Characterization and Classification of Printed Text in a Multiscale Context

Véronique Eglin, Stéphane Bres, Hubert Emptoz
Laboratoire de Reconnaissance de Formes et Vision,
INSA de Lyon, Bat. 403 - 20, avenue Albert Einstein
69621 VILLEURBANNE CEDEX
Phone : (33) 4 72 43 80 93 Fax : (33) 4 72 43 80 97
E-mail : eglin@rfv.insa-lyon.fr

Abstract :
In this paper, we present a new method for printed text characterization. This method is based on a text *visibility* and *legibility* criterion. The text is analyzed through its typographic form. So, we propose to label different kinds of text according to their *visual* aspect and their *textural* contents (especially their size, their density but also the line and letter spacing). A scale of legibility and of structural relief of forms has realized this discrimination. The texture is characterized with a statistical analysis of density, which is impervious (insensitive) to our multiscale approach. This statistical analysis is at the basis of the text labeling. This work is a part of a complete scheme of physical and logical document segmentation. It is dedicated to the classification of texts according to their eye-catching properties.

Keywords:
Texture analysis, multiresolution approach, legibility, text composition.

1. Introduction

Most of font classification methods (and most of logical document structuring) need elaborated recognition steps for data labeling. So far, a lot of methods are based on *local* text analysis for a precise characterization of local data (connected components). We have chosen a less common approach that consists in considering the global problematic of printed text classification as a texture characterization.
Although there is no universal definition for the notion of texture, a lot of authors agree that the texture is a set of *visual* and homogeneous characteristics of surface, whose dimensions (also called visual primitives) are perceived and processed instantaneously by the human visual system, [Ni89], [Ju85], [Bo89]. In this context, we consider the text as a texture, insofar as we define the character as the elementary entity of texture. More precisely, a text can be considered as a set of little symbols (graphics), which creates a « macroscopic » impression of texture. Those visual characteristics of texture depend on the letters dispositions, their frequency in the text, the font, and the language.
The statistical analysis that we propose in the following sections is a first step to a semantic interpretation of document, *i.e.* to its significance. The relevant information of a document is often brought out by typographical compositions (location of titles, particular font of text...). In this context, the statistical analysis is based on parameters of density, which display some visual characteristics of font according to different scales. Those statistical parameters allow us to reconsider the great variety

of different writing styles in some groups (or families) with stable and generic properties. Those shape characteristics are mostly connected to the text content. Richeaudeau states in [Ri89] that the eyes are caught by visually great, bold, and large elements and that the hierarchy of information (titles, paragraphs, abstracts...) is highlighted by typographic marks, so as the relevant information in a document is reflected. Therefore, the effect of this composition influences the text exploration (and the document reading); visibility and legibility criteria influence the understanding.

2. Framework

2.1 The great specificity of writing classifiers

Most of current systems for document physical and logical layout analyses are organized around knowledge databases, which describe precisely all different styles of edition and all textural entities. For instance, some searchers have worked on specialized methods of font recognition, which use stochastic models in [An92] or projection profiles of characters in [Zr94]. We can notice that most of writing classifiers are very specific, and used for precise applications in documents analysis (library notices, newspaper pages in standard format...). A solution would consist in multiplying all classes of styles and increase the databases with new requirements and specific data. Those classifiers will nevertheless stay very specific to particular text formats. On the other hand, the original idea (already explored by Charlaix in [Ch96]) consists in finding some *less specific* and *more generic* measurements for each new type of writing based on *relative* and *discriminating* variations of fonts.

2.2 Introduction to the multiresolution

Instinctively, our perception of things leads us to « differentiate » elements rather than to « recognize » them *one by one*, [Le92]. This primary perception is called « preattention » (parallel, instantaneous, without scrutiny, covering a large visual field, as in texture discrimination). This instinctive discrimination is essentially based on a set of local variations of light intensity, of color and of orientation. This phenomenon can be generalized to the perception of a scene however complex it may be, but also to the perception of a document with its great variability. Some more precise physiological measurements show especially that our perception of things is not homogeneous over the whole visual field, and that our perceptive mechanisms are based on a space-variant pavement of the scene, which yields to a mutiresolution perception, [Bo91]. Moreover, we are able to tackle the complexity of document structures with a mutiscale perception of things: a global vision in low-resolution gives us the global structure of the page, while a local vision in high-resolution reveals the local structure of little connected components (characters). In the same way, the Gestalt psychologists formulated a number of principles of perceptual organization (proximity, similarity, common fate, good continuation, closure, prägnanz..). Some of their principles concern primarily the grouping of sub-regions of figures, and others concern the segregation of figures from background

(Wertheimer 1923). Especially, things that look « similar » (as characters of a word, or words of a line of text, and also as lines of a paragraph) are grouped together as a figure. We have been inspired by all those perceptive phenomena in our approach of text characterization.

A multiresolution representation of a text consists in considering the text according to various frequency contents. When we *closely* perceive an image, we can observe its whole original spectrum; the image is in *high* (or full) *resolution.* So, under lowest resolutions, the loss of information leads to a global perception of the text. Our aim is to find a characterization of the different text zones that is invariant to a change of resolution of the picture. What we want to distinguish and characterize is the fact that some part of a document contain images and some other parts contain text, even if the scale makes it no longer readable. As a matter of fact, we do not have to be able to read a text to recognize it as a text, because of its very typical texture. Therefore, it is natural to establish a bound between text and texture. In this particular domain of document and text analysis using a texture approach, we can mention Jain who proposed in [Ja92] a new approach for the location of text zones. This approach considers the text as a textured entity. It appeared to be an interesting way but still few exploited in document analysis. Then, a text is no longer considered as a juxtaposition of related binary entities, but as a unique pattern, and we will consider parameters coming from its texture like visibility, legibility or relative importance of a police to the other.

3. Our contribution

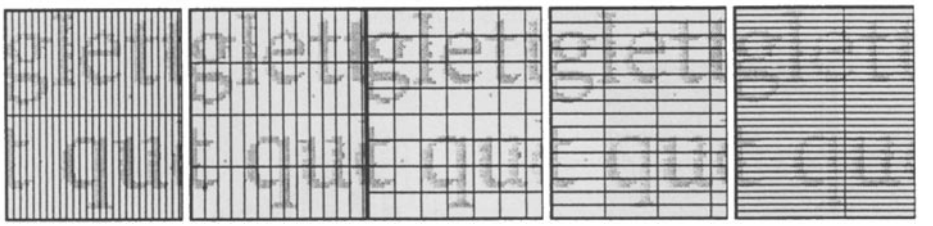
3.1 Statistical approach and multiresolution context

The main idea of our approach is to use a treatment as simple and as generic as possible in order to establish a hierarchical relation of visual importance between the different text zones of a whole page of document. We do not want to characterize precisely the different types of fonts used in these different zones, because this work needs a special characterization effort, with a lot of very accurate parameters. We just want to establish relations of resemblance or dissimilarity using simple but discriminative statistical parameters. Our method is directly applied on portions of homogeneous text: it is based on the evaluation of density parameters computed on a set of pavements. This set of pavements allows to consider the text according to its two preferential directions: horizontal and vertical. These measures lead to a very informative description of the fonts. The results are plotted on 3D-graphics. Each font provides a characteristic signature from which, we extract parameters of bold type, spaces inter-characters and inter-lines. We will finally show the stability of these values through scale changes on under-sampled pictures.

3.2 Use of the density for the characterization of fonts

We do not presuppose any knowledge on what the inspected zone contains. It can be text or picture. It is just assumed that it is usable as it, that is to say that typeface

families (if it is a text zone) are of a sufficient size to be legible or identifiable. Once again, our aim is neither to recognize these families in a precise way nor to recognize the text itself, but just to classify them according to their texture, which must be discernable. This density computation consists in measuring the mean gray level of the pixels in boxes of constant surface and variable shapes. These boxes are part of a regular pavement that covers the image and this can be done on binary or gray level images. The boxes are of a constant surface but of variable measurements in height and in width in order to go from greatly vertical boxes (basis of one pixel and height equals to the surface) to greatly horizontal boxes (width equals to the surface and height of one pixel). This allows an inspection of the text in term of densities and according to its own preferential directions: the horizontal (the one of lines) and the vertical (the one of paragraph sides and characters). The choice of the boxes surface is motivated by the range of character sizes that one wants to analyze, and therefore by the level of resolution of the document. In practice, we used two surfaces that appeared to be interesting for common resolutions of text documents, but also, suitable to a multiscale treatment, as it will be shown in what follows. The first surface is 64 pixels and the second is 256 pixels. We use seven (respectively thirteen) different kind of boxes configurations, from 1x64 to 64x1 (respectively 1x256 to 256x1) as it is illustrated on figure 1.



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

Figure 1 Representation of isotropic and anisotropic boxes involved in the density calculation. 1x64 (not represented), 2x32 (a), 4x16 (b), 8x8 (c), 16x4 (d), 32x2 (e), 64x1 (not represented).

For each analyzed picture, we represent the histograms of the density variation for the 7 (or 13) kind of boxes. These histograms constitute a signature that characterizes the inspected font. It allows distinguishing one font to another. Examples of results we obtain on printed text and on pictures are given on figures 2 and 3. These histograms are able to distinguish two text portions made of identical font text but different bold type. The bold text presents very represented weak density groups for horizontal boxes a more spread out distribution for vertical boxes. On the opposite, non-bold text presents histograms shifted toward the strong densities (figure 2). If the inspected image does not contain a text but an image, these curves are close connected to the histogram and the influence of the boxes orientation is very weak. Images are more isotropic than texts (figure 3). This fact can be a way to discriminate text from images in most of the time, but it can not be generalized in all situations. In this paper we will only consider text blocks.

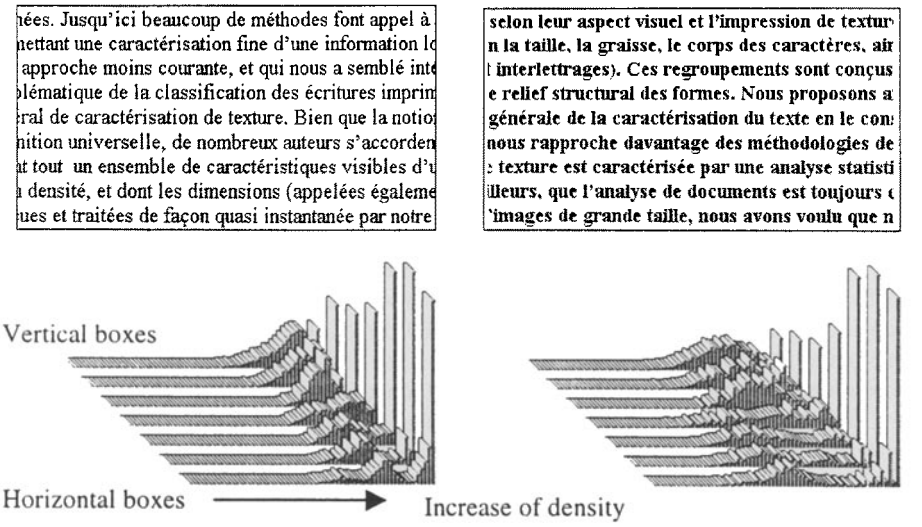


Figure 2 Profile differences between normal and bold polices. Each histogram represents the computation of density on different pavements (from vertical boxes in the background to horizontal boxes in the front).

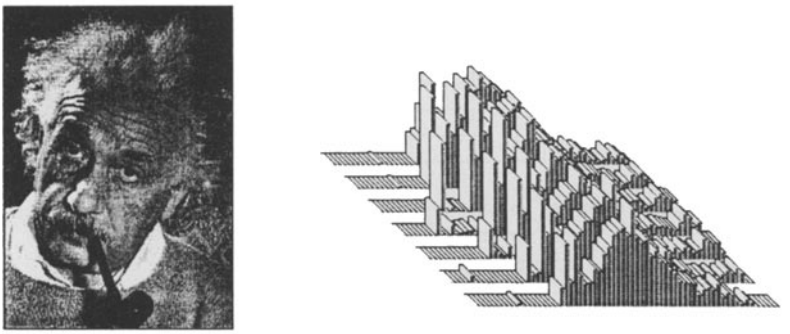
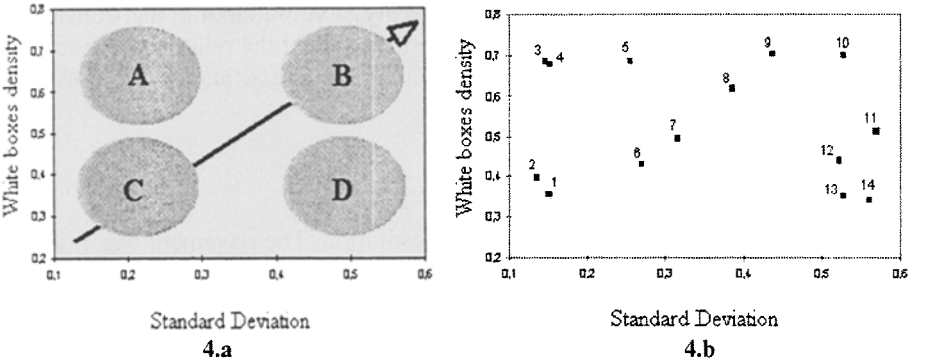


Figure 3 Profile obtained for a gray level image.

3.3 Legibility and form criteria using these measures

These density measures are sufficient enough to distinguish images and text, and even to distinguish different types of printed text. From these measures, it is possible to extract information of relative importance of a block of text in the image (according to the visual interest that one can take in it) as well as its legibility. One agrees to say that a text is legible when the spaces inter-lines are proportional to the size and the bold type of characters (proportion ratio of 1 to 2). We chose these two parameters of attractive power on the reader and legibility, because they have a direct influence on the research of information in the document. A combination of these two parameters is presented on a 2D-space (figure 4). In this figure, we have represented the most significant text portions.



Domain A : Large line-spacings. Small and/or thin characters. **B :** Large line-spacings. Bold and/or big characters. **C :** Thin line-spacings. Condensed characters. **D :** Thin line-spacings. Bold and/or big characters. The arrow represents the increase of structural relief from low to high importance of typefont families.

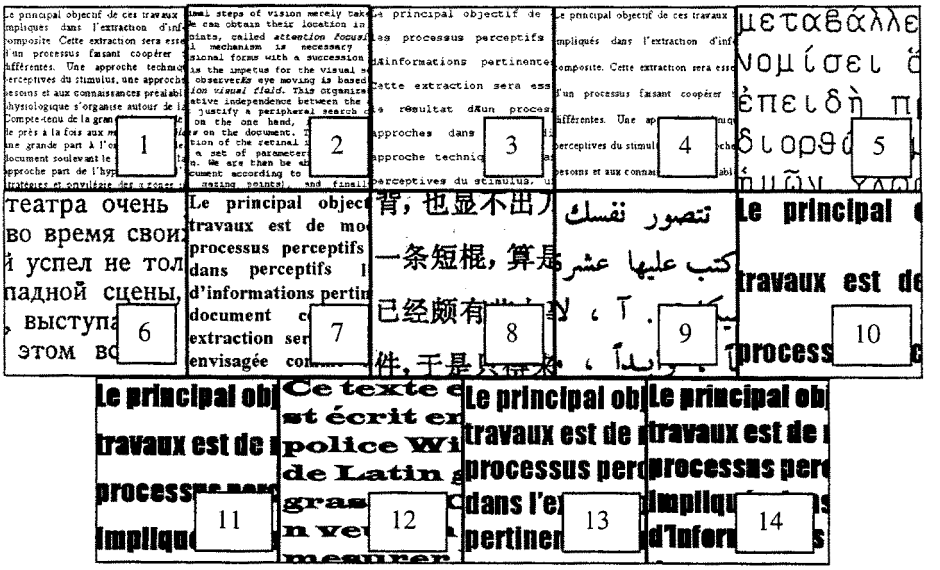


Figure 4. **a.** Theoretical typeface families by legibility and visibility criteria. **b.** Distribution of a font sample according to their structural relief. The numbering of the samples and the numbering of the points in the graphic are matched.

It is the representation of the rate of white boxes in the pavement as a function of standard deviation of text density measures. A white box is a box that contains more than 95% of white pixels. We do not take 100% to decrease the influence of impulse noise that appears on figure 5a. We consider that this impulse noise can be as high as three black pixels on a sixty four pixels pavement. Those 5% have been obtained from this limit. If the impulse noise is higher than 5%, we should modify the rate 95%/5%. Nevertheless, the results are relatively representative of the type font until 10% of noise. A value of standard deviation is calculated for every type of pavement

used. This value is representative of the density distribution on the considered pavement. These statistical parameters are able to highlight the relative tendencies to the legibility and importance of the fonts. Examples of classification are presented on figure 4.

3.4 Invariance through scale changes

The previous classification was made on full resolution. The pavement was made of square boxes of surface 64 pixels or 256 pixels. The choice of these square boxes was motivated by the fact that it is possible to recover exactly the same results and thus, the same fonts classification using much smaller gray level images. These images are built by the classical multiscale technique of averaging 4, or 16, or 64, or 256 points belonging to the same square zone of a previous step to get a pixel of the next. Each step gives an image. Results obtained on the full resolution image, with a pavement of square boxes of surface 64 pixels (respectively 256 pixels) are exactly the same as the ones obtained on an image of size 64 times (respectively 256 times) weaker while using pavements 1x1 (therefore square). It means in practice that our analysis gives exactly the same results on a text image in full resolution of 4096x2048 and on its reduced version 512x256 (respectively 256x128), if the first analysis is achieved using square boxes of surface 64 pixels (respectively 256 pixels) (see figure 5). It is of a great interest for computer treatment of text documents since the number of points to analyze is thus highly reduced. However, it is important to notice that achieving an analysis on reduced pictures decreases the number of possible box shapes. Therefore, it is less and less possible to have information on the anisotropy feature of gray level distribution, and at a surface of 1 pixel for the boxes, it only remains the square pavement.

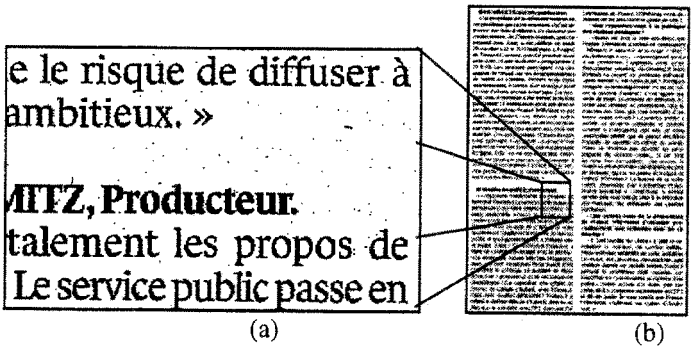


Figure 5 A part of the original document (a) and the analyzed image reduced by a factor of 64 (b).

4. Conclusion and further work

In this article, why we have chosen to characterize the global aspect of a text with *macroscopic* features of texture. A text gives an impression of unity because of the

repetition of its elements (the letters) according to a precise spatial organization. We have characterized this unity by the exploitation of statistical parameters.

This work is a part of a complete project of document structuring. That is the reason why, we do not have presented the preliminary step that corresponds to the segmentation of document in blocks. A method of physical segmentation has been proposed in [Eg97]; it is an essential part of this work. It is also possible to use other techniques of document structuring, if they allow to separate the text blocks and the images, [Ak93], [We93]. So, the textural classification of blocks is a first step for the document logical structuring, which consists in identifying all components of structure (titles, abstract, paragraphs, captions, commentaries...). Finally, this method can be adapted to a change of scale. So we can quickly process large images by considering their reduced representations. Moreover, those images do not presuppose any knowledge about the text. A more complete exploitation of the statistical parameters (we only have presented an aspect of it) allows us to characterize even more precisely other types of printed text.

5. References

[Ak93]Akindele, Belaid, *A labeling approach for mixed Document Blocks*, Proceedings of ICDAR 93, 2nd International Conference on Document Analysis and Recognition, Japan, pp.749-752, 1993.

[An92]J.C. Anigbogu, *Reconnaissance de textes imprimés mulitfontes à l'aide de modèles stochastiques*. Thèse de doctorat : Université de Nancy, 1992.

[Bo91]C. Bonnet and B. Dresp. Psychophysique de l'extraction des contours en vision humaine, *RF IA* 3,102-109, 1991.

[Ch96]E. Charlaix, D. Derrien-Peden, *Reconnaissance de la structure logique de documents par programmation par contraintes sur les styles*,CNED'96, pp.61-68, 1996.

[Eg97]V. Eglin, Structuration de documents par une modélisation floue de l'information visuelle. *Logique Floue et Application, LFA'97*, 10p., 1997.

[Ja92] A.K. Jain, K. Bhattacharjee, *Test segmentation using Gabor filters for automatic document processing*, MVA'92, pp. 169-184, 1992.

[Ju85] B. Julesz, *Preconscious and Conscious Processes in Vision.* Bell Laboratories. Murray Hill, New Jersey. Pattern Recognition Mechanisms, pp. 333-359, 1985.

[Lec92] J.C. Lecas, *L'attention visuelle*, Liège : Pierre Mardaga, 1992, 310p.

[Ni91] J. Ninio, *L'empreinte des sens, Perception, mémoire, langage.* Odile Jabob, 310p., 1991.

[Ri89] F. Richeaudeau, *Manuel de typographie et de mise en page*, 1989.

[We93]J. Wey-Wen, H.E. Meadows, *Classification and compression on digital newspaper images*, VCIP, pp.96-105, 1993.

[Zr94] A. Zramdini, R.Ingold, *Optical font recognition from projection profiles.* Third International Conference on Raster Imaging and Digital Typography, Darmstadt, Allemagne, 1994.