

# Ancient Printed Documents indexation: a new approach

Journet Nicholas<sup>1</sup>, Mullot Rémy<sup>1</sup>, Ramel Jean-Yves<sup>2</sup>, Eglin Veronique<sup>3</sup>

<sup>1</sup> L3I, 17042 La Rochelle Cedex 1 - FRANCE  
{njournal, rmullot}@univ-lr.fr

<sup>2</sup> LI, 64 Avenue Jean Portalis 37200 TOURS - FRANCE  
Jean-Yves.Ramel@univ-tours.fr

<sup>3</sup> LIRIS INSA de Lyon, Villeurbanne cedex - FRANCE  
eglin@rfrv.insa-lyon.fr

**Abstract.** Based on the study of the specificity of historical printed books and on the main error sources of classical methods of page layout analysis, this paper presents a new way to achieve an indexation of ancient printed documents. We have developed an approach based on the extraction and the quantification of the various orientations that are present in printed document images. The documents are initially splitted into homogenous areas in which we analyze significant orientations with a directional rose. Each kind of information (textual or graphical) is typically identified and labelled according to its orientation distribution. This choice of characterization allows us to separate textual regions from graphical ones by minimizing the a priori knowledge. The evaluation of our proposition lies on a document image retrieval using layout extraction criteria and can also be used to precisely localize graphical parts in various types of documents. The system has been tested with success over several ancient printed books of the Renaissance.

## 1 Introduction

In this paper, we present a work corresponding to a collaboration between three research laboratories dealing with document image analysis and the “Centre d’Etude Supérieur de la Renaissance” of Tours. The CESR is a training and research centre which receives students and researchers wanting to work on all the different domains of the Renaissance using a rich library of historical books. The CESR wants to create a Humanistic Virtual Library but, until now, only bitmap versions of historical books that have been scanned or photographed are accessible. So, since a few years, the center is trying to build a more powerful system to index and diffuse their collections through the web.

In this context, we present first a study of historical book specificities in order to infer some invariant characteristics used during the automatic analysis of their layout. Then, we describe the classical extraction methods that are usually applied on such documents by focussing on their drawbacks.

In a second part, we present a new approach of ancient printed documents layout characterization that is robust to noise (in the background but also in the printed text or the graphical areas) and completely independent of the employed typography, the characters size, the presence of graphical parts and of any particular editorial chief. It means that interest regions can be localized everywhere in the page with a total typography free consideration. Finally, we show how it is possible to separate text from graphics without any a priori knowledge on the nature of the books and of the typographical tools. We only need some strong hypotheses of text alignment and regularity. In the last part of the paper, we present a synthetic document image retrieval that illustrates the relevance of our layout analysis method.

## **2 Context of the study**

### **2.1 Characteristics of the study**

The historians have accumulated many European ancient printed document collections. Those books come from different countries (France, Germany, Italy, Holland,...) and different centuries (from the middle of the 15th to the end of the 17th). Books from the 16<sup>th</sup> century are characterized by a wide variety of forms and contents, even if it was the beginning of the rationalization and codification of texts, of typography that provided an improved reading comfort.

At the beginning of printing, the fonts and the layouts of the pages were very close to the handwritten books [1]. Later, these printed documents were handcrafted and the technical constraints of the past reduced the regularity of book production (variations in spacing and margins, random alignment, etc.). These documents contain many defects due to the manufacturing process and the conditions in which these books were conserved. The variability of page layouts is due to either technical inaccuracies or liberties taken by the printer. There are no exact rules, but most of the time a body text part covers the majority of the page area with generally some notes in the margins. The page can also contain graphical parts of various sizes and some ornament 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. The styles used can alternate, with normal style, justified or aligned on the left. Another characteristic of old printed books comes from weak separations between blocks of text (notes in the margins and body text for example). Lastly, we can notice that on some documents layout rules are always not respected. For example, an illustration can overflow into the margins. To adjust all the lines on a page, the printer could vary the line spacing and the margins. This lack of regularity makes automatic layout analysis difficult. On the other hand, such century books are printed using only strokes and line-art graphics, which can be more easily segmented. Moreover, from one country (or century) to another the printed techniques and layout edition rules were quite different. That is why our corpus presents great variability in pages layout that does not exist in contemporary books.

## 2.2 Evaluation of the existing methods on ancient books

The quality of new segmentation methods has significantly increases those last ten years. Consequently, the software that are devoted to contemporain documents recognition are often unsuitable to the processing of books of the Renaissance period even if we corrected the skew and the curvature due to the book binding using Book-restorer software for example.

The methods of structure extraction employed by software dealing with contemporary books can be classified in 3 main categories: bottom-up, top-down and mixed methods [2].

Typical segmentation and information retrieval approaches lie on usual features analysis that are generally based on a connected components analysis or morphological and directional filtering as it is presented in other existing works [3, 4]. These traditional bottom-up methods are not adapted to the historical books with all their specificities presented in section 2.1 (especially non constant spaces between shapes). Thus, the great difficulty of their use lies on the introduction of many parameters that often lead to prohibitive processing times.

In the same way, the top-down and mixed methods [5,6,7] (horizontal and vertical projection, multi-resolution analysis) endures the same weakness: they need to much a priori knowledge about the documents to be effective (number of columns, width of margins, kind and place of ornamental letters...). So, whatever the method is, our tests have shown that these methods are not robust to the variability of our corpus.

That's why, we have decided to use texture-based approaches. Gabor filters, autocorrelation function, fractal or wavelet analysis are interesting methods because they allow a text/graphic separation without using any kind of structural information. In that context, we have finally chosen to use the autocorrelation function that allows us to characterize large area of text and graphics despite digitalization defaults, text skewness and other kinds of ancient document noises (ink dots, background spots...).

## 2.3 Overview of our approach

The aim of our system is to realize a robust indexation system that is adapted to large heterogeneous collections of ancient books. It must be adapted to an end-user, non-specialist in document or image analysis. Thus, no threshold, document model and explicit structure have to be taken into account in the user interface. Due to this high level of constraints, we reduce the indexation process to help the user to build its own indexation. This one is finally based on both his own expertise and on the results of image analysis process. This process has been divided into three parts:

- Unsupervised automatic extraction of homogenous areas in the images using only orientation features (by labelling pixels).
- Classification of the different page layouts for one book using the position and the labels of the different extracted areas. This classification should also underscore different book layout styles.
- Interactive process allowing precise segmentation and semantic labelling of historical book elements.

This paper describes mainly the first stage of the automatic process. This first stage is a labelling process that organizes pages into different large classes of areas: background, text and non-text. This stage does not need any a priori information (except that text is horizontal).

After being resized, a simple separation of foreground and background based on homogeneous high grey level pixels estimation is applied. It can not be considered as a document binarization but it must be seen as a page separation into two main classes. Then, the image is crossed by a window that extracts specific orientations information and marks each foreground pixel in two main classes: the text and the graphics classes.

### 3. Page layout characterization

Our approach lies on the estimation of relevant directions of significant parts of the image: the initial image is then splitted into homogenous regions in which we analyze significant orientations with a directional rose that has been initially proposed by Bres[8].

#### 3.1 Directional rose computation

The directional rose computation lies on the use of the autocorrelation function, which correlates the image with itself, highlights periodicities and orientations of texture. This function has been widely used in a context of texture characterization, [9]. Its definition for a bi-dimensional signal is the following.

$$C_{xx}(k, l) = \sum_{k'=-\infty}^{+\infty} \sum_{l'=-\infty}^{+\infty} x(k', l') \cdot x(k'+k, l'+l) . \quad (1)$$

The autocorrelation function  $C_{xx}(i, j)$ , applied to an image I, combines this image I with itself after a translation of vector (i,j). The different translations that are considered by the function give information on the different privileged directions in the image. With this principle, it is possible to detect orientations of the texture in the different parts of the image. For example, the translation of a line in the same direction leads to a great correspondence and is expressed by a great value of autocorrelation in the line direction. Inversely, in the orthogonal direction of this line the resulting value will be low. The autocorrelation underlines the objects' overlapping that is obtained by translation. This principle can be generalized to a set of objects having a common direction: in our work, we use it to show that text lines can be characterized by a horizontal privileged direction and can also be considered with a possible skew variation. The determination of relevant orientation is based on the application of successive gliding masks in the original image in which we compute an autocorrelation function that reveals periodicities and orientations in image. The principle of the autocorrelation computation lies on the frequencies decomposition of the analyzed

image with a Fourier Transform (FFT) which avoids the highly complex development of the correlation. The Plancherel Theorem [10] is at the basis of this simplification.

The autocorrelation result can be analyzed by the construction of a directional rose that reveals significant directions in the analyzed block image. The rose computation is based on the mean value that is computed from the autocorrelation result. Let's consider  $I'$  the block of the image and  $\{(x,y)\}$  the set of coordinates in this image. We also consider  $\theta$  as privileged direction in the area. The mean value  $E_\theta$  is then defined by the following formula:

$$E_\theta = \{I'(x, y).I'(x + a, y + b)\} \text{ with } \text{Arctg}(b/a) = \theta . \quad (2)$$

$$R(\theta_i) = \sum_{D_i} C_{xx}(a,b) . \quad (3)$$

The directional rose represents the sum  $R(\theta_i)$  of different values  $C_{xx}(i, j)$  (defined in formula 1) in a given  $\theta_i$  direction. So, the directional rose corresponds to the polar diagram where each direction  $\theta_i$  that is supported by the  $D_i$  line, is represented by the sum  $R(\theta_i)$ . For all points (a,b) of the  $D_i$  line we have the following relation: From this set of values, we only keep relative variations of all contributions of each direction. So, the relative sum  $R'(\theta_i)$  is the following :

$$R'(\theta_i) = \frac{R(\theta_i) - R_{\min}}{R_{\max} - R_{\min}} . \quad (4)$$

Figure 1 shows some examples of directional roses corresponding to 3 different initial images.

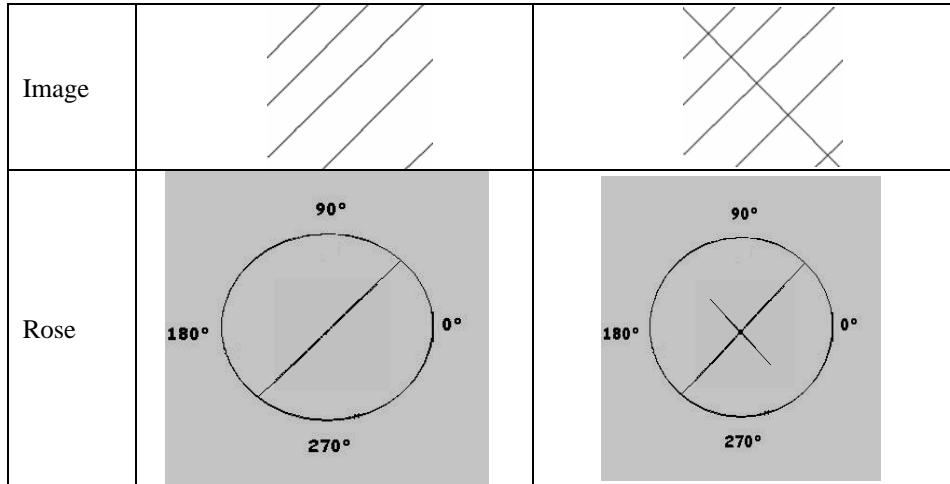


Fig 1.: Directional roses

### 3.2 Directional rose analysis

Due to its mathematical definition, the rose has interesting properties. In a situation of homogenous directions' repartition of the pixels in an image, the rose is a perfect bowl (fig 2 part a). If we add a horizontal line (fig 2 part b), the fact that we only keep relative variations (formula 4) implies that the rose has two invariant characteristics: one "peak" for the horizontal orientations ( $0^\circ$  and  $180^\circ$ ) and the same intensity for all the other directions (it gives an impression of "bowl").

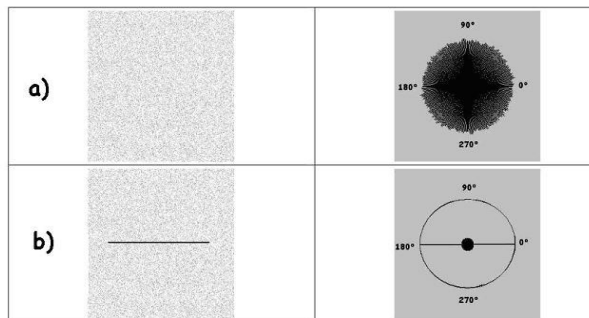


Fig. 2: Text characterisation

A homogenous text area has the same property: all directions are perceptible with a great regularity (that is systematically quantified by a variance measure) and the most important are the horizontal direction that is characterized by a significant characteristic peak with an invariable maximal amplitude (figure 3a). If the text is not homogenous: different sizes of text, non constant spaces....) in the studied area, the formula 4 implies that the size of the "bowl" can decrease or even disappear (figure 3b).

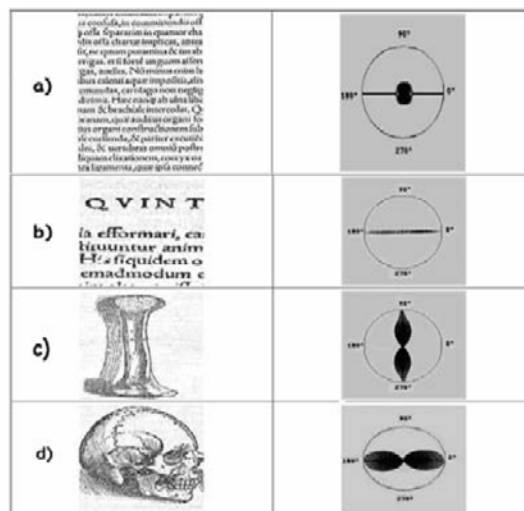


Fig. 3. Graphics characterization

The detection of graphical parts lies on this two following observations: if there are no horizontal significant directions, then we assume that the region contained in the analysed window is a graphical part with a great confidence rate (figure 3c). In the same way, if the main direction is horizontal without a characteristic "peak", then we can conclude that the region is a graphical part (figure 3d). All the other kinds of roses do not allow pixels labelling: it is often due to border zone analysis or to an area where there is not text enough to produce a significant "peak" or a "bowl".

### 3.3 Evaluation of the robustness of the rose

In our method, we always resize the images to be processed on a constant sized image. Our tests confirm the mathematical theory: the resolution of the image does not change the shape of the rose.

We have also brought some tests about the robustness of our rose analysis algorithm on degraded parts of some images as shown in figure 4.

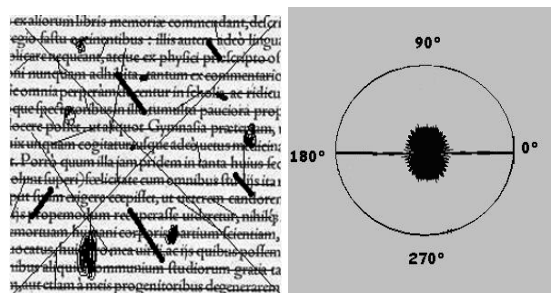


Fig. 4. Rose of a degraded text region

Our tests have shown that the rose is not sensitive to noise or image distortion but a minimum of 4 text lines is necessary to detect a text area. That is why in our method, the size of the analysis area is constant (128X128 pixels) while image sizes can be variable. Actually, the users have the choice to resize the image manually in order to have the minimum lines required in a 128X128 pixel analysis area or to use a constant size for the image (800x900 pixels). Then, the analysis area (a window) is moved iteratively all over the pixels of the image so as to give a label to each of them (background, text, non-text, unknown).

## 4. Targetted applications and experimental results

### 4.1 Labelling results

In most images, all the different parts are well detected using the directional rose analysis. The main problem comes from non constant transitions that exist between blocks and from very huge characters sizes in some titles. Except these two special

cases, our method gives quite good results and allows us to realize our purpose: a relevant separation of the pages into pre-labelled regions that also highlights the document visual content with a robust extractor (with same results whatever are the typography, the fonts, the background noise and the resolution).

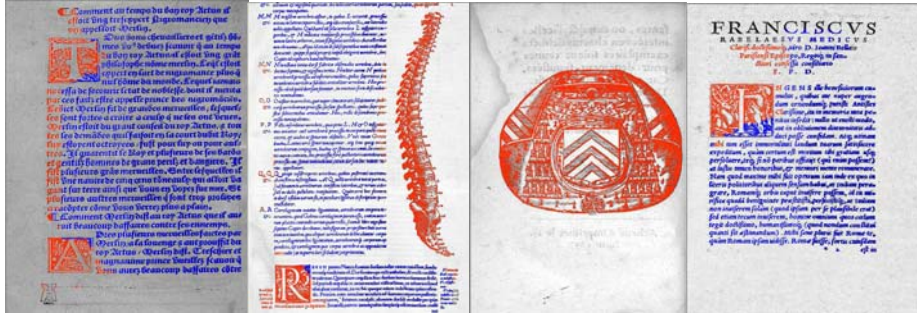


Fig. 5. Examples of results

In the following results, the colors of the pixels are organized as follow: for text and graphics parts we choose respectively blue and red colors. The figure 5 shows a short panel of typical results.

A more consequent test has been realized on over 100 pages from five different books of the Renaissance. The results provided by our algorithm have been compared with a perfect labelling according to a human (our ground truth). So, for text parts, the number of letters which were correctly, wrongly or miss detected has been manually enumerated. For graphical parts, the number of graphical pixels which are correctly, wrongly or not detected has also been manually enumerated. Detection results are summarized in table 1.

Table 1. Detection results (% of correct/wrong/missed detection parts)

Text parts	Graphical parts
95/3/2	88/10/2

#### 4.2. Page comparison

We think that the complex problem of ancient document images indexation can be simplified by a global analysis of all pages of a book before taking any kind of physical or logical conclusion. As Shin and Maderlechner in [11,12] we want to compute a page classification.

With this page layout classification, we hope to extract editorial rules, special text (with specific typography) or graphical features to adapt the following processing.

By using only the percentages of text and non text pixels obtained with our algorithm, we compute a similarity measure between images. Then, it is possible to find pages with similar layout by providing a request image (as in content based image retrieval). A simple distance is computed by the comparison of the labelling of two images pixel by pixel using a Hamming distance (figure 6).






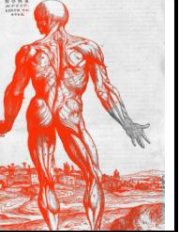


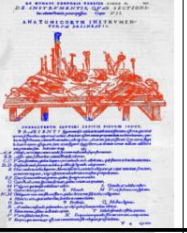




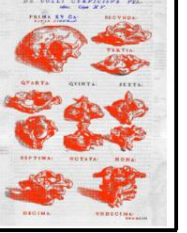
Request Image	Anser by similarity order		
	Top 1	Top 2	Top 3
			
			
			

Fig. 6. Pages comparison using only percentages of text and non text pixels

### 4.3 Precise page layout analysis

The other possible use of the labelling deals with the location of illustrations that must be as precise as possible (sharpness of contours and in the frontiers of graphical parts). Our approach provides valuable information about the possible position of the graphical parts as shown in figure 7.





Labelling result		
Labelling correction		

Fig. 7. Graphical parts segmentation using our labelling

While being focused successively near all the red areas our experiments show that it is possible to merge this information with an algorithm of extraction of connected

components. Thus, by analyzing the labels of pixels contained in all the selected connected components, it is possible to extract in a very fine way the graphical entities without any parameters about the size of the graphical elements in historical books.

## 5. Conclusion

In the first part of the article, we have highlighted the sources of errors of the traditional methods of page decomposition using a characterization of page layout in the historical books. Then we present an efficient method for pixel labelling adapted to ancient printed documents. The originality of our approach lies on the development of a new extraction and analysis tool which separates textual and graphical areas from different kinds of ancient document images without any knowledge like thresholds, models, or structure information. The resulting document labelling is employed as a new feature for a relevant pages comparison. Current results are very promising. The content based classification of an entire book is a direct perspective of this contribution.

## 6 References

- [1] HJ Martin, *La naissance du livre moderne*, Editions du Cercle de la Librairie, 2000.
  - [2] A. Belaid, Computer aided design of models of page for their use in recognition of documents, Workshop on Electronic Page Models, LAMPE' 97. 1997.
  - [3] L O'Gorman, The Document Spectrum for Page Analysis Layout, *Trans IEEE One PAMI*.15(11), 1993 P1162-1173.
  - [4] F Lebourgeois, H Emptoz, E Trinh, Compression and accessibility with the images of digitized documents – Application to the Debora project, *Numerical Document, Flight 7n°3-4*, 2003 p103-127.
  - [5] Jie Xi, J Hu, L Wu, Page segmentation of chinese newspaper *Pattern recognition 2002* 2695-2704
  - [6] D. Malerba, F Esposito Oronzo, Adaptive Layout Analysis of document. Università degli Studi di via Bari, *ismis 2002*.
  - [7] P. Duygulu, V. Atalay A Hierarchical Representation of Form Documents for Identification and Retrieval, *International Journal on Document Analysis and Recognition IJDAR 5 2002* 1, 17-27.
  - [8] S Bres, Contributions à la quantification des critères de transparence et d'anisotropie par une approche globale. PhD Thesis, 1994.
  - [9] W.K. PRATT, *Digital Image Processing*, 2nd edition New- York : Wiley, 1991, p230.
  - [10] [mathworld.wolfram.com/PlancherelsTheorem.html](http://mathworld.wolfram.com/PlancherelsTheorem.html)
  - [11] C. Shin, D. Doermann, Classification of document page images based on visual similarity of layout structures, *Language and Media Processing Laboratory Center for Automation Research University of Maryland*. 2000
  - [12] G. Maderlechner, P. Suda, T. Bruckner, classification of documents by form and content, Siemens AG, Corporate Research and Development, Otto-Hahn-Ring 6, D-81730 Munchen, Germany
- Acknowledgement to CESR-CNRS (France) for providing the images [www.cesr.univ-tours.fr](http://www.cesr.univ-tours.fr)