AGORA: the Interactive Document Image Analysis Tool of the BVH Project

J.Y. Ramel*, S. Busson**, M.L. Demonet**
** Lab. d'Informatique, Ecole Polytechnique de l'Université de Tours,*
64, avenue Jean Portalis 37200 Tours- France
Tel : +33.2.47.36.14.26 Fax : +33.2.47.36.14.22
***CESR - UMR 6576 du CNRS –*
59, rue Néricault-Destouches - BP 11328 - 37013 Tours-France
ramel@univ-tours.fr

Abstract

In this paper, we describe how meta-data of indexation can be extracted from historical document images using an interactive process with a software called AGORA. The algorithms involved in AGORA use two maps to segment noisy images: a shape map that focuses on connected components and a background map that provides information on white areas corresponding to block separations in the page. Using a first segmentation result obtained by using these two maps, meta-data can be extracted according to scenarios produced by the users. 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 desired meta-data. Finally, we describe different experimentations that have been done during the BVH project to test the usability and the performances of AGORA software.

1. Introduction

In this paper, we present a work achieved in collaboration with the "Centre d'Etudes Superieures de la Renaissance" of Tours (CESR / http://www.cesr.univ-tours.fr). The CESR is a library, 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 rare books. The CESR wants to create a Humanistic Virtual Library (BVH in French); so, the CESR asks 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 even through the web.

To achieve this goal, we have implemented a new interactive method for the extraction of meta-data in images of historical documents based on the construction of two representations of the contents of the images. A map of the shapes and a map of the background are computed automatically. By exploiting this information, our segmentation 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 and allow to label precisely the extracted blocks.

According to its needs (extraction of ornamental letters, margin notes, titles,…) and using user-friendly interfaces, the user (not expert in image processing) builds scenarios allowing to label, to merge, to remove the blocks contained in the intermediate representation and obtain the desired meta-data. One can thus extract the desired information 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. The BVH project and the needed meta-data

2.1. Overview of BVH indexation scheme

Since its creation, the CESR has its own collection of 3000 rare books (sixteenth and seventeenth century) and now digitizes a set of 2000 rare books coming from

regional collections (Orleans, Blois, Bourges, …). 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. This set is pan European: coming from France, Germany, Italy, Switzerland, and Holland. The languages used, Latin, Greek or French, bring an additional factor of variability to the books. The Renaissance typographies used have significant variability. Some examples of images of these rare books are presented in Figure 1.

The first objective of the BVH project was to index and to diffuse through the web more than 85 historical books (that is to say 24 261 digitalised pages) before the end of 2005. These books should be consultable through the web with the help of a search engine. It should be possible to visualize the pages of a book in different modes: bitmap, colours, pdf, and ascii transcription.

Figure 1: Examples of pages of historical books

Since 1999, the photographic and computer science services of the CESR have built a complete digitalisation, indexation and diffusion process. A web server hosts the BVH Web site with 1,5 To of disk space in order to manage more than 15 000 books in bitmap mode. The web server use Active Server Page (ASP) technology associated to a relational databases and XML technologies. In order to proceed to a very good digitalisation, the CESR possesses a specific scanner dedicated to rare book digitalisation : the Digibook Suprascan 10 000 RGB (from I2S firm) has been installed in 2005.

2.2. Meta-data management

2.2.1. The BVH catalogue. The book descriptions are identical to those typically found in classical library catalogues: (author, book title, place of publication, publication date, publisher, language, secondary author, subject, etc.). These book descriptions (catalogue) are stored in a relational database. It allows classical search among titles, authors, dates and some more precise queries to retrieve a particular book or element of book. Figure 2 shows a part of the stored meta-data for an example of book.

Authors : Colonna, Francesco
Title : Le tableau des riches inventions couvertes du voile des feintes amoureuses qui sont représentées dans le songe de Poliphile, desvoilées des ombres du songe et subtilement exposées par Beroalde.
Other title : Hypnerotomachia Poliphili
Edition : A Paris : Chez Matthieu Guillemot
Date : 1600
Translator : Martin, Jean
Scientific editor Béroalde de Verville, François

Figure 2: Part of the description of a book in the BVH catalogue

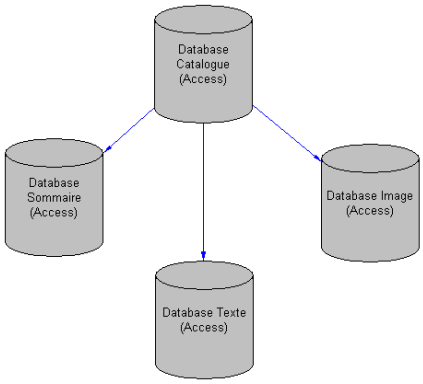
In addition to the catalogue, different databases enable to structure all the information, resources and web tools associated to a given book. Using this information, users can know, for example, if the ASCII transcription of a book is available or not.

2.2.2. Description of book content. Three other databases are used to manage book content description: one for the tables of contents, one for the graphical parts and one for the textual parts of a book. The users can then formulate different queries to access to this meta-data, extracted from a specific book.

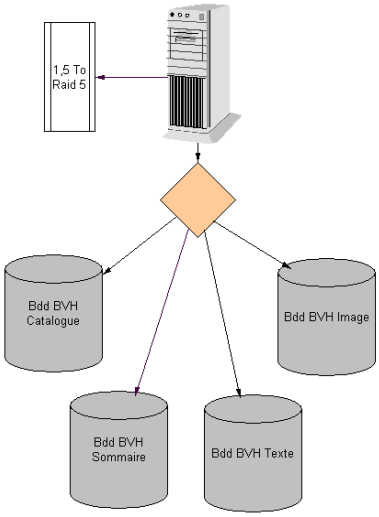
The meta-data relate to the page layout and the logical structure of the book using concepts like: title pages, preliminary pieces, table of contents, final pieces). Due to the high variability of Renaissance books, an automatic recognition of the different structures is a difficult task. Specific scenarios are required to extract and link these structures with the digital image content.

In the BVH project, we have chosen to use both XML technologies and relational databases to manage the meta-data. An XML file is associated to each analyzed pages of a book and describes the physical and logical structures extracted using AGORA. For each extracted block, it is also possible to associate a precise description in the XML file. The meta-data are duplicated in a relational database in order to

accelerate the search of a specific element (for example a special ornamental letter) using a simple sql query. Only ASCII versions (transcriptions) of text blocks are not stored in the database (but only in XML files). Then, ASCII parts are indexed, using a specific search engine during a batch processing. This tool creates an index of all the different words present in a book. Figure 3 shows the 3 necessary databases to store the meta-data associated to each book.



(a) Database structure



(b) Directory structure

Figure 3: Databases and directory structure

For the text, the meta-data concern the position of the text block in the image, the URL of the transcription file, a brief description, and the type of the text block (main body, footnotes, marginal notes, titles, etc.).

In the same way, for the graphics, the information deals with the position of the block in the image, the URL of the image, a brief description and the type of the block (ornamental letter, drop capital …).

By associating the catalogue to these 3 databases with the catalogue, precise queries become possible: for example, searching for all the ornamental letters of a specific printer. One can also notice (figure 3) that an inventory number allows, for each book, to link the information stored in the database to the information stored in the disk directories (XML files and images)

3. Document image analysis algorithms implemented in AGORA

3.1. Pre-processing

Many efforts have been spent on the development of pre-processing algorithms, so that today, commercial solutions exist and are even regarded as satisfactory tools. They relate to the lack of lighting inside the binding, to the curve of the text lines, to page distortion, and to the imperfect elimination of the stains. Many researches were undertaken, in particular at the time of the Debora project, to correct these defects [1] [2]. In the BVH project, the CESR has decided to use Book Restorer commercial tool [3] to deal with low level processing like skew and lighting correction.

3.2. Segmentation

For the segmentation, no software was corresponding to the CESR needs because the physical and logical layout of rare books have very specific characteristics:

• Complex page layout which can present several columns with irregular sizes
• No Editorial Style or identifiable logical structure
• Presence of printed or handwritten notes at the margins
• Presence of location indicators: line numbers, page numbers, catchwords …
• Use of specific and multiple fonts
• Frequent use of ornaments (non textual blocks) such as borders, ornamental letters…
• Variable location of the 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

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

So we have developed a software called AGORA that allow an interactive analysis and segmentation of the layout in historical document images. AGORA segmentation algorithm 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, it is possible to use simultaneously the information provided by these 2 representations (foreground and background) to segment the image. The following sections give more details about these segmentation methods.

3.2.1. Map of the foreground. 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, their size, 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 components representing the letters of a word are extremely close. Unfortunately, in old documents, the page layout and the spacings 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 binarisation of the image (using [4]), 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 constituting the foreground map (figure 4a). 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. It provides the first labelling which will be checked and which will evolve thereafter during the process.

3.2.2. Map of the background. In the binarised image, the background of the page is represented by white pixels. Normally, a large number of white pixels can be aligned vertically (*lgb_v*) or horizontally (*lgb_h*) in the regions between two blocks of a page (text, graphics...). 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 of the image the sum corresponding to the number of horizontally successive white pixels plus the number of vertically successive white pixels to build a distance map (where $Max = Height \times Width$ of the image) .

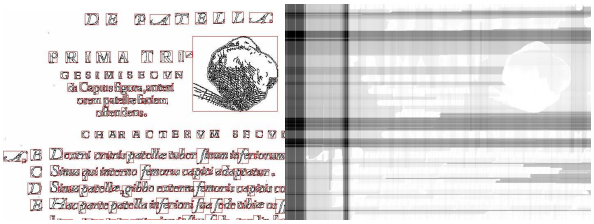
$$Ng(i, j) = Max - [\ lgb\_v(i, j) + lgb\_h(i, j)\ ]$$

If necessary, the *lgb_h* and *lgb_v* values can be weighed respectively by the width and the height of the image in order not to privilege one of the two directions.

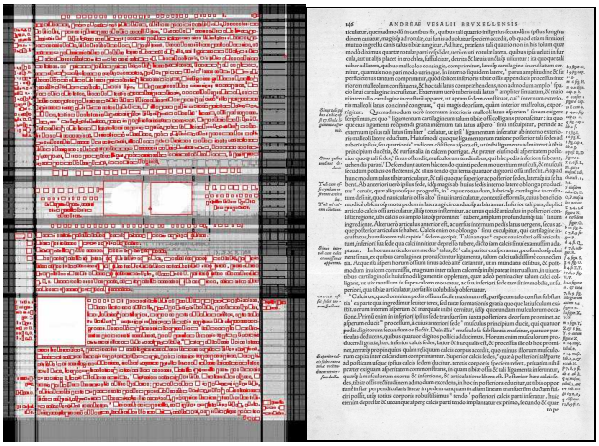
We thus assign a value to each pixel in the background of the page (white parts in the map of the shapes in which the interior of the connected components are blackened). After a normalisation to *255*, we obtain as shown in the figure 4b a map with grey levels (*Ng(i, j)*). As the figure 4b shows, our map of the background translates the separation (more or less underlined) between the blocks of the page.

Baird [5] and Antonacopoulos [6] 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.

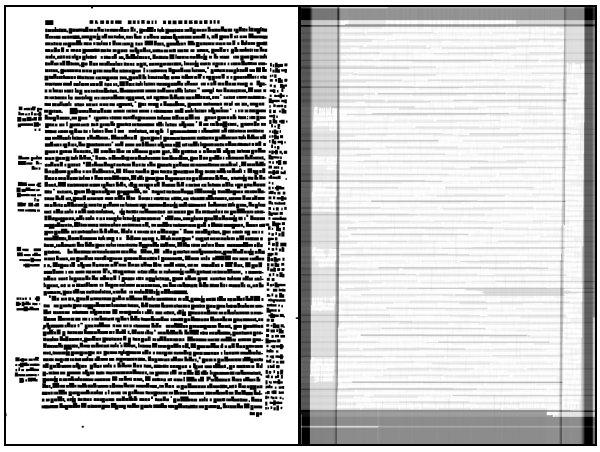
Moreover, in our case, the white blocks constitute only one part of the information which we exploit to carry out the segmentation (and will be combined with the foreground map). In the same way, compared with methods based only on the analysis of the neighbourhood of connected components like [7], our approach uses a more global information because each value in the background map depends on the global layout of the page.



(a) Map of the foreground - map of the background



(b) Global view of the page



(c) 2[d] example of the 2 maps on an other initial image

Figure 4: Maps of the background and maps of the shapes

3.2.3. Fusion of the information provided by the two maps. 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:

$$d(G_1,G_2)\times(256 - \min_{(i,j)\in[G_1,G_2]}[Ng(i,j)]) \leq Fusion\_Threshold$$

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 have been tested (like using the sum of the grey level values instead of the minimum, or the use of the distance between the borders of the 2 areas instead of the centre of gravity, …) and we kept the best one.

Figure 5: Examples of segmentation results (each segmented block has a different gray level value).

The research of the neighbours between text parts is done successively horizontally (horizontal fusion) then vertically (vertical fusion) in a progressive way and stops 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 figure 5. The *Fusion_Threshold* 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), 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 parameters (insertion of the fusion step in the scenarios of analysis created by the users).

4. Interactive extraction of meta-data

4.1. Interactive learning of the model

Except for extremely confined applications [8][9], current block classification or logical layout analysis methods have shown their limitation. Our proposal consists in 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.

As shown in figure 5, the segmentation stage produces an intermediate representation of the image providing for each page a set of blocks labelled 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) allowing a gradual evolution of this intermediate representation according to the desired meta-data and to the characteristics of the images to analyze. 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.

4.2. AGORA interfaces

Once the initial segmentation of the image is achieved, the architecture AGORA makes possible to continue the analysis (model construction) 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 beforehand.

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

• a set of interfaces (instead of a non convivial rule editor) allowing to make evolve the labels given to the various obtained areas (Text, 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 characterisable blocks using simple rules. Then, he can use a more complete context to extract the less easily identifiable parts of the document [10].

• the possibility to apply the algorithms for horizontal or vertical fusion with various thresholds and only on a specific type of blocks defined by the user (for example to merge only the blocks labelled Title which were split at the beginning of the process in several blocks

• the possibility of deletion of the blocks of a particular type (for example the Text labelled blocks if the user is interested only in the ornamental letters).

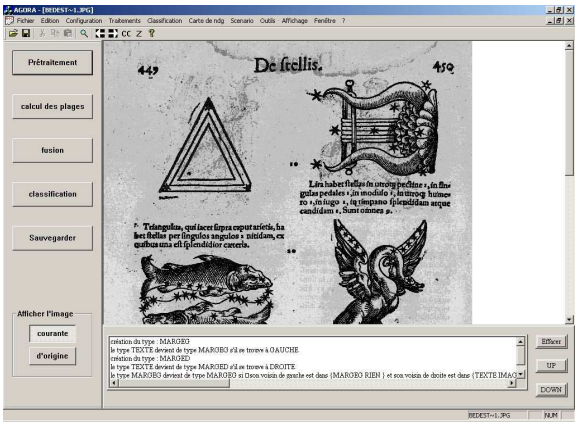


Figure 6: View of the Agora software. The active scenario is displayed at the bottom of the window

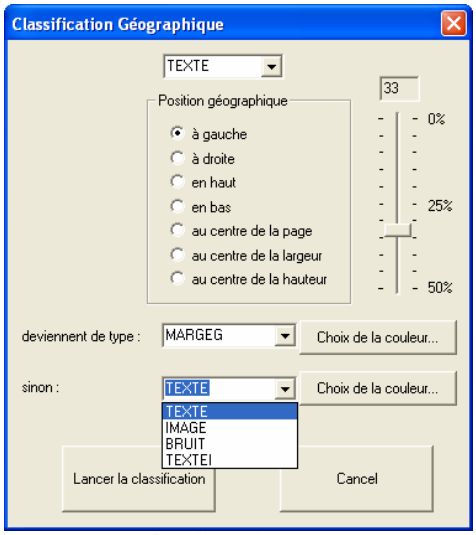
To build a scenario, the user realises the successive actions which have to be recorded on a typical image. These actions are applied to the image and the results are displayed in real time. The user can thus validate their interest. The actions (rules) are translated in a literal list permanently displayed (see figure 6). The user can manage this list to modify the scenario by reordering or removing processing rules. Once the scenario is considered as correct, it can be saved in a file for storage or for application on a more consequent set of images (a complete book) during a batch processing.

Currently, the usable interfaces (rules) to change the labels of the blocks contained in the intermediate representation concern (figure 7):

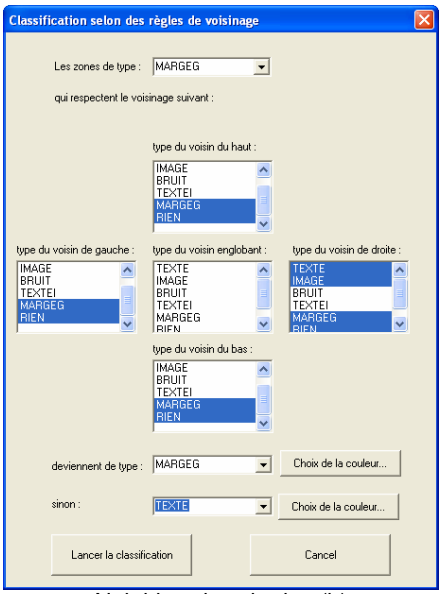
• the geographical position of the blocks

• the neighbourhood relations between identified blocks
• the shape and the content of the blocks

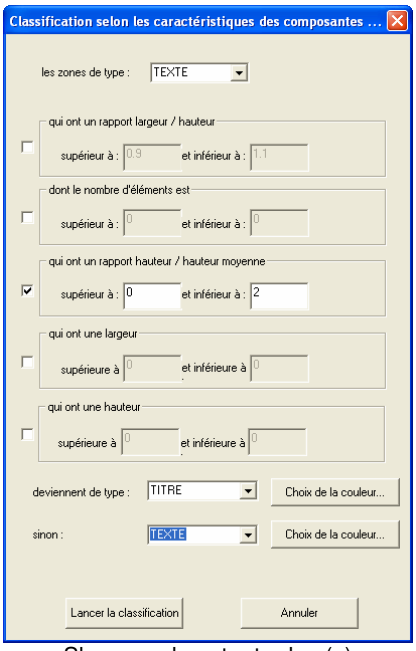
It is possible to identify a number of invariants on the geographical positioning of the objects present in a historical book making it possible to associate a label to them.



Position rules (a)



Neighbourhood rules (b)



Shape and content rules (c)

Figure 7: Interfaces for the installation of the rules rules

The goal is not to use extremely strict rules because page layout can be variable (it is abusive to say that the centre of gravity of a note in left margin is between the pixel of X-coordinate 205 and 213). However, one can reasonably say that the position of the centre of gravity of a left margin text is doubtless located in the first third of the width of the page. The suggested interface (see figure 7a) makes it possible to take in account the geographical position (left, right, top, bottom, centre) of the centre of gravity of the blocks to make the label evolve. Of course, this type of rule will only constitute a first index (provisional label) to correctly isolate an object of a given type.

It is also possible to insert a rule which concern the neighbourhood relations between blocks in a scenario (figure 7b). As shown in this figure, the rule can concern the left, right, above, or behind neighbour of the block. For example, it is possible to use such a rule to change the label of all the Text blocks that have a graphical part above them to make them become a Legend labelled block.

Another interface makes it possible to use shape and content criteria to make the labelling evolve. The non exhaustive list of these criteria is visible figure 7c. The Height / Width ratio allows, for example, location of the square blocks (ornamental letters) or of the blocks much width than high (border). The number of elements corresponds to the connected components counted in the block and can prove to be interesting to

identify certain objects. The zone Height / average Height (of the connected components) ratio is very effective to locate the textual blocks comprising only one line of text (titles, legends...).

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 can be seen in the following part.

5. AGORA usability

Our software was made available to the CESR which currently uses it in an intensive way to process, analyse, index and publish their books on the web. Training in software using (segmentation and creation of scenarios) was given to the potential users (in humanities) 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 possible to improve and complete the interfaces of the software taking into account the needs of the final users.

Many experiments have been carried out in 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.

In the CESR, all the tests have been carried out on images resulting from a numerical capture that provides 1200 X 2000 grey level pixels images. The used segmentation parameters were as follows:

- automatic binarisation
- 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

5.1. Extraction of the table of contents

The first use of AGORA has the objective to analyse text block in order to extract titles, signatures and page numbers in order to automatically build the table of contents of the old books. Figure 8 shows an example of title block extracted using this scenario. The used scenario was very simple:

- Vertical Fusion of the text with 2000
- Horizontal Fusion of the Text with 2000
- Pagination Type: Text in top with 10% + a number of elements 0<<4
- Signature Type: Text in bottom with 25% + a number of elements 0<<6
- Signature has Text below, on the left or on the right + if it is on the left to 50%, it becomes Text type
- Vertical Fusion of the Text with 3000
- Horizontal Fusion of the Text with 3000
- Titles Type: height in pixel 0<<100 + Text in top with 15% to include the running head and the Pagination
- Horizontal Fusion of the Title
- Suppression of the type Title if the number of elements is lower than 8
- Suppression of the type Title if the Width/Height Ratio is lower than 1
- Suppression of the type Title if it is on the left to 10%
- Suppression of the Text type

All kinds of rules (geographical, shape, neighbourhood) are equally used. Fusion step is used very often to merge specific block like Margin and Text. This scenario 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...)

Figure 8: Example of extracted Title.

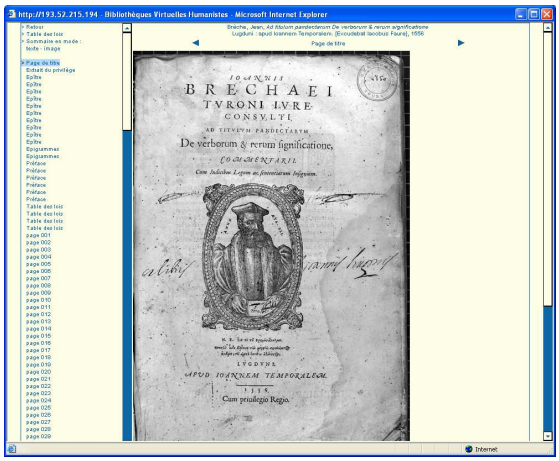


Figure 9: Example of the displayed Table of content

The extracted tables of contents are stored in a specific table of the database. The URL of the title images, the text transcription and the associated page number are stored. The web site can then propose, in a dynamic way, one frame with the table of content of the book and an other frame to display the selected page (figure 9).

5.2. Extraction of graphical regions

The second important use of AGORA is for the automatic extraction and labelling of graphical parts. A more complex scenario is needed to correctly extract and label the different graphical parts. The objective of the scenario presented below is to identify the various types of the graphical blocks present in five different books:

- Left Margin Type: Text on the left with 15%
- Right Margin Type: Text on the right with 15%
- Vertical Fusion of the Left and Right Margins with 100000
- Ornamental Letter Type: Images close to Left Margin or Nothing at the Left, close to Text at the Right + Width/Height Ratio 0,8<<1,2
- Border Type: Image ratio Width/Height Ratio 3<<10
- Portrait Type: Image in the centre of the page + if Portrait in top or bottom with 20% becomes again of Image type
- Floret Type: Image close to Text or Nothing above, close to Nothing behind + centred on the width
- Suppression of the Text and the Left and Right Margins

Here, we can notice that the first rules often use only geographical properties of the blocks to change their labels. This probably means that these kinds of rules are easy to understand and to 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 1452 images were processed in 11h22 and the results obtained are provided table 1.

Table 1: Graphical parts labelling results

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

Taking into account the diversity of the selected old books (size, layout...) 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.

In all cases, AGORA send back to the databases all the coordinates of each extracted graphical blocks so that, after indexation, users can zoom on a specific image of a book (figure 10). As said before, the set of meta-data extracted by AGORA are stored in the Image table of the database. This table stores the URL of the image file, the label of the image block (ornemental letter, portrait, …) the position of a block in the page, eventually some keywords added manually.

Figure 11 shows how a user can interact with the BVH Web site to find a specific indexed graphical element.

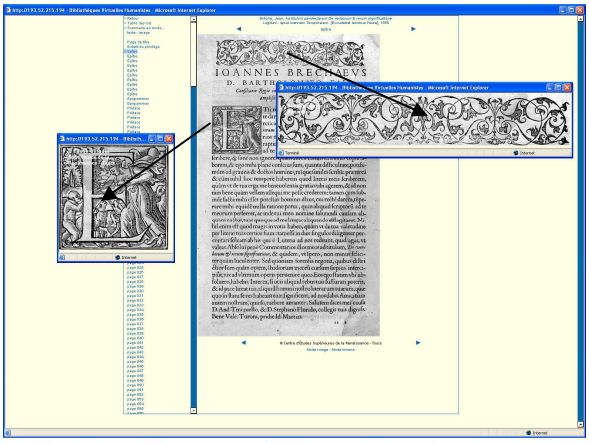


Figure 10: Example of zoom on graphical regions

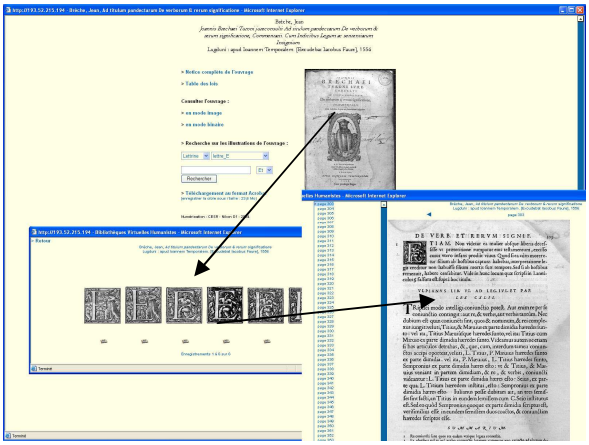


Figure 11: Example of query on graphical elements

5.3. Transcription of text blocks

To conclude, we want to talk about the analysis of text blocks using AGORA. As for graphical parts, it is possible to label each text blocks in an historical book (for example: margins notes, legend, main text, …) by using a specific scenario. An example of scenario is provided just below :

- Suppression of the Image
- Horizontal Fusion of the Text with 3000
- Left or right Margin Type: Text on the left or on the right to 25%
- Vertical Fusion of the Margins
- If the Margins with a Width/Height Ratio 4<<30 becomes Text

- Signature Type: Text in bottom with 25% of which the number of elements 0<<8
- Vertical Fusion of the Text with 3000
- Title Type: Width/Height average in 0<<2 + text in top with 15% to include the running head and the pagination + Text in which the number of elements is lower than 50
- Horizontal Fusion of the type Titles

During our experimentation of this scenario on a specific book, all the Titles and Signatures were detected correctly. The extraction of the Margins notes worked but produced many spurious elements (shades, spots...)

Once the text blocks have been labelled, it is possible to use a specific OCR engine to make their transcription. A specific engine or a specific learning set can be used to recognize the characters in the different types of blocks extracted using AGORA. It is interesting because, for example, characters in margin are very different from the one in the titles. We have also developed a transcription system to deal with old characters recognition but we will not speak about it in this paper (figure 13). Nevertheless, let us say that AGORA provides the low-level-data to the OCR software using XML files identical as the one shown in figure 12.

```
- <bloc name="C:\PFE\ImagesCESR\Lot\vesale_0150.jpg"
    coord="798,107,2023,155">
  - <ligne>
    - <mot>
        <cc value="2">798,123,831,155</cc>
        <cc value="3">849,123,881,154</cc>
        <cc value="4">899,123,934,155</cc>
        <cc value="5">951,121,981,153</cc>
        <cc value="6">997,121,1024,153</cc>
        <cc value="7">1038,119,1069,151</cc>
      </mot>
    - <mot>
        <cc value="8">1161,119,1195,152</cc>
        <cc value="9">1212,117,1240,149</cc>
        <cc value="10">1257,117,1277,148</cc>
        <cc value="7">1296,116,1327,147</cc>
        <cc value="11">1344,117,1376,148</cc>
        <cc value="12">1393,116,1406,147</cc>
        <cc value="13">1422,114,1438,147</cc>
      </mot>
    - <mot>
        <cc value="14">1508,115,1528,146</cc>
        <cc value="5">1547,113,1576,146</cc>
        <cc value="9">1594,114,1629,148</cc>
        <cc value="15">1646,113,1683,146</cc>
        <cc value="8">1699,112,1727,145</cc>
        <cc value="16">1743,111,1777,143</cc>
        <cc value="12">1790,112,1822,143</cc>
        <cc value="6">1839,111,1866,143</cc>
        <cc value="3">1884,109,1916,141</cc>
        <cc value="14">1970,107,1986,140</cc>
      </mot>
  </ligne>
</bloc>
```

Figure 13 : XML files corresponding to text block

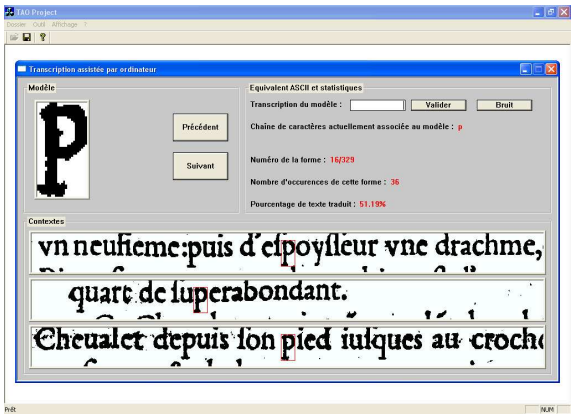


Figure 13: Our transcription software

5.4. Discussion

The results provided by AGORA are dependant on the initial segmentation and on the robustness of the rules constituting the scenario built by the users. In most of cases, the initial segmentation 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 of the use of the strict fusion thresholds. But, one can notice that more fusions can be achieved later during scenarios of classification to correct this over-segmentation.

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 we may think.

Seeing these experiments and results, we think 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 blocks are not rectangular and if the layout is variable

• Robustness to normal skew

• 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 components (ornamental letters or catchwords in historical books, titles and paragraphs in current books) and not only paragraphs, tables and graphics like with other classical systems [11] [12] [13] [14].

The only weak point of AGORA shown by our experiments comes from the binarisation step that can

be sensitive to bad illumination or to noise coming from the degradations in the initial documents.

6. Conclusion

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 to 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 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,...). These limitations appear too restrictive with respect to historical documents and new approaches need to be developed.

So, our recommendation consists of using first an hybrid algorithm based on the construction of two representations of the image: the map of the shapes which is focused on the connected components and the background map which provides information on white spaces separating the blocks constituting the page. The joint analysis of the contents of these two maps makes 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 easy 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 analyzed. 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, this providing to the specialists of historical books, a useful tool which they had never imagine.

7. References

[1] E. Trinh, De la numérisation à la consultation de documents anciens. Thèse de doctorat en Informatique. Insa de Lyon. 2003

[2] F. Lebourgeois, H. Emptoz, E. Trinh, Compression et accessibilité aux images de documents numérisés / Application au projet Debora. Document Numérique. Vol 7(3-4). p103-127. 2003

[3] http://www.i2s-bookscanner.com/fr/

[4] Sauvola, J., Pietikainen, M.: Adaptive Document Image Binarisation. Pattern Recognition Vol. 33, p225-236. 2000

[5] H Baird. Background structure in document images. In Advances in Structural and Syntactical Pattern Recognition, ED. H. Bunke. p253-269. 1992.

[6] A. Antonacopoulos, Page Segmentation Using the Description of the Background. Computer Vision and Image Understanding, Special Issue on Document Image Understanding and Retrieval, Vol. 70, No. 3, p350-369, 1998.

[7] K. Kise, A. Sato, M. Iwata. Segmentation of page images using the area Voronoi diagram Computer Vision and Image Understanding Special issue on document image understanding and retrieval. Vol. 70(3). p370–382, 1998.

[8] J. He, A. Downton: User-Assisted Archive Document Image Analysis for Digital Library Construction. Proceedings of the 6th International Conference one Document Analysis and Recognition.. p498-502, 2003:

[9} Simone Marinai, Marco Gori, Giovanni Soda: Artificial Neural Networks for Document Analysis and Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol.27(1): p23-35 2005.

[10] JY Ramel, N. Vincent, H. Emptoz 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). p453-462. 1998.

[11] K Hadjar, O Hitz, R. Ingold. Newpaper page decomposition using Split and merge approach. Proceedings of the 5th International Conference one Document Analysis and Recognition. p1186-1191, 2001.

[12] K Hadjar, O Hitz, L Robadey, R. Ingold. Configuration REcognition Model for Complex Transfers Methods Engineering: 2(CREM). Proceedings of the 5th International Workshop one Document Analysis Systems. p469-479, 2002

[13] S.-Y. Wang and T. Yagasaki. Block Selection: A Method for Segmenting Page Image of Various Editing Styles. In Proc. of the 3th International Conference on Document Analysis and Recognition, Montreal, Canada, p128-133, 1995.

[14] L. O'Gorman and R. Kasturi. Document Image Analysis. IEEE Computer Society Press, Los Alamitos, CA, 1995.