

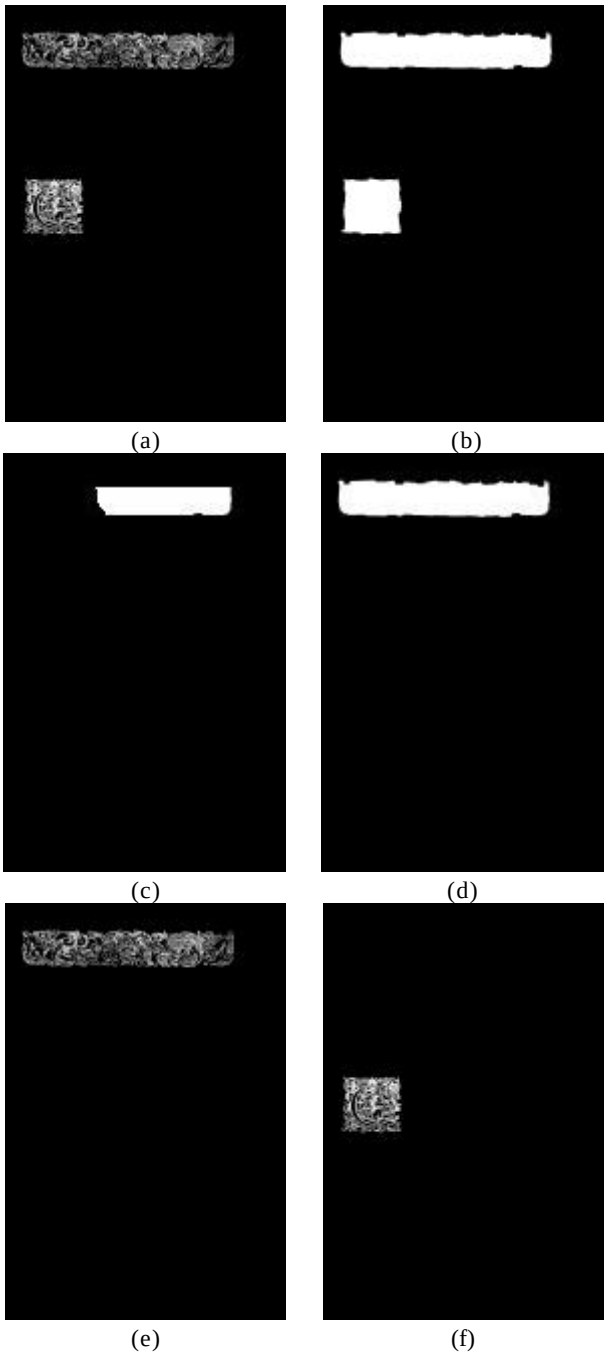


Figure 6 – Algorithm for separating stripes from drop capitals: (a) Initial image; (b) Closing and Hole filling; (c) Linear erosion; (d) reconstruction of strip; (e) Intersection between (a) and (d) (strip final result); (f) subtraction between (a) and (e) (drop capital final result)

4.3 Processing the text sets

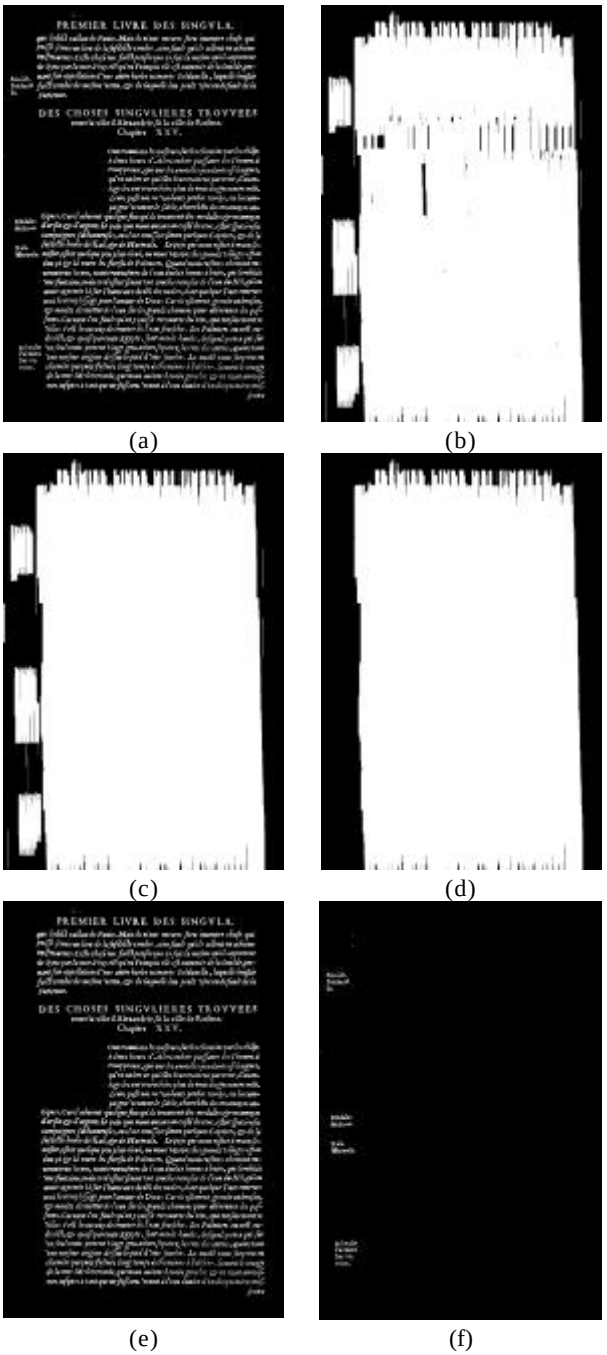
4.3.1 Separating annotations from text matter (level 2)

The separation of these two kinds of text explores the disposition of the annotations and text matter in the pages and also its relative dimensions: the annotations are located in the margins (normally only on one side, on the left or on the right side of the pages) and consist on few lines, while the text matter occupies the central regions of the pages, normally from top to bottom (except the regions occupied by figures and illuminated letters). In order to deal with these geometric features presented by these sets in the pages, the following algorithm was developed:

1. Creation of vertical pseudo-convex hulls: It consists on creating hulls or blocks of the different regions of text. The application of directional closings in the normal direction of the main orientation of the text (horizontal) permits constructing these sets (figure 7b).
2. Creation of markers of text matter: The annotations consist on few lines, being the blocks obtained smaller than the similar blocks of text matter. The application of a suitable erosion permits obtaining markers of text matter (figure 7c).
3. Identification of text matter: The reconstruction of the blocks of text matter produces the masks (figure 7d). The set intersection of these masks with the departure image gives text matter images figure 7e).
4. Identification of annotations: The set difference between the previous images and the initial ones gives the annotations (figure 7f).

4.3.2 Text recognition (level 3)

The recognition of the text in the pages is out of the scope of this paper. The fonts used in printed books of XVI century are not stable, presenting high variations not only from one book to another but also within the same book. The OCR techniques that give good results in recent typographic fonts cannot be applied successfully in these types of books. Several recent studies under development are proposing novel approaches to deal with this difficult task. Readers are advised to consult some preliminary results on [Muge00] and the most recent ones on [CaldasPinto01] where the advances and robustness of the algorithms are successfully increasing the classification rates.



(a) (b) (c) (d) (e) (f)

Figure 7 – Algorithm for separating annotations from text matter: (a) Initial image; (b) Directional closing; (c) Hole filling; (d) Erosion and reconstruction of marker of text matter; (e) Intersection between (a) and (d) (text matter mask); (f) Set difference between (a) and (e) (annotations)

5. CASE STUDY: APPLICATION TO PIERRE BELON'S BOOK

The developed algorithms were tested on an entire digital version of a book previously segmented. The book that was chosen is *Les observations de plusieurs singularités et choses mémorables trouvées en Grèce, en Asie, en Judée* from *Pierre de Belon*, dates from 1554 and belongs to the *Bibliothèque Municipale de Lyon* in France. The available input consisted on 451 pages/binary images of 2084 x 3244 pixels.

The automatic application of the developed algorithms followed the sequence presented in figure 3. In the following, each one of the 451 pages was manually analysed in order to verify the result of application of each algorithm in order to quantify the errors committed. The results obtained at each level for each type of separation are presented in table 1, where the rate concerns the successful cases in number (Nb) and in percentage (%).

Level	*Type of separation*	*Rate*	
		Nb	*%*
1	Text ↔ Non-text	441/451	97.78
2	Non figures ↔ Figures	254/280	90.71
2	Annotations ↔ Text matter	222/298	74.50
3	Stripes ↔ Drop capitals	5/5	100.00

Table 1 – Classification rates obtained on Pierre Belon's book (451 pages)

At level 1, the separation of text from non-text is excellent producing very few errors (classification rate of 97.78%). The situations where the algorithms failed (10 pages out of 451) are very particular and are related to figures that contain text characters inside them.

At level 2, the separation of non-figures from figures was successful in 90.71% of the cases, which is also excellent. The separation of annotations from text matter also gives good results (classification rate of 74.51%), residing the problem from a not very satisfying previous geometric correction. This is problematic once some pages do not present a similar distortion along the pages. Thus, the global geometric correction must be corrected by using other algorithms that can act differentially from region to region within the same image or page. This modification would enable a more correct application of this algorithm.

At level 3, the separation of stripes from drop capitals is completely correct.

6. CONCLUSIONS

In this paper a general methodology to extract the components of pages of antiques books was proposed. It is based on mathematical morphology operators and explores the geometric features of the structures that are present in the books. The classification rates obtained are excellent and indicate that the approach followed is correct. Although this methodology was intensively applied to a single book, its degree of generalisation was also achieved, once pages from other books of XVI century were also analysed with similar successful results.

Anyhow, there are some improvements that still are under development, namely related to the choice of the parameters of each algorithm that vary from book to book. Presently, that choice is done after a previous analysis of the components of the pages of each book. It is envisaged to develop some approach in order to determine the automatic choice of the values of the parameters of each algorithm.

7. ACKNOWLEDEGMENTS

The methodology presented and the results obtained are being developed under the European Union project named DEBORA (1998-2001, contract no. LB5608/A) under the *DGXIII-Telematics for Libraries Programme* (http://www.enssib.fr/divers/debora/).

8. REFERENCES

[Agam96] G. Agam, I. Dinstein, 1996, Adaptive Directional Morphology with Application to Document Analysis, *in P. Maragos, R.W. Schafer, M.A. Butt (eds.),* Mathematical Morphology and its Applications to Image and Signal Processing, 401-408, Kluwer Academic Publishers, Boston.

[Beucher96] S. Beucher, S. Kozyrev, D. Gorokhovik, 1996, Pré-traitement morphologique d'images de plis postaux, in *Actes CNED'96 - 4^{éme} Colloque National Sur L'Écrit Et Le Document*, 133-140, Nantes, France.

[CaldasPinto01] J.R. Caldas Pinto, A. Marcolino, M. Ramalho, F. Muge, N. Sirakov, P. Pina, Comparing Matching Strategies for Renaissance Printed Words, in *10EPCG – 10º Encontro Português de Computação Gráfica*, in this volume.

[Cumplido96] M. Cumplido, P. Montolio, A. Gasull, 1996, Morphological Preprocessing and Binarization for OCR Systems, *in Maragos P., Schafer R.W., Butt M.A. (eds.), Mathematical Morphology and its Applications to Image and Signal Processing*, 393-400, Kluwer Academic Publishers, Boston.

[Debora00] R. Bouché, H. Emptoz (eds.), *Debora, European Project LB5608A*, ENSSIB, Lyon, France, 177 pp.

[He96] S. He, N. Abe, 1996, A Clustering-Based Approach to the Separation of Text Strings from Mixed Text/Graphics Documents, in *Proceedings of ICPR '96 - 13th International Conference on Pattern Recognition* (Vienna, Austria), 706-710, IEEE Computer Society Press, Los Alamitos, California.

[Mengucci00a] Mengucci M., Granado I., Muge F., Caldas Pinto J., 2000, A Methodology based on mathematical morphology for the extraction of text and figures from ancient books, in *Campilho A.C. & Mendonça A.M. (eds.), Proceedings of RecPad'2000 - 11th Portuguese Conference on Pattern Recognition*, 471-476, Porto.

[Mengucii00b] M. Mengucci, I. Granado, 2000, Morphological segmentation of text and figures in renaissance books (XVI Century) in *Goutsias J., Vincent L.& Bloomberg D.S. (eds.), Mathematical Morphology and its Applications to Image and Signal Processing*, 397-404, Boston, Kluwer Academic Publishers.

[Montolio93] P. Montolio, T. Gasull, L. Corbera, F. Marqués, 1993, Character Recognition and Document Analysis by Morphological Techniques, in Salembier Ph (ed.), International Workshop on Mathematical Morphology and its Applications to Signal Processing, Barcelona.

[Muge00] F. Muge, I. Granado, M. Mengucci, P. Pina, V. Ramos, N. Sirakov, J.C. Pinto, A. Marcolino, M. Ramalho, P. Vieira, A.M. Amaral, 2000, Automatic feature extraction and recognition for digital access of books of the Renaissance, in *Borbinha J. & Baker Th. (eds.), Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science - vol. 1923*, 1-13, Springer, Berlin.

[Serra82] J. Serra, 1982, *Image Analysis and Mathematical Morphology*, Academic Press, London.

[Soille99] P. Soille, 1999, *Morphological Image Analysis*, Springer; Berlin.