# Text/Graphic labelling of Ancient Printed Documents

N. Journet[1] V. Eglin [2] J.Y Ramel[3] R. Mullot[1]

[1]*L3I, 17042 La Rochelle Cedex 1 - FRANCE njournet@univ-lr.fr*

[2]*LIRIS INSA de Lyon, Villeurbanne cedex - FRANCE eglin@rfv.insa-lyon.fr*

[3]*LI, 64 Avenue Jean Portalis 37200 TOURS - FRANCE Jean-yves.ramel@univ-tours.fr*

## Abstract

*This paper presents a text/graphic labelling for ancient printed documents. Our approach is based on the extraction and the quantification of the various orientations that are present in ancient printed document images. The documents are initially cut into normalized square windows in which we analyze significant orientations with a directional rose. Each kind of information (textual or graphical) is typically identified and marked by its orientation distribution. This choice of characterization allows us to separate textual regions from graphics by minimizing the a priori knowledge. The evaluation of our proposition lies on a page classification using layout extraction criteria. The system has been tested over several ancient printed books of the Renaissance.*

## 1. Introduction

Since the beginning of large digitalization campaigns, 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). At the beginning of printing, the fonts and the layouts of the pages were very close to the handwritten books [1]. Thus, 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. 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 (see Section 3).

In this work, 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. That means that interest regions can be localized everywhere in the page with a total typography free consideration.

This paper is organized in 3 sections. After a short survey of previous works in layout analysis we show how it is possible to separate text from graphics without any a priori knowledge on the nature of the books and 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 page classification that illustrates the relevance of our layout analysis method. It can be assessed by a quantification of classification results.

## 2. Ancient patrimonial printed documents

### 2.1. A brief study of the characteristics of ancient printed books

For page layouts, the technical and historical constraints impose particular presentations. 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.

This study enables us to draw up a list of characteristics that are essential and that must be introduced in our method of old documents layout analysis. Here are some of them: complex page layout (several columns with irregular sizes), specific and unknown fonts, frequent use of ornaments (borders, ornamental letters), low spaces between the lines (contacts between characters), non constant spaces between characters and words, superposition of information layers (noise, handwritten notes...).

### 2.2. Evaluation of the existing methods on ancient books

The quality of new segmentation methods has significantly increases those last ten years. But the current books have different standards of presentation from historical books. Consequently, the softwares that are devoted to contemporain documents recognition are often unsuitable to the processing of books of the Renaissance period. The methods of structure extraction employed by this software can be classified in 3 main categories: bottom-up, top-down and mixed methods [2].

Traditional bottom-up methods like RLSA, morphological operations, filtering algorithms, connected components ([3,4]) are not adapted to the historical books with all their specificities presented in section 2 (especially non constant spaces between shapes). Thus, the great difficulty of this study 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, number 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 like [8,9] 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…).

## 3. Characterization of informative areas by extraction of orientations

### 3.1. 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 a final 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 final 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 developped in two stages:

- Unsupervised automatic blocks classification using only orientation features. This classification must take into account different book styles.
- Interactive process allowing indexation and semantic labelling of homogeneous areas.

This paper presents the first stage of the automatic process. This first stage is a labelling process that organizes pages into different large classes of blocks: text or non-text. This stage does not need any a priori information (except that text is horizontal).

After being resized, a simple separation based on a homogeneous high grey level pixels estimation of foreground and background is applied to locate information [10]. Then, the image is crossed by a window that extracts specific orientations information and marks each pixel in two main classes: the text and the graphics class.

### 3.2. Extraction of orientation information

The approach lies on the estimation of relevant directions of significant parts of the image: the initial image is cut into normalized square windows in which we analyze significant orientations with a directional rose that has been initially proposed by Bres [11].

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, [12]. Its definition for a bi-dimensional signal is the following.

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

(Formula 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 of the image. The data that are relative to a same direction will be located in a same line. With this principle, it is possible to detect orientations of the texture blocks. 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 analysed image with a Fourier Transform (FFT) which avoids the highly complex development of the correlation. The Plancherel Theorem [13] is at the basis of this simplification. So, in practice, the autocorrelation function is efficiently computed with the image spectrum and not directly with the previous mathematical definition (formula 1).

The autocorrelation result can be analyzed by the construction of a directional rose that reveals significant directions in the analyzed block image. This rose gives with a great precision all privileged orientations of the block. 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 θ as privileged direction of the block. The mean value $E_\theta$ is then defined by the following formula:

$$E_\theta = \left\{ I'(x, y).I'(x + a, y + b) \right\}$$

With $Arctg(b/a) = \theta$ (formula 2)

The directional rose represents the sum R(θi) of different values $C_{xx}(i, j)$ (defined in formula 1) in a given θi direction. So, the directional rose corresponds to the polar diagram where each direction θi that is supported by the Di line, is represented by the sum R(θi). For all points (a,b) of the $D_i$ line we have the following relation:

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

From this set of values, we only keep relative variations of all contributions of each direction. So, the relative sum R'(θi) is the following :

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

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

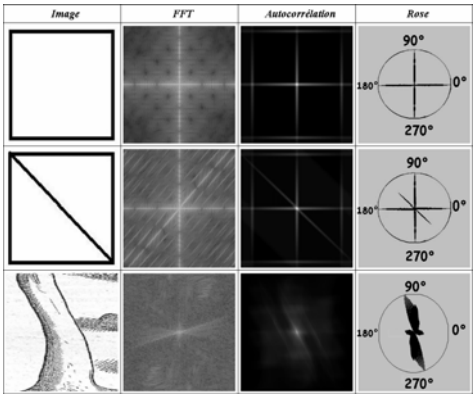


**Figure 1 Example of directional roses**

### 3.3. Characterization of the page layout

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 relatives variations (formula 4) implies that the rose has two invariant characteristics: one "peak" for the horizontal orientations (0° and 180°) and the same intensity for all the other directions (it gives an impression of "bowl").

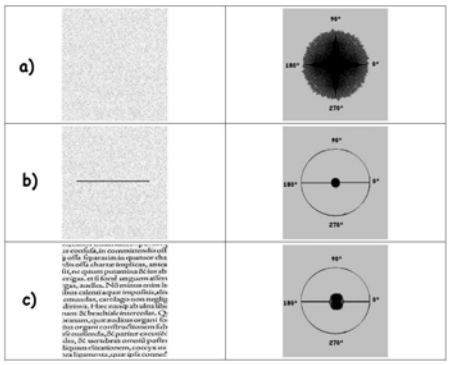


**Figure 2 Text characterization**

A homogenous text area has the same property: all directions are presents 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 2.c). 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 3.a).

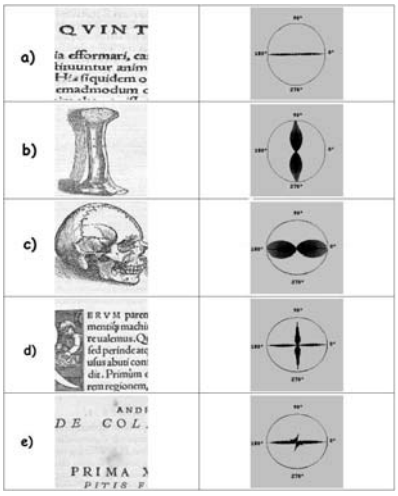


**Figure 3 Graphics and intermediate regions characterization**

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 3.b). In the same way, if the main direction is horizontal without a characteristic "peak" (like in figure 2), then we can conclude that the region is a graphical part (figure 3.c). All the other kinds of roses do not allow pixels labelling: it is often due to border zone analysis (figure 3.d) or to an area where there is not text enough to produce a significant "peak" or a "bowl" (figure 3.e).

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. Figure 4 presents an experimental verification of the predicate. In this study, sizes of analysis area are empirically chosen.

Our tests have shown that a minimum of 4 or 7 text lines is necessary to detect a text block. That is why in our method, the size of the analysis area is constant (128X128 pixels) while image sizes can be variable. Actually, the image is resized manually in order to have the minimum lines required in a 128X128 pixel analysis area.

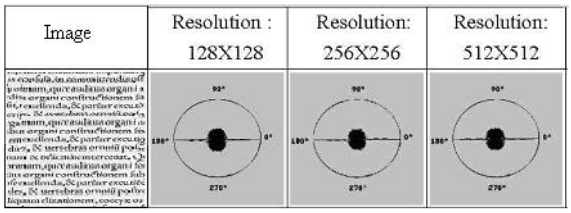


**Figure 4 Influence of different resolutions for the directional rose drawing**

## 4. Experimental results

### 4.1. Quality of characterization

The tests are realized over 100 pages from five different books of the Renaissance. Tests are organized as follow: for text parts, the percent of letters witch were correctly, wrongly or not detected are computed. For graphical parts, the percent of graphical pixels witch are correctly, wrongly or not detected are computed (see tab 1).

**Table 1 layout characterisation results.**
**(% of correct / wrong / forgotten detections)**

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

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.

In most images, all the different parts are well detected using this algorithm. 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).

**Figure 5 Examples of results**

### 4.2. Quality of the layout classification

The results are presented in figure 6 through a page classification based a the visual page content, as Shin and Maderlechner in [14,15] . This page classification presents two advantages. From the one hand it illustrates the quality of the layout characterization, and from the other hand it is used as a robust tool for our current indexation works.

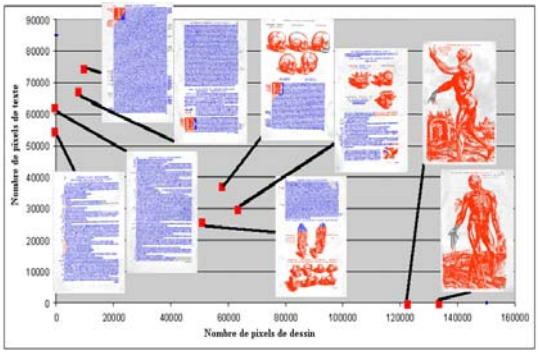


**Figure 6 Example of a pages classification**

Indeed, we think that a complex indexation problem of ancient document images can be simplified by a global analysis of all pages of a book before taking any kind of physical or logical conclusion. For example, as shown in figure 8, pages can be gathered in different classes (with percentages of text and graphics) before achieving the final indexation. By analysing deeply our layout classification, we hope to extract more easily editorial rules, special text (with specific typography) or graphical features to adapt the following processing.

## 5. Conclusion

In this paper we present an efficient method for text/graphic 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 marking is employed as a new feature for a relevant pages classification. Current results are very promising. The content based classification of an entire book is a direct perspective of this contribution. It needs still some more investigation to achieve a complete indexation system especially on non specialist user backtrackings.

## 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 one 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 Orabona 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] P.B Pati,S Raju N Pali, AG Ramkrihnan. Gabor filters for document analysis in bilingual documents. Bangalore INDIA

[9] Memory Efficient Quadtree Wavelet Coding for Compound Images, P. Cosman, T. Frajka, D. Schilling, K. Zeger D.E.C.E. University of California , San Diego

[10] J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen. Adaptive Document Binarization. In ICDARecognition, volume 1, pages 147–152, 1997.

[11] S Bres, Contributions à la quantification des critères de transparence et d'anisotropie par une approche globale. *PhD Thesis*, 1994.

[12] W.K. PRATT, Digital Image Processing, 2nd edition *New- York : Wiley*, 1991, p230.

[13]mathworld.wolfram.com/PlancherelsTheorem.html

[14] C. Shin, D. Doermann, Classification of document page images based on visual similarity of layout structures, *L.M.P. Laboratory*, University of Maryland. 2000

[15] G. Maderlechner, P. Suda, T. Bruckner, classification of documents by form and content, Siemens AG, *Corporate Research and Deĺelopment*, D-81730 Munchen, Germany



Acknowledgement to CESR-CNRS (Tours – France) for providing image bases.http : www.cesr.univ-tours.fr