IJDAR (2007) 9:243–261
DOI 10.1007/s10032-007-0040-6

ORIGINAL PAPER

User-driven page layout analysis of historical printed books

J. Y. Ramel · S. Leriche · M. L. Demonet · S. Busson

Received: 10 February 2005 / Revised: 12 January 2007 / Accepted: 30 January 2007 / Published online: 2 March 2007
© Springer-Verlag 2007

Abstract In this paper, based on the study of the specificity of historical printed books, we first explain the main error sources in classical methods used for page layout analysis. We show that each method (bottom-up and top-down) provides different types of useful information that should not be ignored, if we want to obtain both a generic method and good segmentation results. Next, we propose to use a hybrid segmentation algorithm that builds two maps: a shape map that focuses on connected components and a background map, which provides information about white areas corresponding to block separations in the page. Using this first segmentation, a classification of the extracted blocks can be achieved according to scenarios produced by the user. These scenarios are defined very simply during an interactive stage. The user is able to make processing sequences adapted to the different kinds of images he is likely to meet and according to the user needs. The proposed "user-driven approach" is capable of doing segmentation and labelling of the required user high level concepts efficiently and has achieved above 93% accurate results over different data sets tested. User feedbacks and experimental results demonstrate the effectiveness and usability of our framework mainly because the extraction rules can be defined without difficulty and parameters are not sensitive to page layout variation.

J. Y. Ramel (✉) · S. Leriche
Lab. d'Informatique, Ecole Polytechnique de l'Université de Tours, 64, avenue Jean Portalis, 37200 Tours, France
e-mail: ramel@univ-tours.fr

M. L. Demonet · S. Busson
CESR, UMR 6576 du CNRS, 59, rue Néricault-Destouches, BP 11328, 37013 Tours, France

Keywords Page layout analysis · Segmentation · Digital libraries · Analysis strategy · Zone classification

1 Introduction

High-level analyses of document images are mainly based on the output of a page segmentation process. For example, the extracted text regions can be the input for an OCR system to retrieve the ASCII characters printed on the pages. The spatial relationships between segmented blocks along with other features can be used in logical page organization analysis to group the extracted components appropriately and recover the correct reading order. Many techniques for page segmentation have been proposed in the literature [3,21,14,15,22], but most of them are based on the assumption that an input document image consists of a set of rectangular blocks. Furthermore, the classification step is generally domain specific and uses static rules to automatically determine, for each block, the coherent label selected from a predefined list (title, paragraph, graphic, table, etc.). These limitations appear too restrictive with respect to historical documents and new approaches need to be developed.

In this context, we present a work achieved in collaboration with the "Centre d'Etude Superieur de la Renaissance" of Tours (CESR/http://www.cesr.univ-tours.fr). The CESR is a training and research centre, which receives students and researchers who wish to work on various domains of the Renaissance using a rich library of historical books. The CESR wants to create a humanistic virtual library; however, until now, only bitmap versions of several books that have been scanned or photographed are accessible. The initial objective of the CESR was to obtain an ASCII version of the text

contained in the pages of these historical books. The centre first tried to use the commercial OCR software to index their books, but they quickly realized that, applied to historical documents, this procedure had been an avowed failure. So, the CESR asked our Pattern Recognition and Image Analysis research team to help them to define a new system adapted to their needs. They have appreciated our efforts, as our collaboration will lead to a system able to bring a better description and indexation of the content of their books and would also make the search and the reading of these precious historical books easier.

In this paper, we first study the old book specificities in order to infer some invariant characteristics used during the automatic analysis of documents with complex structures. Next, we describe the layout extraction methods applicable to such documents by focusing on their quality and their drawbacks. We also specify the adaptations we need to make in the classical methods in order to analyse complex documents. The second part of the article proposes a new hybrid method for the extraction of the layout of the documents based on the construction of two representations of the contents of the images. A mapping of the shapes and a mapping of the background are computed. By exploiting this information, our algorithm produces and sends back a list of blocks constituting a first segmentation result.

Then, this initial representation of the image is used during a more sophisticated analysis. Having an aim of genericity, the architecture of the system that we carried out authorizes an interactive installation of scenarios for analysis of the image contents. Scenarios work on the initial representation provided by the first step of the segmentation. According to its needs (localization of the ornamental letters, the notes at margins, titles, etc.) and using user-friendly interfaces, the user (not expert in image processing) builds scenarios allowing to label, merge or remove the blocks contained in the intermediate representation. One can thus locate the desired entities without taking care of the other areas of the image. The elaborated scenarios can then be stored, modified and applied to various sets of images during batch processing.

2 A brief study of the characteristics of historical books and of processing methods

2.1 Characteristics of historical books

2.1.1 Production values

Since its creation, the CESR has established a collection of precious historical books currently numbering around 3,000 copies, dating from the middle of the 14th century to the beginning of the 17th century. The first books belong to the beginning of the printing era when the fonts used and the layouts of the pages were very close to those of the handwritten books. The collection of the CESR is pan European: coming from France, Germany, Italy, Switzerland and Holland. The languages used, Latin or French, bring an additional factor of variability to the books. The Renaissance typographies used have significant variability. Some examples of images of these precious books are presented in Fig. 1.

The technical materials used impose particular presentations in the layout of the pages. The variability of page layouts is much larger than in the current books due either to inaccuracies or to liberties taken by the printer. Most of the time, the text part represents the majority of the page, with notes in the margins on either side of the text. The page can also contain graphical parts of various sizes and some *ornamental patterns*. In the text, we can find known structures like the titles and the subtitles, the paragraphs, the page numbers and other more particular structures like the *catchwords*. These are the words that are placed at the bottom of the page and represent the page number; rather they act as a connection between two pages and usually correspond to the first word of the next page. The catchwords help the printers to organize the manuscript correctly. The styles used can alternate, normal style, justified or aligned on the left. Another characteristic of old printed books comes from weak separations between blocks of text (notes at the margins and body text for example). Lastly, on some pages, current page layout rules are not complied with: for example an illustration can overflow into the margins (Fig. 1). In the Renaissance books, the illustrations were generally printed using metal or wood plates, engraved and inked to reproduce the image. They were generally included in a white rectangle often surrounded by text.

2.1.2 Characteristics due to image acquisition techniques

Other characteristics are due to the digitalization processes. They relate to the lack of lighting in the binding, to the curve of the text lines, to page distortion and to the imperfect elimination of the stains. Much research was undertaken, in particular at the time of the Debora project, to correct these defects [29,20].

2.1.3 Selected invariant characteristics

This study enables us to extract a list of characteristics, which are essential to take into account while designing algorithms for layout analysis of complex documents.

Fig. 1 Examples of pages of historical books illustrating close spaces between blocks, complex layouts with texts and graphics at random positions

Here is the prior knowledge that we used to implement our algorithms:

characteristics depending on printing materials and layout conventions of the Renaissance:

- Complex page layout, which can present several columns with irregular sizes.
- No editorial style or identifiable logical structure.
- The presence of printed or handwritten notes at the margins.
- The presence of location indicators: line numbers, page numbers, catchwords, etc.
- Use of specific and multiple fonts.
- Frequent use of ornaments (non-textual blocks) such as borders, ornamental letters, etc.
- Variable location of graphical illustrations and the associated legends.
- Absence of leading bringing contacts between characters.
- Non-constant spaces between characters, words and blocks of text.
- Text blocks are not rectangular.

Image characteristics connected to the nature of the documents:

- Images still degraded even after restoration (appearance of the characters on the back of the page).
- Presence of superposition of information layers (noise, handwritten notes, etc.)

2.2 Evaluation of existing methods on images of historical documents

Typography and printing technology have progressed enormously and current books meet quite a different standard of presentation than historical books. Consequently, the software built to recognize current documents is often inadequate when processing books of the Renaissance period. The methods of layout extraction employed by this software can be classified into three main categories: bottom-up, top-down and hybrid methods [5].

- To be classified among the bottom-up methods, a method must work starting from the image material in order to obtain more and more interpreted structures using the information provided by the image rather than information about the model of the document. So, most often, such data-driven methods work using an element aggregation strategy. In this category, we find the methods based on different morphological or filtering techniques, methods based on texture analysis and methods based on the study of the connected components in the image. Thus, the problems involved in the bottom-up methods come from many parameters, which are difficult to set and from prohibitive processing times. Moreover, these methods do not allow the analysis of noisy images of documents.
- In contrast, the top-down approaches use knowledge about the document to drive the segmentation of the image (model-driven methods). The segmentation process is very different from the one used by the previous methods because algorithms look for precise information specified in the model of the documents. So, image analysis is achieved by using mostly a split strategy forced by a priori knowledge. Most of the attempts at creating such systems have come to very complex architectures or to applications adapted to a restricted set of documents. Nagy et al. [25] have, for example, developed a syntactic based system working with technical journals. They define a set of appropriate context-free grammars, each defining rules to locate more and more structured entities till up to the logical objects. From the

grammars, parsers are automatically obtained. They are then used to perform segmentation and labelling simultaneously. A set of alternative grammars can be used to allow different document structures to be extracted and checked.

- The hybrid approaches try to use advantages of both top-down and bottom-up methods. Such systems generally include a set of image analysis and page segmentation algorithms that work without requiring any prior information about a specific model of a document. In contrast, a controller has to be defined to select which algorithm should be run at a given moment, according to rules that are valid within a specific category of documents.

Document analysis systems of these two last categories employ, often in an implicit way, a set of typesetting rules and have been designed to operate in a specific domain, mainly due to the test images used. This fact gives rise to a variety of methods, which are characterized by different objectives and different assumptions about input type, layout structure and expected data types. A weak aspect, which is common in all these algorithms is the lack of information about the handling of regions that do not belong to one of the expected categories.

Thus, before describing our system, we present the results of some experiments, which enabled us to evaluate the various traditional methods used for segmentation of traditional documents on historical printed documents.

2.2.1 A few words about pre-processing

Despite the many efforts spent on the development of skew estimation algorithms, every year new algorithms are proposed in literature. This is mainly due to the need of both accurate and computationally efficient algorithms, and of methods that do not make strong assumptions about the specific class of documents they can deal with. Today, commercial solutions exist and are even regarded as satisfactory tools (Book Restorer [8]), although some problems still remain. For example, the correction of page distortion (e.g., the skew) can involve degradation on the borders of the text and the image areas. This is probably due to the general assumption that the text represents the most relevant part of the document image. So, performance decreases due to the presence of other components like graphics or pictures. Furthermore, a major part of the algorithms assumes dealing with documents with a clearly dominant skew angle, and only a few methods can deal with documents having multiple skew angles [26,18,31,23]. In this work, we decided not to spend time on working with these low-level problems, but rather to set up algorithms able to deal with these specificities.

2.2.2 Morphological and differential filtering algorithms

Morphological methods have been implemented and frequently tested mainly for the fusion of characters in words, then in lines and for the elimination of the noise [7,27]. Figure 2 illustrates the results we can obtain using this type of method. The traditional Run Length Smearing Algorithm (RLSA) [35] can also be used; it works and provides similar results to all those morphological methods.

It is also possible to work directly on grey level images by using differential filters [19]. The idea is to use filters that detect and agglomerate the variations of intensity, periodically produced by the contours of the characters, and then to look for horizontal alignments corresponding to text areas.

2.2.3 Methods based on the analysis of the connected components

These methods transform each page into a set of connected components [33]. Figure 3 shows that the size, the proximity and the relative position of the connected components can be used to extract the physical structure of a document. The bounding boxes of the connected components frequently overlap themselves in the graphical parts, but not so much in the text parts. Furthermore, the graphical areas correspond to the connected components whose dimensions (width or height) exceed a fixed threshold.

A text block is a set of small, connected components in close proximity to one another. Two characters are known as neighbours if the distance separating them is lower than a maximum threshold. O' Gorman proposed to use only the neighbourhood between components to locate the text areas in an image [26]. Considering the same idea, we implement a method, which seeks for the components in the four principal directions of each text component. A component that does not have a neighbour in at least one direction is considered as a part of the contour of the block. Then, a closed circular chaining translated the border of a text area (Fig. 4).

K. Kise [17] proposes to use an area Voronoï diagram instead of directly connected components to extract text blocks. The area Voronoi diagram enables in building a neighbour-graph, which represents neighbours of the characters. Then, the task of line extraction is considered to be a path selection problem in the graph.

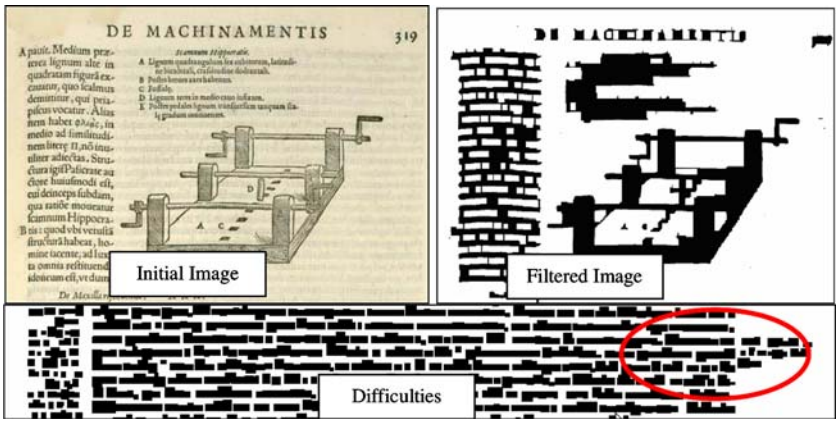

Fig. 2 Use of morphological filtering or RLSA is not sufficient on historical books because of horizontal and vertical proximity between different text blocks

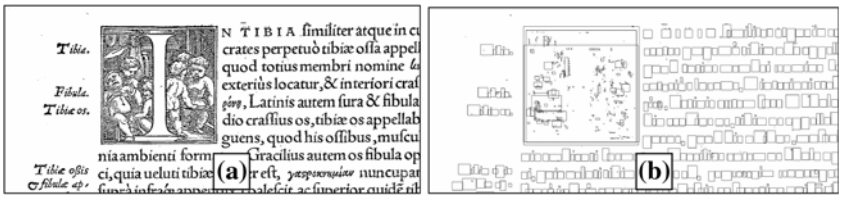

Fig. 3 Use of connected components for segmentation. **a** Initial image. **b** Obtained connected components

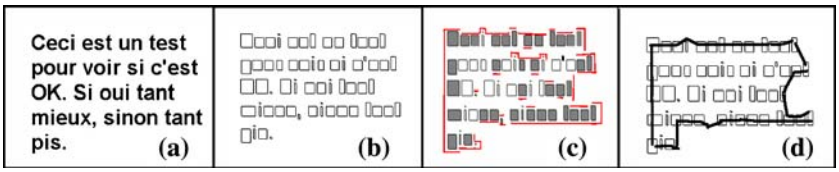

Fig. 4 **a** Initial image. **b** Connected component. **c** *In grey*: components of contour, *in red*: sides without neighbour. **d** Circular chaining obtained

Like the methods based on filtering, these methods are not suitable for historical books processing because of the proximity between some of the text areas in the pages. For the localization of the graphical parts, the thresholds on the size of the connected components must correspond to a size slightly larger than the greatest character contained in the book. However, even if statistical studies can make it possible to automate the selection of these thresholds, these methods hardly support variability in size of the characters and variability in page layout, which can frequently appear in the old books. The bad quality of the images (holes, spots, etc.) can also raise many problems when using such methods.

In conclusion, it seems quite difficult to segment correctly the text paragraphs and the graphical parts in the old documents by using bottom-up approaches, mainly because they use only the local criteria.

2.2.4 Background analysis

Some other methods are based on the assumptions that the white spaces act as separators of the different regions in the document [2,6,3]. In practice, they consist of a hierarchy of constraints on the relative size of spacings between different objects, e.g., spacings between characters are smaller than spacings between words, which, in turn, are smaller than spacings between lines and so on. In [6,3], the search for the white separations could be achieved by using horizontal and vertical projections. Then, the localization of white separations is done by an analysis, of the histogram, of the number of black pixels of the lines and the columns of the image. Next, one can regard the significant difference between successive values in the histogram as delimitation between two blocks of text. The problem is to estimate the significant differences between the histogram values. Generally, a priori knowledge about the document to be analysed (number of columns, margins, etc.) is used to locate separations more easily. In the old documents, the tests we have done using such methods showed that the quality of the results is increased by blackening the bounding boxes of the connected components of the image before calculating the histogram.

The main drawback of these two methods is that they suppose the existence of rectangular text blocks. It is not the case for historical books. Moreover, the problem with this type of techniques relates to the separation between the local and global minima in the histogram. In historical books, certain separations between blocks correspond to non-significant minima (case of badly deskewed images). So, it is difficult to use these methods when page layout is variable or when the pages are badly rectified by pre-processing.

The method described in [2] attempts to locate in a more local way the inter-block spacings instead of the maximum size white rectangles themselves. It used white tiles to extract contours of regions. But this always supposes that two blocks are necessarily separated by a white space from significant surface in either a horizontal or a vertical direction. Unfortunately, as shown in Fig. 1, it is not the case with historical documents, which we have to process.

Among all the top-down methods suggested to segment complex page layouts of composite documents (newspapers, pages of magazines, etc.), we decided to adapt the Split and Merge method suggested by Hadjar [13] during the "segmentation contest" of ICDAR '01. Its goal is to locate the homogeneous white blocks by splitting the image into smaller and smaller areas (Fig. 5). A priori information is needed particularly to define the stop criterion in the recursion. This method can be compared with other more or less advanced techniques like the *Recursive XY-Cut* (*RXYC*) [24,1] method. Readers interested in such a method can also look at [32], where Wang and Srihari compare the RLSA and RXYC approaches. In all the cases, it is the analysis of the neighbourhood (realized with the assistance of graphs or trees) of each white area discovered by splitting that will make possible to locate and characterize the blocks contained in the page in a more or less fine manner. The major problem of such techniques is that they are suitable only for layouts that are decomposable by a sequence of horizontal and vertical subdivisions.

The top-down methods seem to be less sensitive to the noise than the bottom-up methods. When the images are well deskewed, they partly solve the problem of proximity between blocks and characters. Nevertheless, they are not easily applicable to documents having a variable or a non-rudimentary page layout, since they require the use of a priori knowledge [14] about the document (number of columns, splitting criteria, etc.). They are thus not adapted to the processing of old documents, which do not respect a precise editorial style.

2.2.5 Multi-resolution and texture analysis

Jain and Al [11,12] present a multi-channel filtering approach to texture segmentation. The basic assumption is that regions of text in a document image define a unique texture, which can be easily captured by a small number of Gabor filters. As said in [21], one major problem of texture-based approaches is their high time complexity, since different filters are needed to capture the desired local spatial frequency and orientation of the regions. Many masks are needed to extract local features and small masks do not allow one to detect large-scale textures. To avoid this problem, some researches [10] have proposed using a multi-scale analysis. Such methods improve the performance, but are quite time-consuming due to the dependence on thresholds and parameters. Furthermore, in cases of historical documents, different regions can be incorrectly merged always due to proximity between the different text blocks.

To test the texture-based approaches, we implemented a method based on the extraction and the quantification of the various orientations present in different areas of the document images, based on the use of the autocorrelation function [16]. The conclusion of our experiments (Fig. 6) is that text/non-text segmentation in Renaissance documents can be viewed as a texture segmentation problem. We also conclude that the above methods, which consist in analysing the various textures contained in the image, even with a multi-scale approach, are not easily adaptable to obtain precise contours of the text blocks when the text blocks are too close to one another. Small white spaces become undetectable when the resolution decreases and two different text blocks do not always present any significant differences in their texture. These techniques are nevertheless interesting when we try to characterize globally pre-segmented blocks using particular indices.

3 A new method of segmentation

The preceding tests stressed the limits of the traditional methods to segment historical documents as well as the reasons of their failure. On this basis, we developed a new method to exploit the benefits of the top-down methods and the bottom-up methods at the same time. Our method uses a map of the background of the images to highlight the separation between blocks and a map of the shapes present in the image (foreground). Then, we propose to use simultaneously the information provided by these two representations (foreground and background) to segment the image. Thus, we solve most of the difficulties mentioned in the previous chapter.

3.1 Map of the foreground (of the shapes)

The connected components provide relevant information about shapes (graphical and text parts). They correspond to one or more characters, noise or graphical parts. Their positions, size and the overlapping of their bounding boxes provide precise information about the contents of the pages. They can provide local information on each shape present in the image. The connected

Fig. 5 Use of the background for page segmentation of historical books (with a modified version of Split and Merge algorithm)

Fig. 6 Examples of the results obtained with a texture-based approach (Text regions are in *blue* and non-text are in *red*)

components representing the letters of a word are extremely close. Unfortunately, in old documents, the page layout and the spacing between shapes may change frequently. For example, the last letter of a line can be closer to a note in the margin than to the letter, which precedes it on the line. So this information should not be used solely.

To obtain the map of the shapes, after a binarization of the image (using the method described in [30]), we carry out a contour tracking of the shapes which allow us to extract the bounding box from each component. The position and the size of the rectangles are stored in a list, which correspond to the foreground map (Fig. 7a). According to its dimensions, each shape is labelled using one of the following labels:

- *Noise* (connected components of small size),
- *Graphic* (connected components of large size),
- *Text* (connected components of intermediate size).

The thresholds used during this phase are chosen by the user according to the maximum and minimum size of the characters in the book. As discussed in Sect. 2.2.3, the methods based on connected components while using simply their dimensions could not fulfil the whole task. In fact, in our approach, this step provides only

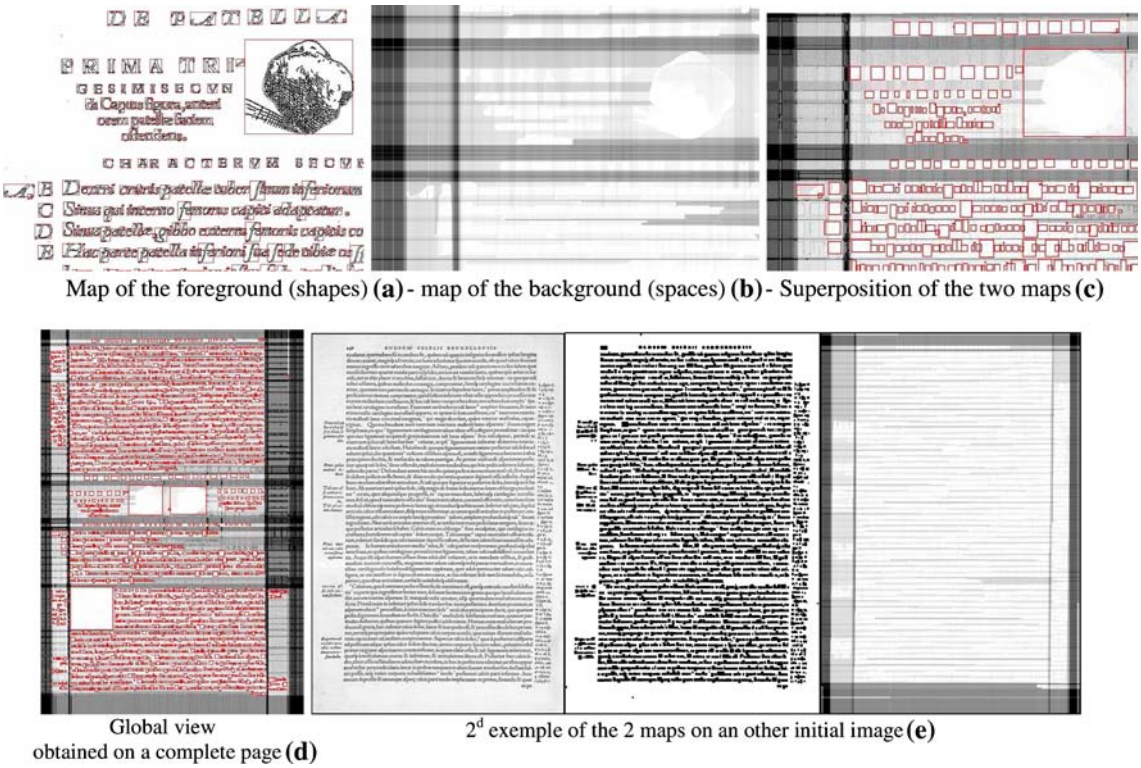

Fig. 7 Maps of the background and shapes for two different images

a first-hand labelling, which will be checked and evolved thereafter during the interactive analysis. Our experiment on old documents shows that it is sufficient to separate text shapes from graphics, but to extract and label precisely different parts of a document (paragraphs, illustration, etc.), more elaborate processing steps are needed. These processing steps are described in Sect. 3.3.

3.2 Map of the background (separations between blocks)

In the bilevel image, the background of the page is represented by white pixels. Normally, a great number of white pixels can be aligned vertically or horizontally in the regions between two blocks of a page (text, graphics, etc.). In the same way, a large number of white pixels are aligned horizontally in the regions between two paragraphs. In contrast, the number of white pixels that we are able to align vertically or horizontally in a paragraph between two letters of a word and between two words of a sentence are small, the same between two lines of the same paragraph.

Consequently, we propose to associate to each pixel $P(i,j)$ of the image the summation of $lgb_h(i,j)$, the number of successive white pixels on the horizontal segment containing the current pixel (line), with $lgb_v(i,j)$, the number of successive white pixels on the vertical segment containing the current pixel (row), to build a distance map. The maximum value for $Ng(i,j)$ is Max = Height × Width (size of the image) and if necessary the lgb_h and lgb_v values can be weighted, respectively, by the width and the height of the image in order to not privileged one of the two directions. This summation corresponds to formula (1) that allows assigning of a value to each pixel in the background of the page (white parts in the map of the shapes in which the interior of the connected components are blackened).

$$Ng(i,j) = Max - \left[\lg b_v(i,j) + \lg b_h(i,j)\right]. \quad (1)$$

After a normalization to 255 of $Ng(i,j)$, we obtain as shown in Fig. 7b a map with grey levels. As Fig. 7b shows, our map of the background translates the separation (more or less underlined) between the blocks of the page.

Baird [4] and Antonacopoulos [2] have proposed a method based only on information similar to that provided by our map since they proposed to extract the white areas from maximum size in a page. But, because of the use of rectangles or of tiles, these methods are more sensitive to the distortion and noise problems than the one proposed here. Moreover, in our case, the white blocks constitute only one part of the information, which

we exploit to carry out the segmentation (and will be combined with the foreground map). In the same way, compared with methods based only on the analysis of the neighbourhood of connected components [17], our approach uses a more global information because each value in the background map depends on the global layout of the page.

3.3 Fusion of the information provided by the two maps

3.3.1 Text/graphic separation

The graphical parts do not always correspond to only one connected component in the map of the shapes. On the contrary, most of the time, they correspond to the superposition of the bounding boxes of several connected components labelled *Text* or *Graphic* In our processing, when two bounding boxes labelled *Graphic* overlap, they are merged in order to produce only one rectangle including the whole of the detected graphical areas. It is also possible that connected components labelled *Text* are present inside an area labelled *Graphic*. These components are then labelled *Text_Graphic* and would be the subject of a particular analysis to determine if they belong indeed to the *Text* class or to the *Graphic* class. For that, the algorithm that we developed uses, once again, the concept of map of the background to carry this classification. The algorithm used to distinguish the *Text* components *close to* or inside a *Graphic* component is as follows:

1. *Text* areas inside *Graphic* areas are relabelled *Text_Graphic.*
2. The *Text* and *Text_Graphic* labelled blocks are coloured white, in the binarized image.
3. Calculation of the background map in the same manner as done previously.
4. *Text_Graphic* blocks positioned over the dark parts of the map of the background are preserved; the others are removed because they are regarded as pieces of *Graphic* blocks.

The removed areas are located in strongly textured areas of the binarized image. We assume that these areas correspond to the presence of graphical parts in the initial image (see Fig. 9).

3.3.2 Extraction of the text areas

By simultaneously using the information provided by the maps of the shapes and the background, we can extract *Text* areas. We start from the list of the connected components labelled *Text* to rebuild the paragraphs of text by association of connected components likely to be characters. To carry out an association, it is necessary that the two components labelled *Text* are rather close and the segment between their centres of gravity (G_1 and G_2) does not cross an important transition area in the map of the background (low grey level values). This multi-criterion constraint can be expressed by:

$$d(G1, G2) \times \left(256 - \underset{(i,j)\in[G1,G2]}{\text{Min}} [Ng(i,j)]\right) \leq \text{Fusion_Threshold}, \quad (2)$$

where d indicates the Euclidean distance between the centres of gravity of the two sets to be associated. When this criterion is checked for two close areas, we give them the same label (attribution of an identical number to each character of the same block of text). If the criterion is not respected, the association is refused. Some other multi-criterion constraints were tested (such as using the sum of the grey level values instead of the minimum, or the use of the distance between the borders of the two areas instead of the centre of gravity, etc.) and we kept the best one. The *Fusion_Threshold* can be chosen manually by the users. The selection of this value is quite straightforward. First, the users are advised to start with a low value (which will result in less fusion of the text part) and then later during the interactive analysis, if necessary, the fusion step can be reactivated for difficult text parts. Such strategies are discussed in detail in Sect. 5.2 (description of the different scenarios).

The search for the neighbours between text parts is done successively, horizontally (horizontal fusion) and then vertically (vertical fusion) in a progressive way, and stopped when no more fusions are possible. These areas are not always rectangular since they correspond simply to a gathering of the shapes having the same labels. The obtained areas can thus have an arbitrary shape. Two examples of the results are provided in Fig. 8. The *Fusion_Threshold* (2) used is the same for the two images. In order to visualize the obtained segmentation, each label given to a component of a given area corresponds to a different grey level value. In example (a), all the blocks were correctly segmented in spite of their proximity. In example (b), the layout of the page and the typesetting rules are much more complex (presence of ornamental letters, images, legends, titles) and our method tends to over-segment the image (there are too many blocks). But, we will see later, this over-segmentation caused by the choice of a strict initial threshold for the fusion is only temporary, since the algorithm of fusion can be applied several times with different

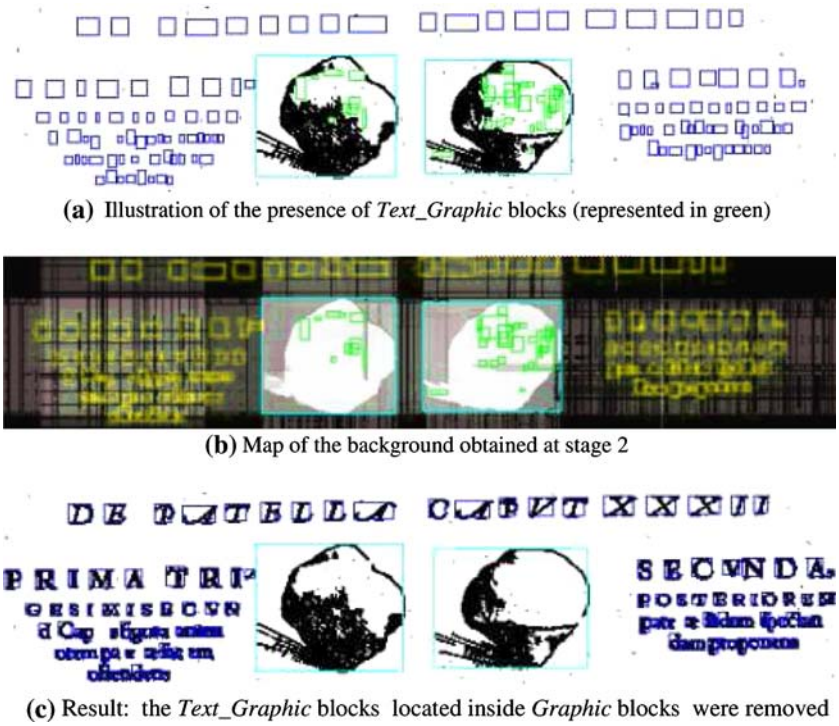

Fig. 8 Illustration of Text/Graphic separation in graphical parts

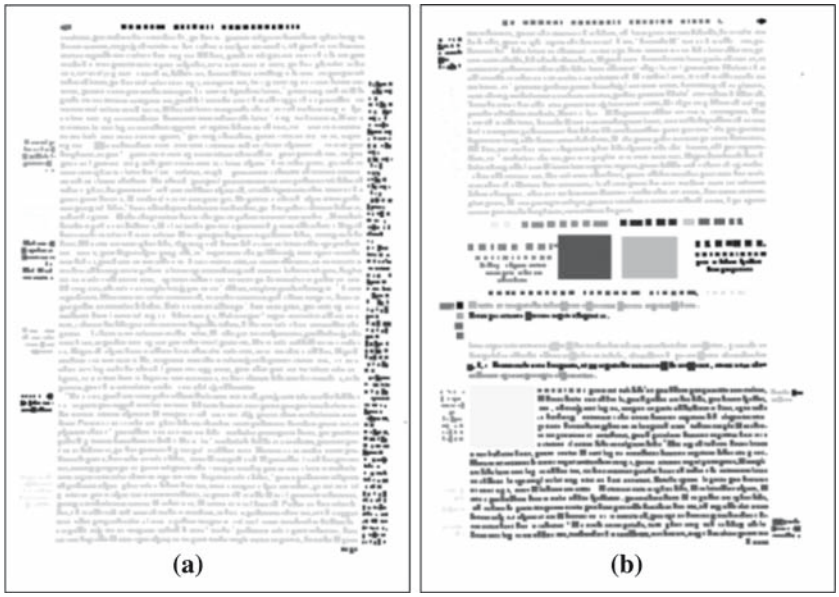

Fig. 9 Examples of segmentation results (each segmented *block* has a different *grey* level value)

parameters (insertion of the fusion step in the scenarios of analysis created by the users).

4 User-driven analysis of the layout

Except for extremely specific applications, current block classification or logical layout analysis methods have shown their limitation. Our proposal consists of a general revision of document recognition methods to design a widely usable system (i.e., not being dedicated to a particular model of documents). For this purpose, block classification needs to be addressed in a more flexible way. So, we are proposing an architecture that drops the strict processing chain to offer an adequate level of human–machine cooperation. An internal representation of the document structure associated with a set of rules can enable an interactive learning of the model of the documents. Furthermore, the users can use a given example of image to define this new model.

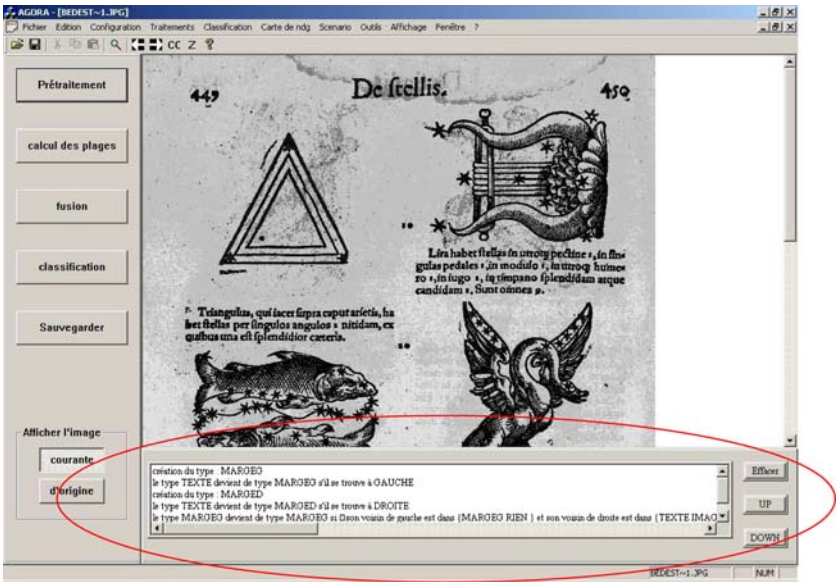

Fig. 10 View of the Agora software. The rules of the active scenario are displayed at the bottom of the window

As shown in Fig. 8, the segmentation stage produces an intermediate representation of the image, providing for each page a set of blocks labelled *Text*, *Graphic*, *Text_Graphic* or *Noise* At this stage of the process, it is possible to ask the user to build different processing sequences (we will name them scenarios of structural analysis), allowing a gradual evolution of this intermediate representation according to the user aims and to the characteristics of the images analysed. He and Downton have already shown that it is sometimes interesting to propose user-assisted document analysis systems [15]. In the long term, our goal is to obtain automatically the most precise labelling of the contents of the digitized pages of complex documents by the application of a scenario that the user builds interactively and easily (user-driven analysis).

4.1 Interactive setting of scenarios for analysis

Once the initial segmentation of the image has been achieved, the architecture of the system (called AGORA) that we propose, makes it possible to continue the analysis in an interactive way on a typical image. For that, we conceived a set of interfaces that allow the user to build, in his own way, scenarios of analysis. The application of a scenario will allow the progressive evolution (incremental analysis) of the contents of the intermediate representation (initial segmentation) obtained before-hand.

The tools placed at the disposal of the user to build the scenarios are:

- A rule editor allowing the refinement of the labels given to the various obtained areas (*Text*, *Text_Graphic*, *Graphic*, *Noise*). These rules are observed in a sequential way according to a strategy defined by the user who can, for example:
  - be interested in the *Graphical parts* only or, on the contrary, in the *Text* parts only;
  - choose to label first the easily characterizable blocks using simple rules; then, he can use a more complete context to extract the less easily identifiable parts of the document [28].
- The application of the algorithms for horizontal or vertical fusion with various thresholds and only on a specific type of blocks defined by the user (e.g., to merge only the blocks labelled *Title*, which were split at the beginning of the process into several blocks, as in the example of Fig. 8).
- Deletion of the blocks of a particular type (e.g., the *Text* labelled blocks if the user is interested only in the ornamental letters).

To build a scenario, the user performs (as for a macro) successive actions, which have to be recorded on a typical image. These actions are applied to the image and the results are displayed in real time. We can thus validate their effectiveness. The actions (rules) are translated in a literal list permanently displayed (see Fig. 10). The user can, then, manage this list to modify the scenario by reordering or removing processing rules. Once the scenario is considered to be correct, it can be saved in a file for storage or for application on a more subsequent set of images (a complete book) during batch processing.

Currently, the usable rules to make the labels of blocks contained in the intermediate representation concern (Fig. 11):

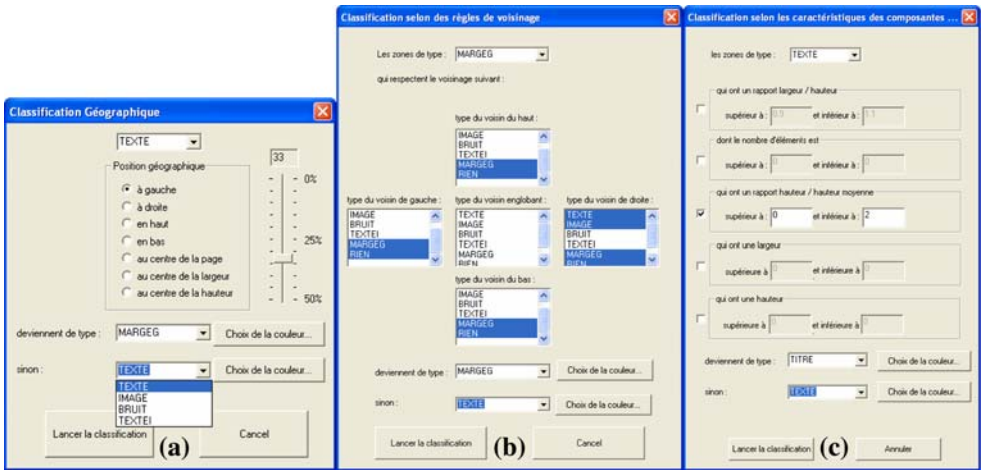

Fig. 11 Interfaces for the installation of position (**a**), neighbourhood (**b**) and intrinsic properties (**c**) rules

- *the topological position* of the blocks,
- *the neighbourhood relations* between identified blocks,
- *the shape and the content* of the blocks.

4.2 Study of the topological position

It is possible to identify a number of invariants on the topological positioning of the objects present in a historical book, making it possible to associate a label with them. The goal is not to use extremely strict rules because page layout can be variable (it would be unwise to specify that the centre of gravity of a note in the left margin will be between the pixel of X-coordinate 205 and 213). However, one can reasonably specify that the position of the centre of gravity of the left margin text is doubtless located in the first third of the width of the page. The suggested interface (see Fig. 11a) makes it possible to take into account the topological position (left, right, top, bottom, centre) of the centre of gravity of the blocks for the evolution of the label. Of course this type of rule will only constitute a first index (provisional label) to correctly isolate an object of a given type.

4.3 Study of the neighbourhood

It is also possible to insert rules in a scenario, which concern the neighbourhood relations between blocks. Examples of rules that allow labelling and extraction of different elements are provided in Fig. 12. For example, the second interface (Fig. 11b) allows in defining rules like:

A *Graphic* block will be labelled *Ornamental letter* if it has:

- at the left, a block labelled *Left Margin* OR *Column* OR *Graphic* OR *Nothing*
- AND at the right, a block labelled *Legend* OR *Column* OR *Graphic*
- AND above, a block labelled *Legend* OR *Column* OR *Title* OR Text OR Graphic
- AND below, a block labelled *Legend* OR *Column* OR *Graphic*

4.4 Study of shapes and contents

Another interface (Fig. 11c) makes it possible to use intrinsic properties of shape to refine labelling. The actual and non-exhaustive proposed criteria are

- the *size* of the block (*width* and *height*);
- the *Height/Width ratio* of the block, for example, for the location of square blocks (ornamental letters) or of blocks having a width larger than the height (border);
- the *number of connected components* (elements) counted in the block that can prove to be of interest in identifying certain objects;
- the ratio of *block Height/average Height of the connected components, which can be usually* effective to locate the textual blocks comprising only one line of text (titles, legends, etc.).

To finish this part, let us recall that, in addition to these rules allowing evolution of the labels of blocks, it is possible to insert rules of fusion and suppression of blocks in the scenarios. Examples of scenarios built by users,

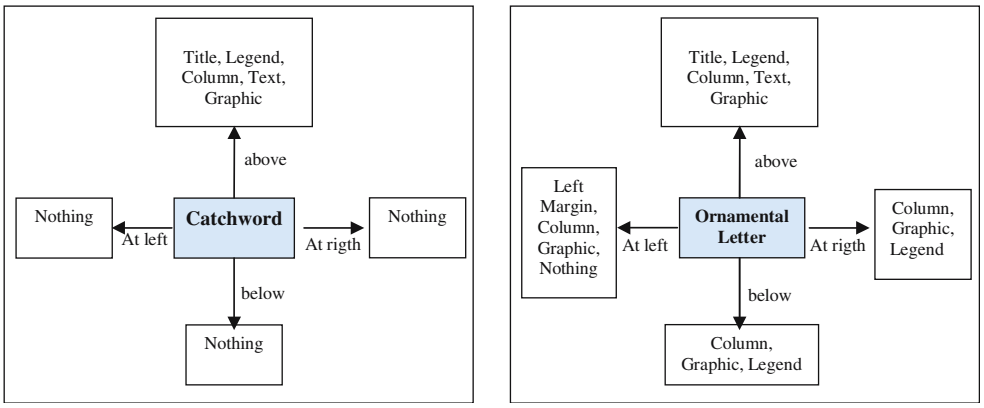

Fig. 12 Examples of neighbourhood rules

maybe not specialists in image processing, can be seen in the following part.

5 Experimentation and results

Our software was made available to the CESR, which currently uses it in an intensive way to process, analyse, index and make available on-line their historical books. Training in the use of the software (segmentation and creation of scenarios) was given to the potential users (historians, librarians, keyboard operators, etc.), so that they could produce scenarios and consequently test our user-driven system on numerous images. This collaboration between the CESR and our laboratory makes it possible to improve and complete the interfaces of the software, taking into account the needs of the final users.

Several experiments were undertaken either by the staff of the CESR or within our laboratory. The first experiment was devoted to an analysis of the robustness of the block segmentation with respect to the skew. This test was achieved using ten images of current documents and ten images of historical books on which we applied rotations from 1° to 20°. A second experiment was carried out in our lab on a complete historical book with 250 pages. The objective of this test was to validate the usability of the proposed interfaces to build different scenarios of analysis for text/graphic separation and classification.

AGORA was then provided to the real end-users at the CESR where numerous experiments have been done. The results of some of them are presented in this article: the first one was aimed at evaluating whether only one scenario could be used to analyse different books or if it was preferable to use one scenario for each book. To conclude this part, two different scenarios are used to illustrate the flexibility of the proposed system. All these experiments have led to the results presented in the next sections.

5.1 Initial segmentation

The result obtained after the first application of the algorithm of fusion on the connected components, according to the foreground and the background maps, which provide a set of blocks labelled *Text*, *Graphic*, *Text_Graphic and Noise*, is called "initial segmentation". This segmentation is not the final one (since it will evolve during the scenarios of incremental analysis). So, we evaluate, in this part, only the resistance of the first step to the skew problems. Next, the results obtained after the complete application of a scenario will prove the global performance of the AGORA system.

We tested the resistance of our segmentation algorithm to the skew of images on old documents and on images of current periodicals. The results obtained using 20 images are summarized in Figs. 13 and 14 and indicate a sufficient tolerance for the old documents and a strong tolerance for more structured documents. If the documents are badly positioned at the time of scanning or if the images are badly illuminated, it is preferable to apply a correction of the geometrical defects before using the segmentation algorithm. When the image scans are correct, the obtained results are very satisfactory without pre-processing.

5.2 User-driven structural analysis

During a first experiment carried out in our laboratory, a scenario allowing the labelling of seven types of historical elements was set up: "Left Margin (MG)", "Right

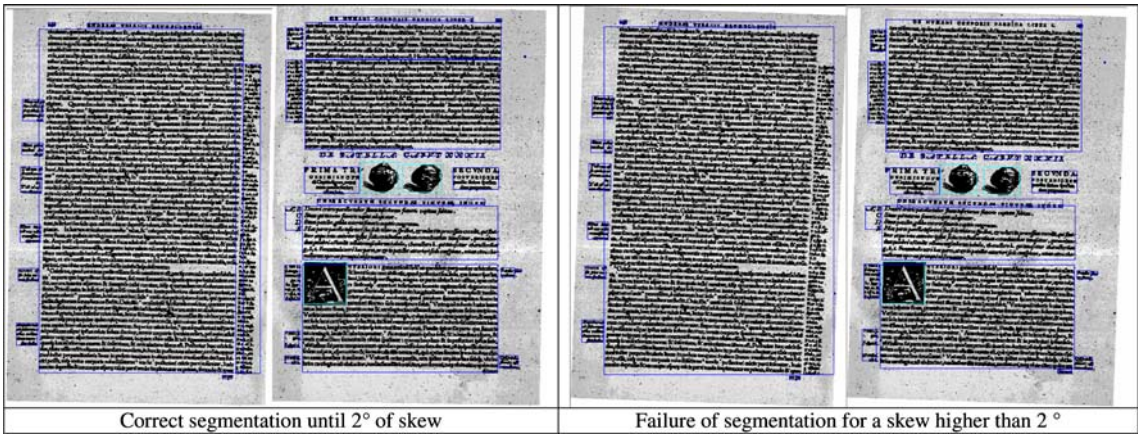

Fig. 13 Sensitivity to rotation on images of old documents

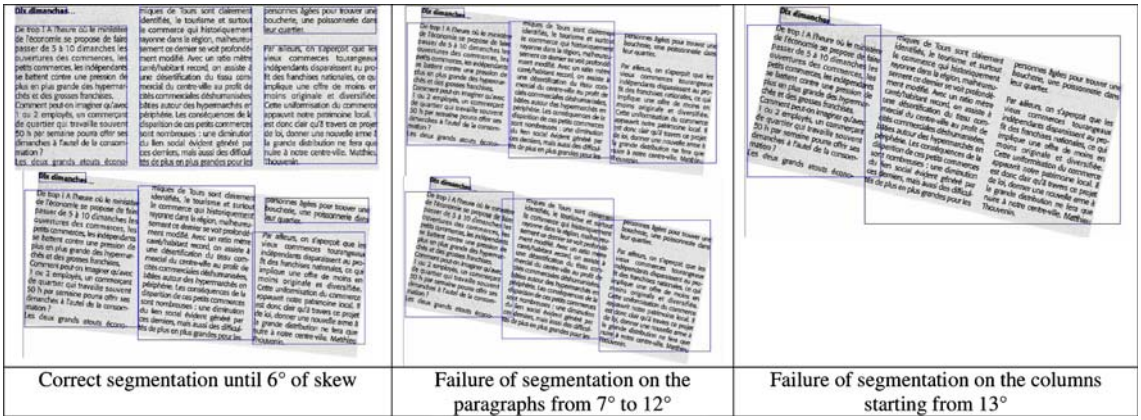

Fig. 14 Sensitivity to skew on images of current documents

Margin (MD)", "Ornamental letter (Ltr)", "Right Page Number (NPD)", "Left Page Number (NPG)", "Principal Title (TP)" and "Title (T)". For this test of block classification, we used an image database of 250 pages coming from a complex historical book provided by the CESR. Previous examples of pages used to illustrate this article were extracted from this book. The obtained results are provided in Fig. 15. The tiresome work of checking the assigned class for each zone was carried out in a visual way on each analysed page. Table 1 shows the results obtained during the batch processing, which lasted 9h and 20min on an Athlon barton 2.5GHz with 768MB of RAM.

These results are dependent on the initial segmentation and on the robustness of the rules constituting the scenario built by the users. The aim of the user who conceived this scenario was not to have any bad classification (minimal risk) for the *Ornamental Letters* and the *Margins*.

The scenario used was thoughtful, but did not implement "redundant rules", which could have involved additional risk on certain criteria. These "redundant rules" aid in the verification of the high-level elements (e.g., ornamental letters) that have already been extracted with the help of the processing rules. For example, once ornamental letters have been located using rules describing their size and positions in the image, the user can specify a redundant rule to verify that no ornamental letter is found to the right of another ornamental letter (impossible in real images). If found, for example, the redundant rule will change the label "ornamental letter" with the label "unknown graphic". Indeed, the labels generated at a given moment can be validated later using other rules. With this scenario, 92% of *the Margins* are extracted correctly without any wrong detection. This scenario could be enhanced in order to locate the missing *Margins*, once again, by adding more rules when the other high-level objects have been localized (evolution of the context to facilitate new extractions). Still with the same scenario, 99% of *the Ornamental letters* were well classified. This is satisfactory since the labelling of *the Ornamental letters* uses the

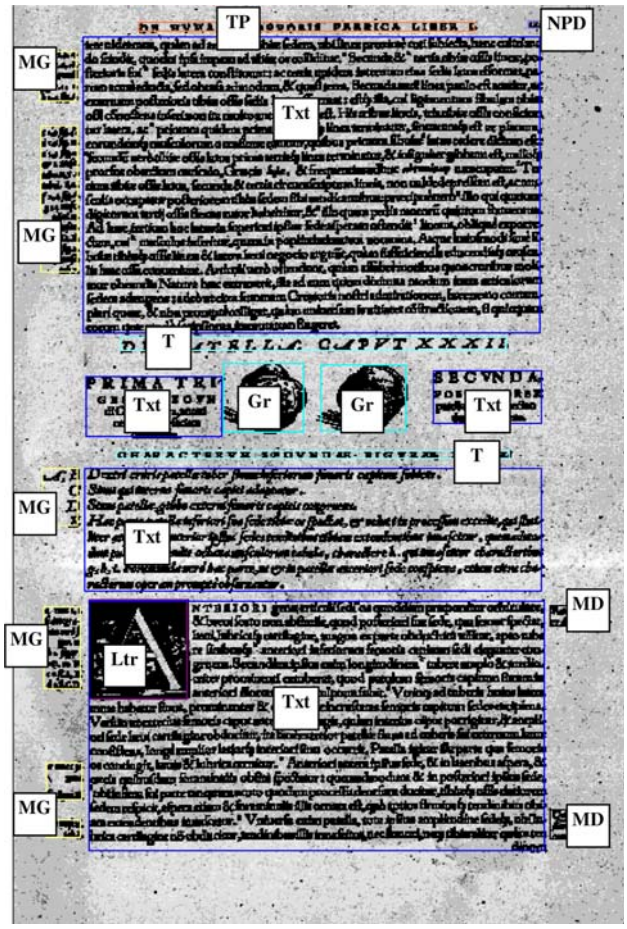

Fig. 15 Example of result obtained with a first scenario 1 [Left Margin (*MG*), Right Margin (*MD*), Ornamental letter (*Ltr*), Right Page Number (*NPD*), Left Page Number (*NPG*), Principal Title (*TP*) and Title (*T*)]

result of the classification of *the Margins* . Table 1 shows that the *Page numbers* were difficult to extract or that the rules used to locate them were not sufficient.

Many other experiments were carried out at the CESR by the staff following the training in the use of the software. In addition to the recognition rates, it is interesting to notice the way in which the produced scenarios are structured and applied.

At the CESR, all the tests were carried out on images resulting from a digital camera that provides 1,200× 2,000 Gy level pixel images. The segmentation parameters were as follows:

- automatic binarization;
- minimum height of a large component: 60;
- minimum width of a large component: 60;
- maximum height of a small component: 5;
- maximum width of a small component: 5;
- threshold of horizontal fusion: 500;
- threshold of vertical fusion: 500.

The first result relates to the initial segmentation, which did not produce any error of labelling. All the *Text* and *Graphic* blocks were correctly detected in all the books selected; however, sometimes an over-segmentation of some blocks resulted from the use of strict fusion thresholds (500). But, one can notice that more fusions can be achieved later during scenarios of classification to correct this over-segmentation.

5.3 First scenario

A first scenario was carried out on 1,452 images from five different books. The objective of this scenario (shown below) was to identify the various types of graphical blocks present in the five different books. As explained in Sect. 3.1, at the end of the first step of the processing, the representations of the image contain a list of blocks labelled as Text, Noise or Graphic. Then, the user has to settle scenarios to make these initial labels evolve using specific rules as shown below:

– *Text on the left of the page with 15% becomes Left Margin (15% of the page width is a threshold defined by the user in this scenario).*
– *Text on the right of the page with 15% becomes Right Margin.*
– *Vertical Fusion of the Left and of the Right Margins with a threshold of 100,000.*
– *Graphic with at the Left : Left Margin OR Nothing AND at the right : Text AND Width/Height Ratio between 0.8 << 1.2 becomes Ornamental Letter.*
– *Graphic with Width/Height Ratio between 3 << 10 becomes Border.*

Table 1 Results obtained with scenario 1

Detected types	Detected	Not detected	False detections	Rate of detection (%)
Margins	207	18	0	92
Numbers of right page	97	32	7	74
Numbers of left page	106	30	10	76
Principal titles	223	27	0	90
Ornamental letters	80	1	0	99

- *Graphic in the centre of the page becomes Portrait.*
- *Portrait with above : Portrait OR below : Portrait OR at the bottom of the page with 20% becomes Graphic.*
- *Graphic with at the Left : Text OR at the right : Text OR above : Text OR below : Text OR Nothing AND in the centre of the page becomes Floret.*
- *Suppression of the Text and the Left and Right Margin blocks.*

Here, we can notice that the first rules often use only topological properties of the blocks to change their labels. This probably means that these kinds of rules are easy to understand and manage (but perhaps not very efficient). When *Margins* are detected, it is possible to run for another time the fusion step (used during the segmentation) to merge only *Margin* blocks. In this scenario, neighbourhood and shapes rules are used to locate *Ornamental Letters* and *Borders* . The images were processed in 11h and 22min and the results obtained are shown in Table 2.

Taking into account the diversity of the selected old books (size, layout, etc.) and the simplicity of this scenario, the results appear quite good. Nevertheless, to obtain better detection rates and to extract more ambiguous objects, it is often preferable to adapt the scenario to each book. Thus, the continuation of the tests was achieved on images coming from a unique book. Some pages of this book are presented in Fig. 16.

5.4 Second scenario

A second test was carried out on 180 pages of the selected book. Because some images were badly illuminated or some text lines had an important curvature, Book Restorer [8]) was used for the pre-processing of the images (lighting correction for the spots and the shade of the binding, geometrical correction of the curvature):

- *Graphic with Width/Height Ratio between 3 << 10 becomes Border.*
- *Graphic with at the left : Left Margin OR Nothing AND at the Right : Text AND Width/Height Ratio between 0.7 << 1.3 becomes Ornamental Letter.*
- *Vertical Fusion of Text with a threshold of 2,000.*
- *Text on the left of the page with 25% becomes Left Margin—Text on the right of the page with 25% becomes Right Margin.*
- *Vertical Fusion of the Margins with a threshold of 100,000.*
- *Horizontal Fusion of the Text with a threshold of 2,000.*
- *Text on top of the page with 10% AND with a number of elements between 0 << 4 becomes Pagination.*
- *Text on bottom of the page with 25% AND with a number of elements between 0 << 6 becomes Signature.*
- *Signature with Text below, on the left or on the right AND on the left of the page with 50% becomes Text.*
- *Vertical Fusion of the Text with a threshold of 3,000.*
- *Horizontal Fusion of the Text with a threshold of 3,000.*
- *Text with height in pixel between 0 << 100 AND on the top of the page with 15% becomes Title.*
- *Horizontal Fusion of the Title with a threshold of 3,000.*
- *Suppression of the Title if the number of elements is lower than 8.*
- *Suppression of the Title if the Width/Height Ratio is lower than 1.*
- *Suppression of the Title if it is on the left of the page with 10%.*
- *Suppression of the Text.*

This scenario is again quite simple. All kinds of rules (topological, shape, neighbourhood) are equally used. Fusion steps are used very often to merge specific block

Table 2 Results obtained on the 1,452 pages of the five books

Type		Detected	False detections	Not detected	Rate of detection (%)
Border		81	4	8	90.6
Floret		36	2	0	100
Ornamental letter		294	56	11	95.6
Portrait		89	31	0	100

The books were mainly textbooks, containing a variety of 85 borders, 34 florets, 249 ornamental letters and 58 portraits

Fig. 16 Example of pages from the selected book after the pre-processing with Book Restorer

like *Margin* and *Text*. The processing lasted 2h and 32min and produced the following results:

- Detection of the *Borders* with 100%.
- Detection of the *Ornamental Letters* with 96%.
- Detected *Pagination* with 79%.
- Identification of 100% of the Titles but with 4% of noise elements (spots, noises, etc.)

5.5 Third scenario

A last test was applied on 1,202 images without pre-processing, with the objective of only labelling the *Text* blocks. The following scenario was used:

- *Suppression of the Graphics*
- *Horizontal Fusion of the Text with a threshold of 3,000*
- *Text in the left of the page with 25% becomes Left Margin - Text on the right of the page with 25% becomes Right Margin*
- *Vertical Fusion of the Margins with a threshold of 3,000*
- *Margins with a Width/Height Ratio between 4<<30 becomes Text*
- *Text on bottom of the page with 25% AND with the number of elements between 0<<8 becomes Signature*
- *Vertical Fusion of the Text with a threshold of 3,000*
- *Text with Width/Height average in 0<<2 AND on top of the page with 15% AND number of elements is lower than 50 becomes Title*
- *Horizontal Fusion of the type Titles with a threshold of 3,000*

After 10h and 52min of processing time, all *the Titles* and *Signatures* were detected correctly. The extraction of *the Margins* functioned, but produced many spurious elements (shades, spots, etc.)

5.6 Discussion

To conclude this experimental stage, we can first notice that the average processing time for one image is 30s. The obtained results are better on pre-processed images. The automatic binarizatrion (by Sauvola algorithm [30]) increases the processing time by 25–35% as well as increasing the precision of the detection. Lastly, it is preferable not to merge the blocks too much at the beginning (initial segmentation) in order to be able, in certain cases, to isolate small elements (e.g., *Signature* or *Pagination*) before carrying out additional fusions later in the scenario of analysis. The presented scenarios show that the users prefer to implement several scenarios (one for the text part and one for the graphical part) instead of using only one global scenario, even if it would probably have provided better results. It probably means that it is not as easy to build thoughtful scenarios as one may initially think.

Seeing these experiments and results, we think that the main advantages coming from the use of AGORA are:

- Robustness to proximity between blocks during the segmentation step. Our hybrid method of segmentation avoids wrong fusion of close text blocks.
- Working even if the blocks are not rectangular and if the layout is variable.
- Robustness to normal skew (for a skew higher than 2°, a skew correction step is needed)

- Powerful separation of Text/Non-text areas even for characters in graphical parts
- Of course, genericity due to the use of a "user-driven method". With the help of scenarios, users can label every type of component (ornamental letters or catchwords in historical books, titles and paragraphs in current books) and not only paragraphs, tables and graphics like with other systems [9,13,22,34].
- The only weak point of AGORA shown by our experiments comes from the binarization step that can be sensitive to bad illumination or to noise coming from the degradations in the initial documents.

6 Conclusion

In the first part of this article, we highlighted the sources of errors in the traditional methods of page decomposition using a characterization of page layout in the historical books. We noticed that each type of method (bottom-up and top-down) required different information, which one should have been aware of to reach a segmentation of quality. For that, our recommendation consists of using first a hybrid algorithm based on the construction of two representations of the image: the map of the shapes that is focused on the connected components and the background map that provides information on white spaces separating the blocks constituting the page. The joint analysis of the contents of these two maps makes it possible to lead to a robust initial segmentation of the image. The results obtained with this method are very interesting; the adjustment of the necessary parameters is straightforward and not sensitive to variations.

Second, the originality of our approach lies in the opportunity, which we offer to the users to be able to build, in an interactive way, scenarios of incremental analysis. We propose to call this new method "user-driven analysis" in opposition to data-driven or model-driven methods. The goal is, on the basis of the initial segmentation, to be able to make the representation of the images evolve in a progressive way to lead to the finest possible characterization of its contents according to the user objectives and to the type of images to be analysed. The CESR has processed several complete books using AGORA prototype and their own scenarios of block classification. Thus, the CESR has increased the number of books offered to the users in its virtual library (see http://www.bvh.univ-tours.fr). Even if the system produced some errors, the processing and time saved as compared to manual processing is considerable (e.g., the manual indexation of the page layout of an historical book of 300 pages lasts $\sim$2days instead of only 2h when using Agora), thus providing to the specialists of historical books, a useful tool, which they had never imagined.

References

1. Akindele, O., Belaid, A.: Page segmentation by segment tracing. In: Proceedings of the 2nd International Conference One Document Analysis and Recognition, Tsukuba pp. 341–344 (1993)
2. Antonacopoulos, A.: Page segmentation using the description of the background. Comput. Vis. Image Underst. Spec Issue Doc. Image Underst. Retrieval **70**(3), 350–369 (1998)
3. Baird, H.S., Jones, S.E., Fortune, S.J.: Image segmentation by shape-directed covers. In: Proceedings of International Conference on Pattern Recognition, Atlantic City, NJ, 820–825 June 1990
4. Baird, H.: Background structure in document images. In: Bunke, H. (ed.) Proceedings of the Advances in Structural and Syntactical Pattern Recognition, Singapore pp. 253–269 (1992)
5. Belaïd, A.: Computer aided design of models of page for their use in recognition of documents, Workshop one Electronic Page Models (LAMPE' 97), Lausanne (1997)
6. Breuel, T.M.: An algorithm for finding maximal whitespace rectangles at arbitrary orientations for document layout analysis. International Conference for Document Analysis and Recognition (ICDAR 2003), Edinburg vol. 1, pp. 66–70 (2003)
7. Chen, S., Haralick, R.M.: Recursive erosion, dilation, opening and closing transforms. IEEE. Trans. Image. Process. **4**(3), 335–345 (1995)
8. Book scanner web site: http://www.i2s-bookscanner.com/fr/default.asp
9. Dori, D., Doermann, D., Shin, C., Haralick, R., Phillips, I., Buchman, M., Ross, D.: The representation of document structure: a generic object-process analysis. In: Bunke, H., Wang, P.S.P. (eds.) Handbook of Character Recognition and Document Image Analysis, chapter 16, pp. 421–456. World Scientific, London (1997)
10. Etemad, K., Doermann, D., Chellappa, R.: Multiscale segmentation of unstructured document pages using soft decision integration. IEEE Trans. Pattern Anal. Mach. Intell. **19**(1), 92–96 (1997)
11. Jain, A.K., Bhattacharjee, S.: Text segmentation using gabor fillters for automatic document processing. Mach. Vis. Appl. **5**(3), 169–184 (1992)
12. Jain, A.K., Zhong, Y.: Page segmentation using texture analysis. Pattern Recognit. **29**(5), 743–770 (1996)
13. Hadjar, K., Hitz, O., Ingold, R.: Newpaper page decomposition using Split and merge approach. In: Proceedings of the 5th International Conference one Document Analysis and Recognition, Seatle pp. 1186–1191 (2001)
14. Hadjar, K., Hitz, O., Robadey, L., Ingold, R.: Configuration recognition model for complex transfers methods engineering: 2(CREM). In: Proceedings of the 5th International Workshop one Document Analysis Systems, Princeton pp. 469–479 (2002)
15. He, J., Downton, A.: User-assisted archive document image analysis for digital library construction. In: Proceedings of the 6th International Conference one Document Analysis and Recognition, Edinburg pp. 498–502 (2003)

16. Journet, N., Eglin, V., Ramel, J.Y., Mullot, R.: Text/graphic labelling of ancient printed documents. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'05). Séoul, Corée, pp. 1010–1014 (2005)
17. Kise, K., Sato, A., Iwata, M.: Segmentation of page images using the area voronoi diagram. Comput. Vis. Image Underst. Spec. Issue Doc. Image Underst. Retrieval **70**(3), 370–382 (1998)
18. Le, D.S., Thoma, G.R., Wechsler, H.: Automated page orientation and skew angle detection for binary document images. Pattern Recognit. **27**(10), 1325–1344 (1994)
19. LeBourgeois, F., Emptoz, H.: document analysis in gray level and typography extraction using character pattern redundancies. In: proceedings of the 5th International Conference on Document Analysis and Recognition, Bengalore, India, pp. 177–180 (1999)
20. Lebourgeois, F., Emptoz, H., Trinh, E.: Compression et accessibilité aux images de documents numérisés/application au projet debora. Doc. Numérique. **7**(3–4), 103–127 (2003)
21. Lee, S.W., Ryu, D.S.: Parameter-free geometric document layout analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1240–1256 (2001)
22. Marinai, S., Gori, M., Soda, G.: Artificial neural networks for document analysis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **27**(1), 23–35 (2005)
23. Min, Y., Cho, S.-B., Lee, Y.: A data reduction method for efficient document skew estimation based on hough transformation. In: Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, pp. 732–736 (1996)
24. Nagy, G., Seth, S.: Hierarchical representation of optically scanned documents. In: Proceedings of the 7th International Conference on Pattern Recognition (ICPR), Montreal pp. 347–349 (1984)
25. Nagy, G., Seth, S., Krishnamoorthy, M., Viswanathan, M.: Syntactic segmentation and labeling of digitized pages from technical journals. IEEE Trans. Pattern Anal. Mach. Intell. **15**(7), 737–747 (1993)
26. O' Gorman, L.: The document spectrum for page analysis layout. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1162–1173 (1993)
27. O' Gorman, L., Kasturi, R.: Document image analysis. IEEE Computer Society Press, Los Alamitos, CA, (1995)
28. Ramel, J.Y., Vincent, N., Emptoz, H.: Extraction contextuelle d'entités graphiques dans les dessins : du plus simple au plus complexe. . .. Colloque International Francophone sur l'Ecrit et le Document, Quebec, Canada, pp. 453–462 (1998)
29. Trinh, E.: De la numérization à la consultation de documents anciens. Thèse de doctorat en Informatique, Insa de Lyon, (2003)
30. Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern Recognit. **33**, 225–236 (2000)
31. Spitz, A.L.: Analysis of compressed document images for dominant skew, multiple skew and logotype detection. Comput. Vis. Image Underst. **70**(3), 321–334 (1998)
32. Wang, D., Srihari, S.N.: Classification of newspaper image blocks using texture analysis. Comput. Vis. Graph. Image Process. **47**, 327–352 (1989)
33. Wang, S.-Y., Yagasaki, T.: Block selection: a method for segmenting page image of various editing styles. In: Proceedings of the 3th International Conference on Document Analysis and Recognition, Montreal, Canada, pp. 128–133 (1995)
34. Wang, Y., Phillips, IT., Haralick, RM.: Document zone content classification and its performance evaluation. Pattern Recognit. **39**, 57–73 (2006)
35. Wong, K.Y., Casey, R.G., Wahl, F.M.: Document Analysis System. IBM J. Res. Dev. **26**(6), 647–656 (1982)