IJDAR (2004) 7: 28–43
Digital Object Identifier (DOI) 10.1007/s10032-004-0127-2

International Journal on Document Analysis and Recognition
IJDAR

Analysis and interpretation of visual saliency for document functional labeling

V. Eglin, S. Bres

INSA, LIRIS / CNRS FRE 2672, 20 avenue Albert Einstein, 69621 Villeurbanne Cedex, France
e-mail: eglin@rfv.insa-lyon.fr

Received: 6 December 2003 / Accepted: 22 December 2003
Published online: 12 August 2004 –

Abstract. In this paper we propose a complete methodology of printed text characterization for document labeling using texture features that have been inspired by a psychovisual approach. This approach considers visual human-based predicates to describe and identify text units according to their visual saliency and their perceptual attraction power on the reader's eye. It supports a quick and robust process of functional labeling used to characterize text regions of document pages. The test databases are the Finland MTDB Oulu base[1] that provides a great panel of document layouts and contents and our laboratory corpus that contains a large variety of composite documents (about 200 pages). The performance of the method gives very promising results.

Keywords: Texture analysis – Text characterization – Functional labeling – Document layout – Psychovisual exploration

1 Introduction

1.1 The document as message conveyer

A document editorial work is a necessary step to organize data, to represent an ideas hierarchy, and to give readers a global impress of coherence and efficiency in the document exploration. This work constitutes the editorial chief that precisely reveals the author's will to transmit a message. In that context, Maderlechner in [16] claims that the reader's attention and reading speed strongly depend on the layout of a document. We can notice that among the great variability of documents and even normalized page layouts (scientific papers, newspapers, advertisements, etc.), it is not easy to access retrieved information rapidly and correctly. Thus, for an automatic system of information retrieval and page object recognition, it becomes more and more difficult to recognize and analyze document layout: this expanding research field needs an increasing number of dedicated and specific approaches for each class of documents. In that context, we believe that placing human beings at the heart of document decoding process, like Nagy in [18] and Doermann in [5], is an interesting way to characterize documents with a particular focus on attractive and emergent information. According to the document type (Doermann speaks about functional class in [5]), information is not perceived in the same manner by the reader. As for Doermann, when documents are regarded as message conveyers, they can be classified according to the type of message that is conveyed. In a document corpus, we can then be interested in categorizing documents according to their editorial proximity, which is strongly correlated with the message sense. In our work, we propose a functional description of documents based on the interpretation of the physical structure by using texture primitives.

1.2 Functional organization of documents

The functionality concept. In the field of document understanding, documents have traditionally been viewed according to their geometric and semantic organizations. Both organizations have a common content that represents the basic level of data (texts, graphics, and images). The physical organization of a page can be obtained by a low-level characterization of information that leads to a geometric segmentation into blocks. So as to recover the logical organization of a page, we need precise knowledge on the kinds of documents under investigation. This analysis leads to a complete high-level labeling that gives a precise sense to the physical layout. Between these two extremes we can define an intermediate level that is known as *functional organization*. At this level, we are interested in how physical features in the page can be used by the author to organize and convey his message. The functional level relates to the efficiency with which the document transfers its information to the reader. The physical representation of the message is supplementary information to emphasize ideas in the page and to under-

[1] J. Sauvola and H. Kauniskangas (1999) MediaTeam Document Database II, a CD-ROM document image collection, Oulu University, Finland

a **b**

Fig. 1a,b. Examples of header black-surrounded blocks of documents having a common functional description but a different logical meaning

line their hierarchy. The constraints that will be taken into account by the system dedicated to document logical analysis are not the same as those for a functional analysis: in the first case, the system must recognize physical objects according to their location on the page and their conformity to the reference model, whereas in the second case, it must be able to focus on eye-catching and attractive information that will be useful for the reader. The functional organization of documents that have been recently introduced by Doermann in [5] is the starting point of our work. In his research, Doermann has studied the relationships that exist between the physical, the functional, and the logical descriptions of the document.

As an illustration of the relationship between the physical, the functional, and the logical organizations of documents, let us consider a text block at the top of a page. The physical analysis of the block gives its precise dimensions and location on the page in relation to other text blocks on the page. It also informs on the spatial proximity of inner components that form the block. The functional interpretation of the block based on the block's attributes concludes that the block is a *header*. The logical interpretation gives more precise information on the block class: it concludes that the block is a *title*. In another context, a header block can also be a head note, a letterhead, a subtitle, or many different things. In Fig. 1a the heading block represents a head note using a bold font style and in Fig. 1b it corresponds to the main title of the page. In both cases, the functional description concludes that blocks are headers.

In his work [5], Doermann considers that the functional description of a document is independent of the document type: the categories of blocks can be chosen from among *headers, footers, lists, tables, graphics*, i.e., generic categories that are common to many types of documents. In our work, we give more precise functional descriptions of blocks: we can speak about pseudological descriptions of blocks. We have based this description on three visual families: the family of headings (page titles), the family of body paragraphs (standard paragraphs of text), and the intermediate family of salient and/or dense regions of text (like salient abstracts, subtitles). This description is derived from physical and texture properties that are presented in following sections. Applications of the concept of functionality can be found in the works of Schreyer and Maderlechner in [15] and [16]. They propose a method based on the Julesz theory ([11]) to develop hierarchical bottom-up segmentation and a texture-based font-style classifier by defining an attractiveness indicator for text blocks.

Functionality concept for document labeling. We have chosen to base our work of document interpretation on the concept of functionality and pseudologic. The document interpretation module of our system that is based on text characterization leads to a pseudological description of text blocks of documents having a standard editorial chief with a stable description of text components on the same page: for example, typographical tools (size, boldness) used to represent titles are the same on the same page. This principle of editorial stability must be applied not only to the whole page area but also to all pages of the same document (in the case of multipage documents). This situation is often encountered in our test base. Especially here we have focused on Latin documents containing horizontally written text blocks with some a priori knowledge, for example contrasted and bold head titles, small written text paragraphs, the existence of legend beneath (and not above) each image or graphics, etc. We have applied the functionality concept to document labeling by defining generic functional families for text blocks. This concept can be then derived in different applications starting from the text characterization module: for example, we are currently working on a new approach to document classification based on the analysis of the visual layout saliency of the page composition that is given by our functional description. The text characterization process is based on the definition of visual texture-based features that are interpreted as *complexity*, *visibility*, and *compactness* indicators. They are used to characterize text blocks of documents. In our experiments, we consider characters, graphic blocks, and images as basic component units. We also assume that the document has been separated into basic blocks of text, images, and graphics as is represented in the MTDB Oulu test base and in our own laboratory corpus.

1.3 Paper organization

The organization of the paper is as follows. In Sect. 2, we present some psychovisual aspects of text perception including recent works on texture-based document analysis. In Sect. 3, we present the text characterization process by the global description of the successive steps of page processing. Section 4 presents the texture-based features that are applied to functional labeling. Section 5 presents in detail the labeling decision tree and the results obtained in the MTDB Oulu database and our personal corpus. Finally, Sect. 6 is an enlarged discussion

of the proposed method of text characterization and its application to page labeling and document classification. The discussion presents a comparative analysis between existing works in the field of document labeling and our texture-based approach.

2 Text and texture as a psychovisual reality

2.1 Psychovisual approaches of text perception

Some recent approaches that are relevant to the perceptual organization of information present the fundamental rules of "pregnancy", "complexity", and "good form". The gestalt theory has introduced some new concepts dealing with the *unity* and the form *stability*. The principles of element organization and space arrangement have been introduced. Those principles are at the basis of our human perception. In this theory, elements are grouped together according to proximity, good continuation, and similarity principles. The global perception of text units derives from the combination of those principles. For example, when we use white spaces as separators, the principle of proximity, which states that elements that are closer tend to be merged together, is applied (Fig. 2). A more recent formalism has been introduced to characterize the forms according to complexity, unity, symmetry, and continuity. The authors have tried to find objective criteria of "good form" such as the numbers of continuous lines of the contour and the number of corners. Those properties have been developed by David Marr for the primal sketch description [17][3]. Another fundamental work has been proposed by Julesz on texture image that confirms the basic hypothesis of stability, unity, and good form [11]. Thus, because the transfer of information to the reader of a document is done using vision as the privileged medium, documents are often designed in accordance with those perceptual principles. That is why this work is strongly influenced by a physiological and psychological approach to human visual perception. The texture has been chosen as a privileged descriptive tool because its definition relies on visual human-based considerations. The texture is a powerful visual indicator that has often been associated with a macroscopic image analysis [20]. In the document analysis context, the texture has been introduced to underline emergent visual characteristics of text in different resolutions [10]. In this paper, we have tried to characterize the hierarchy of text areas in a document page by analyzing their saliency and pregnancy and by featuring the text structural relief, the complexity, and the local density with appropriate measures.

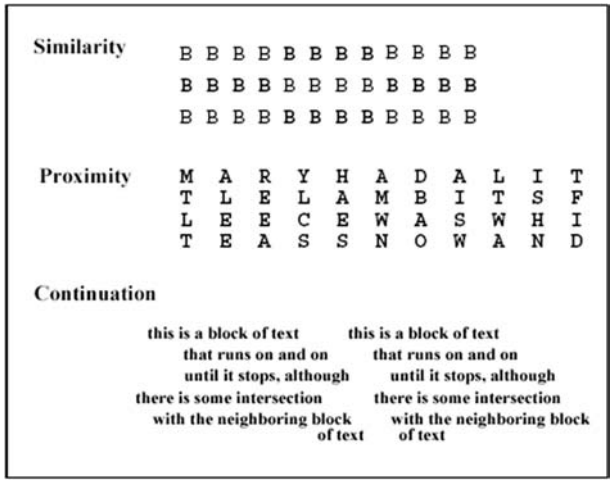
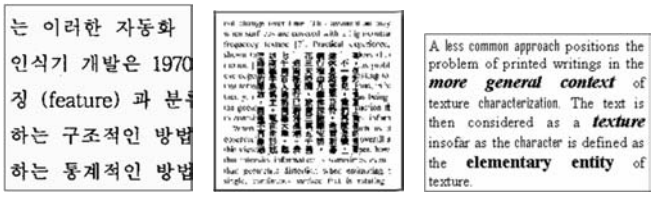


Fig. 2. Application of the gestalt theory to text perception [5]

2.2 Texture-based approaches in document analysis

Currently, most of the font-classification methods (and more generally most methods of document logical-structure analysis) use approaches based on connected components of word images and physical features of text zones [26]. Most studies involve a geometric analysis such as horizontal projections, word shapes [27], or histograms of black pixels for each scan line [24]. These methods of font classification are based on the detection of connected components and on the creation of bounding boxes in the preprocessing phase. This research is specialized in script categorization and it uses a very local characterization of components. It also heavily depends on the initial image quality, and the accuracy of local and geometric methods is generally high [24]. Other studies involve categorizing blocks into text and nontext classes. For example, Bergler [1] uses spatial features such as block size, distribution, and alignment of the bounding boxes of connected components. In [12], the authors propose a multifont classification system based on a local analysis of typographical attributes. In [21], the authors extract features for each text zone such as run length mean, spatial mean, or zone width ratio and use a decision tree classifier to assign a zone class on the basis of its feature vector. Another example of geometric and connected-component-based feature analysis is also proposed in [14], where the authors have developed a feature-based zone classifier using the knowledge of the width and height of connected components. Finally, in [13] a system for automatic text zone labeling using labels such as titles, authors, affiliation, and abstracts is proposed. The page layout and some generic typesetting knowledge for Latin text characterization are used as input data to a neural network.

A less common approach considers the problem of printed writings in the more general context of texture characterization [6,7,19,23]. The text is then considered as a texture insofar as the character is defined as the elementary entity of texture. More precisely, a page of text can be considered as a set of small graphics, the characters, that generate a *macroscopic* impression of texture. Visual characteristics of this texture depend on the arrangement of the letters, their frequency, font style, boldness, italics, and alphabet (Fig. 3).



a **b** **c**

Fig. 3. Examples of mixed texture using two alphabets – Latin–Korean (**a**), Latin–Chinese (**b**) – and an arrangement of boldness, font styles, and italics (**c**)

In our study, the texture elements are the text characters, and our purpose is to analyze their drawings, density, and organization in the blocks. Texture-based methods have been proposed recently: they are more generic, more global, and often content independent, like the font-recognition method based on a 2D Gabor filtering technique proposed in [28]. In that context, we can also mention the work of Chetverikov, who proposes [4] an approach based on the autocorrelation function to characterize blocks. Jain and Zhong have also introduced the concept of texture analysis in a context of text characterization [8]. In those works, texture is a tool used to format text units in segmentation modules or to discriminate text and nontext blocks on the same page, whereas in our work it is used to categorize text blocks in functional families. We have attempted to use as generic a treatment as possible in order to establish a hierarchical and visual relation among the different text areas of the same page. For the page labeling that is the goal of our work, we do not need to precisely recognize the different types of fonts used in text.

3 Fundamental working hypotheses

3.1 Page layout stability

The general principle of text characterization that is the first step in the process of document labeling and classification is based on three fundamental hypotheses of page layout stability:

- Hypothesis H_1: On the same page, text blocks having a common functionality (titles blocks, subtitles, text paragraphs, headnotes, footnotes) are represented with the same typographical tools. Thus, the hierarchy of text blocks (page titles, subtitles, text paragraphs, notes) is highlighted by a hierarchical typographical composition. In that context, it is possible to define *relative scales* for text block representation on the same page. This notion of relativity is fundamental here.
- Hypothesis H_2: In the same category of documents (scientific papers, information newspapers), page layouts are stable. That means that several pages of the same document category can be processed together and text block classification will be made for the whole document. In that case, the classification is generally more accurate because all the different kinds of text categories are represented functionally (titles blocks, subtitles, text paragraphs, headnotes, footnotes). It is useful for the great corpus or multipage documents.
- Hypothesis H_3: The last hypothesis consists in a transversal stability in the whole corpus: the rules that are developed for text block characterization can be applied to diverse categories of documents that also respect the first local stability hypothesis. That means that documents having a stable representation of text hierarchy can be correctly processed by our system.

The diverse categories of page layouts that we have chosen to take into account and that we have encountered in the corpus are characterized by the existence of three main functional families having generic and stable properties: a *titles* family called F_1 (grouping page titles, video inversed text areas, or especially thin titles), an *intermediate* family called F_2 of salient texts including subheadings (also called subtitles) and pregnant paragraphs that often correspond to salient abstracts. This second family presents intermediate eye-catching characteristics in the page layout. The last family, F_3, is represented by text *paragraphs* and contains elements such as standard paragraphs (single or multicolumn) and figure captions and includes all localized information in only one text line such as headnotes, footnotes, or isolated text lines. Figure 4 illustrates this separation in three families.

3.2 Ground truth document structure and ideal page segmentation

Our system starts with segmented pages in homogeneous regions that are then analyzed in their bounding boxes. A region is homogeneous if its entire area is of one type: text, figure, title, etc. Each text line of the page lies entirely within one text region of the layout. In this work, we have chosen to analyze documents that have already been segmented so as to concentrate our efforts on text block characterization (Fig. 5a). In the results presented here, we will use the ground truth document structures of the Oulu database and of our personal corpus. We note here that segmentation greatly influences text characterization as well as text block labeling and page classification. Consequently, segmentation has to be properly realized.

As an illustration of the influence of text block segmentation, we present in Fig. 5b an example of bad segmentation that can lead directly to a wrong text characterization. In these examples, some blocks, indicated by gray arrows, contain information with different visual pregnancy. A texture block analysis will give a unique estimation for the whole block even if it is not homogeneous for that point of view. Those situations are often encountered in complex structured pages, like advertisement or magazine pages [9].

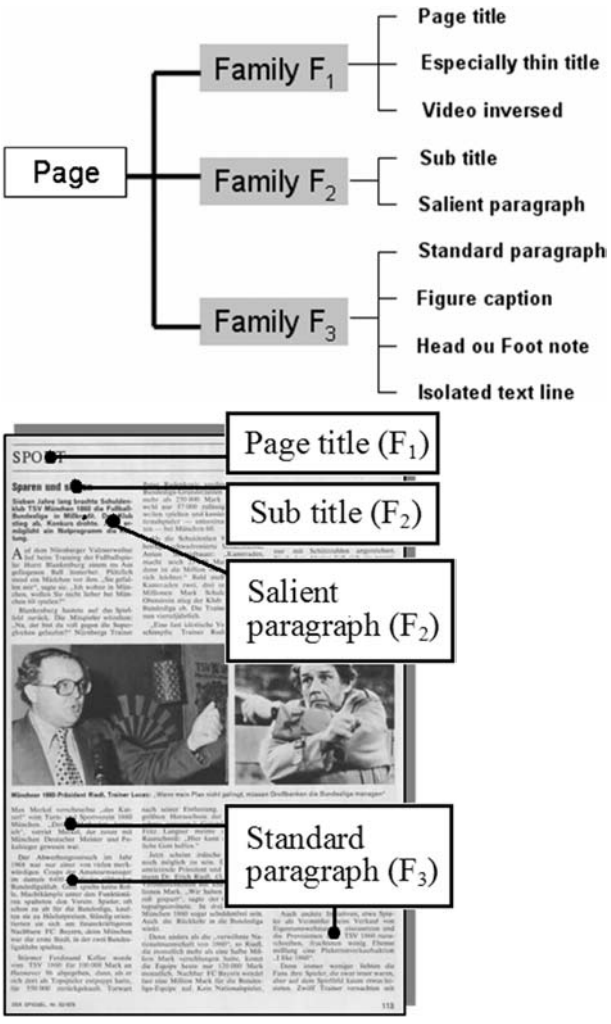


Fig. 4. Families and application on a page of the MTDB Oulu database

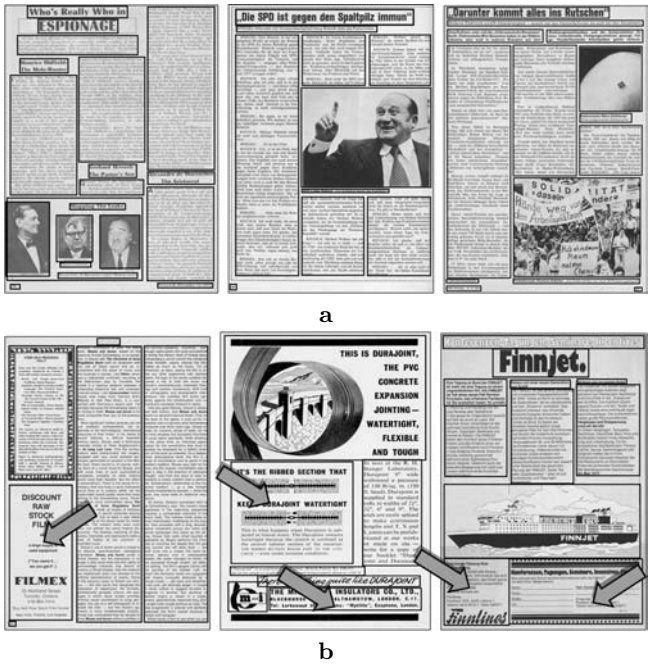


Fig. 5. **a** Examples of well-segmented pages in the OULU database. **b** Example of bad segmentation that can lead to misinterpretations

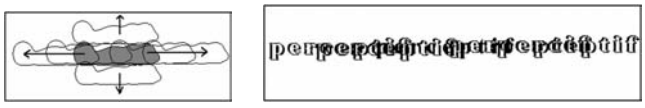


Fig. 6. Privileged orientation of (**a**) a smoothed word and (**b**) a set of connected components by autocorrelation [2]

4 Block characterization process

Before text block characterization, we have to know which blocks on the page are text blocks and which are not. This discrimination is the first step of our labeling process.

4.1 Text block/nontext block discrimination

In this step, we disregard all blocks whose areas are less than 0.5% of the global image area. They are too small to have representative texture features. The text and nontext block discrimination process is based on the analysis of the autocorrelation function, often used for texture characterization. It allows one to determine the main block orientation. We can mention here Chetverikov's works that lead to a classification method based on textural characteristics [4]. Strouthopoulos [23] proposes an approach based on a set of primitives tuned in a neural network to discriminate text and nontext blocks. In our method, we use an autocorrelation function that correlates an image with itself and highlights periodicities and orientations of texture. The definition of the autocorrelation function for a bidimensional signal is

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

The autocorrelation function $C_{II}(i,j)$, 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 the same direction will be located in the same line. This principle makes it 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. Conversely, in the orthogonal direction of this line the resulting value will be low. The autocorrelation underlines the objects' overlapping that is obtained by translation (Fig. 6). 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.

Figure 7 presents two examples of autocorrelation results for two different segmented blocks (a textual block

and an image). The autocorrelation image on Fig. 7a is representative of text lines with a uniform repartition of horizontal gray-level lines. The autocorrelation image in Fig. 7b presents a less uniform distribution of orientations: the second image cannot be assimilated as a text block image. The autocorrelation result can be analyzed by the construction of a corresponding directional rose. This rose gives with great precision the privileged orientations of the block. In [2], we propose an approach to directional rose computation based on the mean value computed from the autocorrelation result. Let us consider I the block image and (x, y) the set of coordinates in this image. We also consider θ as a privileged direction of the block. The mean value E_θ is then defined by the following formula:

$$E_\theta = \{I(x, y).I(x + a, y + b)\}, \tag{2}$$

where $\theta = \arctan(b/a)$.

The directional rose represents the sum $R(\theta_i)$ of different values $C_{II}(i, j)$ (defined in Eq. 1) in a given θ_i direction. Thus, the directional rose corresponds to the polar diagram where each direction θ_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:

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

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

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

Examples of relative directional roses are given in Fig. 7. With this approach, we keep only blocks that are represented with a horizontal principal direction and with isotropic values for all other directions (that are represented in a circular distribution of values in the rose; see Fig. 7a). In the directional roses, we detect local extreme values and keep the values that are greater than the extremes' average. The horizontal extreme value can easily be detected with a tolerance percentage around the horizontal direction. The tolerance angular domains are $[359, 1]$ and $[179, 181]$. All blocks that belong to these domains are considered as text blocks. With this approach, the results of the autocorrelation function in segmented text blocks are illustrated in Fig. 8.

4.2 The general principle

After this first step of nontext block extraction, we consider that we only have text blocks to analyze and characterize. The general principle of text block characterization is summarized in the following scheme (Fig. 9). For each text block, we determine a set of features: geometrical measures, measures of complexity and visibility, directional compactness, and location values as described in Sect. 4.3.

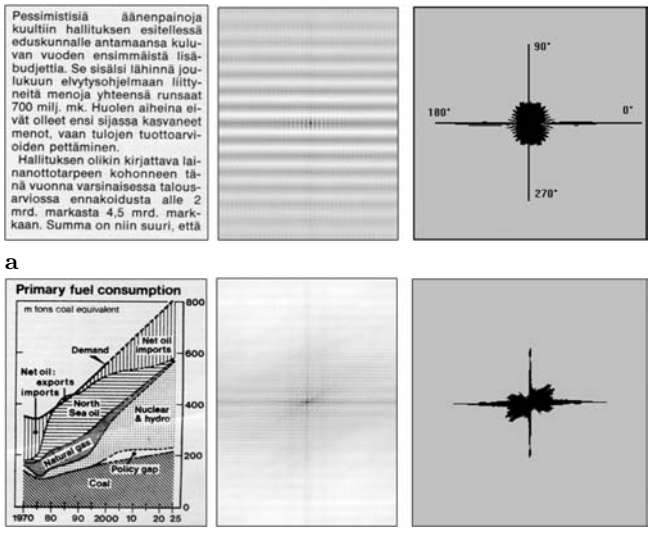


Fig. 7a,b. Two examples of directional roses: initial image, autocorrelation results, and relative directional roses (from *left* to *right*)

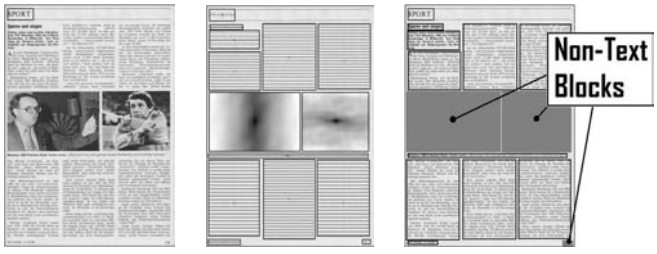


Fig. 8a–c. Results of block discrimination on a segmented page. **a** Original image. **b** Result of autocorrelation in segmented blocks. **c** Text block selection by autocorrelation analysis

On the basis of the two measures of complexity and visibility, we build a 2D-feature space where each block is represented by a point. A k-means method is then applied on that set of points, and each block is classified into a visual cluster defined in the complexity/visibility space. The number k of classes is fixed at 5. Section 5.1 presents this method in detail. This step leads to a decomposition of pages in five visual classes – C_1, C_2 to C_5 – that are strongly correlated to the initial functional families F_i.

4.3 Texture features as expressions of text saliency

Relevant psychovisual text dimensions. In this section, we present the different texture features that have been chosen for their psychovisual properties, their relevance, and their robustness to initial image quality. We have formulated the hypothesis that there exists a hierarchy of text blocks in a page according to their function (see hypothesis H_1). To highlight and quantify this hierarchy, we chose two complementary features: the *complexity* and the *visibility* computed for each text block of a

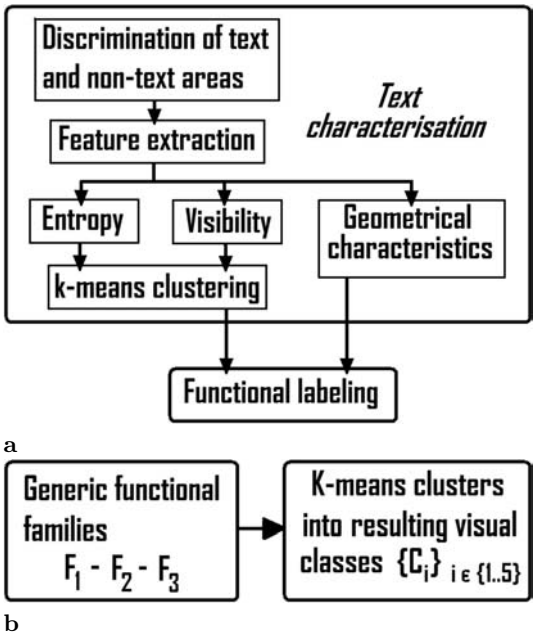


Fig. 9. a Text block characterization, labeling step, and application to page classification. **b** Text classes considered during the process

page. The complexity underlines the frequency of transitions between text components, whereas the visibility estimates the density of these transitions. Complexity and visibility are two complementary features that are a priori not correlated. Nevertheless, a correlation exists in practice: the boldness of a character is often linked to its size and the greater characters are often the less complex ones (in the normalized Latin typographies). The combination of these two complementary measures is expressed by a basic 2D-feature space (called *saliency graph*) in which each text block is represented by a point. It leads to a first classification into visual clusters (the $C_{i\ i\epsilon\{1..3\}}$). In [5], Doermann pointed out the necessity of considering both those dimensions to emphasize what must be eye-catching in a page with a significant boldness (which can be associated with our definition of *text visibility*) and how the hierarchy of ideas must be underlined with varying text character sizes (which is expressed by the *text complexity*).

The expression of text complexity. Our complexity feature is directly correlated to the visual impression of "complexity" we have during the observation. A text made of small letters seams more "complex" than a text with big letters. Our study quantifies this complexity with a measure of entropy. For that purpose, we compute the number of transitions from the background to the text that can be found on horizontal lines. That leads to the estimation of transition probability occurrence on a pixel for each horizontal line. We only keep the maximum probability p in a considered text block because it is representative of how much complex the analyzed text block can be. The texture in the global text block area is called Γ. The entropy $E(\Gamma)$ is then defined for each

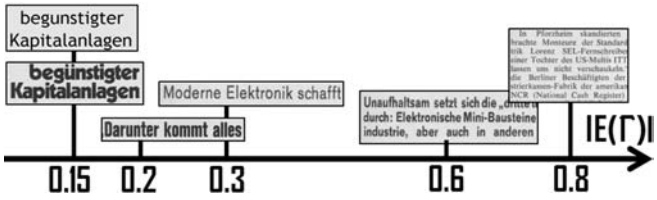


Fig. 10. Entropy scale in a page extracted from the MTDB database

block by the following formula:

$$E(\Gamma) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}. \tag{5}$$

$E(\Gamma)$ always has a positive or null value between extreme normalized values 0 and 1. $E(\Gamma)$ is null if there is no transition between the background and the text, and it is maximal in 1 if p is equal to 1/2. This situation can be encountered when a line is alternately composed with a background pixel and an object pixel. Consequently, the more text is written in small font, the more complex is the curve and, as a result, the higher is the entropy. In the following examples in Fig. 10, we present estimated entropies for different types of text.

The given examples highlight the influence of the size of characters and line spacing. Entropy is a measure of complexity directly influenced by font style and text size. For example, a text with large characters is less complex than a text with small characters. In this example, we also have underlined the miscorrelation that exists between entropy and boldness (see the first examples with $E(\Gamma) = 0.15$). This result illustrates Doermann's hypothesis on significant boldness and hierarchy in a text.

The expression of text visibility. The difference between two characters, one boldface and one lightface, is linked to a perception of visibility. Visibility is the expression of the scriptural stamp that is defined in our method by the width of object segments measured from intersections between multidirectional random lines (called *computation lines*) and the text itself. In Fig. 11, we show an illustration of visibility $V(\Gamma)$ computed in a bold written text block with the following formula:

$$V(\Gamma) = \frac{1}{N_l} . \sum_{j=1}^{N_l} [\frac{1}{Nt_j} \sum_{i=1}^{Nt_j} seg_i] , \tag{6}$$

where N_l is the total number of computation lines used for the estimation of $V(\Gamma)$, Nt_j is the total number of transitions in the j-th computation line, and Seg_i is the width of an object segment (a black transition) as shown in Fig. 11.

In Fig. 12, we propose five samples of texts blocks representative of varying boldness on the same page. For practical purposes, we will normalize this measure by dividing it by the maximal computed value.

The expression of text vertical compactness VCo. We know that the global text structure is essentially charac-

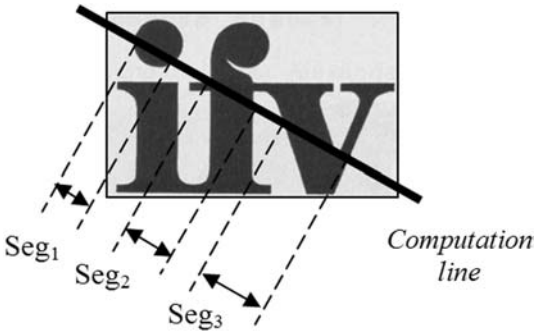


Fig. 11. Visibility computation principle

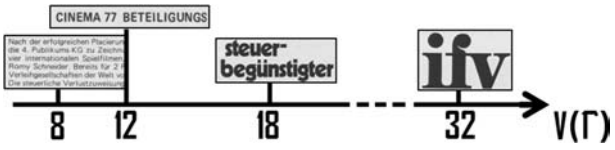


Fig. 12. Visibility scale with text samples on a page of the MTDB database

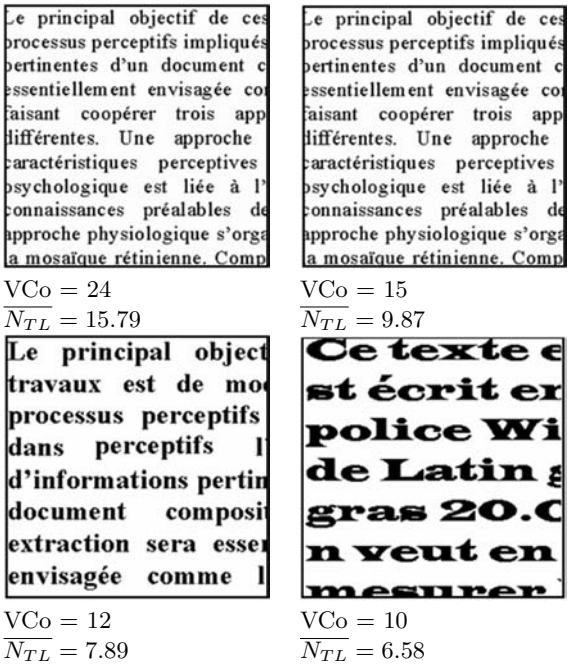


Fig. 13. Examples of VCo and $\overline{N_{TL}}$ values for a set of text samples

terized by two privileged directions: the horizontal and the vertical ones (when skew lines have been detected). The $VCo(\Gamma)$ value is then computed on the basis of vertical computation lines. The VCo feature corresponds to the maximal number of vertical transitions on the height of a block. We do not take into account 1% of the highest values, in case of noise artifacts This approach provides a realistic estimation of the number of lines in the considered block. This number is proportional to the vertical compactness of the entire block. A precise statistical study has shown that the average ratio between the maximal number of vertical transitions and the number of text lines is 1.52. With this principle, the compactness formula is as follows:

$$VCo(\Gamma) = \max_{j\epsilon\{1..width\}}(Nt_j)\,, \tag{7}$$

where Nt_j is the number of transitions in the j-th column. The estimated number of text lines $\overline{N_{TL}}$ is then deduced by the simple relation $\overline{N_{TL}} = VCo/1.52$. Wood [27] and Spitz [22] have proposed a similar approach based on horizontal projections to categorize different scripts. Examples of VCo and $\overline{N_{TL}}$ values are given in Fig. 13.

All these features can be computed at the same time because they are based on the same principle: the use of intersecting lines.

The expression of the relative location of blocks on a page

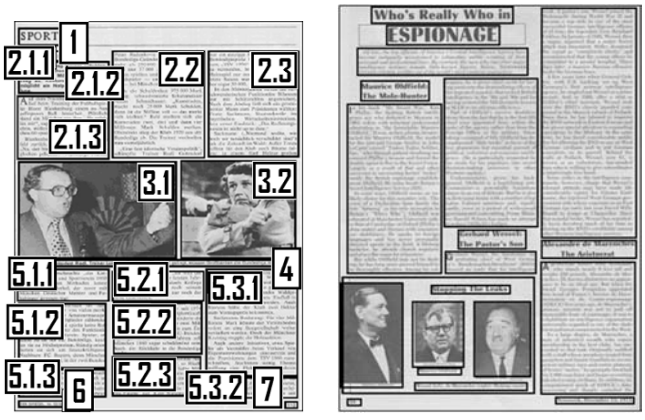
. A text analysis based only on textural features cannot lead to a complete document labeling system without taking into account additional physical information on page organization. For this reason, we propose to introduce geometrical features for each text block corresponding to the *height*, *width*, and *location* on the page. The location model as it is proposed in our work is dependent on the type of document under investigation. We distinguish two categories of pages: the simple linear structured and the complex nonlinear pages (as presented in Fig. 14b).

In this work, the physical location is used to avoid some confusion during the labeling process: the confusion can be linked to the misinterpretation of single text lines (which may be legend figures, headnotes or footnotes, simple isolated lines, titles, or subtitles) and of little text paragraphs (which can also be figure captions, abstracts, or body text paragraphs). In those situations the y-axis is relevant enough to raise the ambiguities. Figure 14 presents the physical segmentation of a document into significant numbered blocks and the block location model based on the description of previous and subsequent block lists (PF-List) according to the y-axis. In our study, we use a simplified tool derived from the XY-tree description when it is suitable, especially for simple document structures (Fig. 14a).

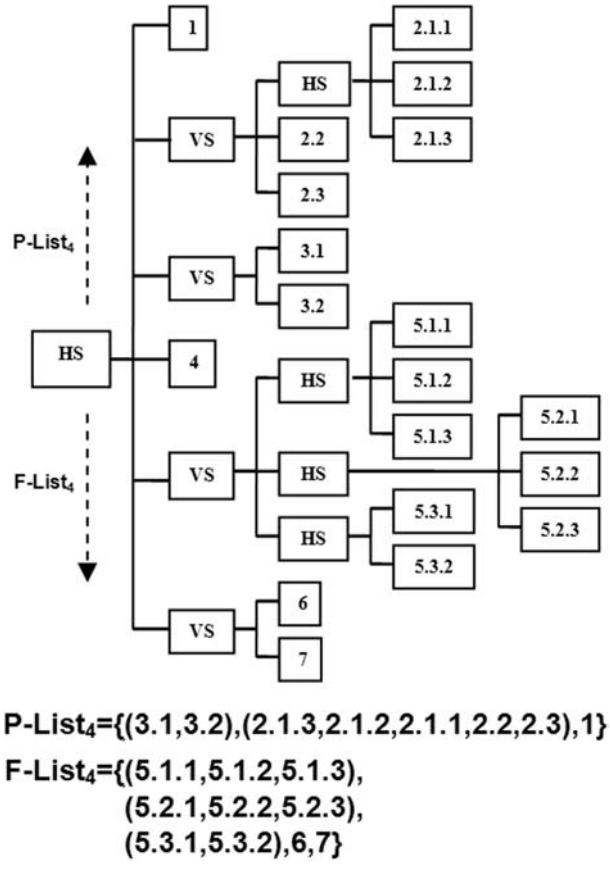
In Fig. 14c we propose the list of previous (resp. subsequent) ordered blocks of block number 4 as $P - List_4$ (resp. as $F - List_4$) and the corresponding XY-tree of the document (Fig. 14a). The two opposite arrows give the sense of the PF-List constitution (from the nearest to the more distant block) that also corresponds to the tree skimming. In more complex pages, blocks are not necessarily vertically and horizontally organized: in those cases, we only keep vertical relations between blocks that give efficient information on block organization (Fig. 14b). The PF-List can be easily completed.

5 Labeling technical description

The functional labeling of a page is based on the exploitation of the 2D space that is obtained with the



a b



c

Fig. 14. Example of (**a**) simple and (**b**) nonlinear document structure. **c** XY-tree decomposition and resulting PF-List for a simple organized page

results of complexity (on the horizontal axis) and visibility (on the vertical axis). It represents the *saliency graph* of the document under investigation. For the labeling process, this graph is then completed by a set of visual features defined in Sect. 4: vertical compactness, physical features, and specific measures that are derived from the saliency graph. These features are also used as transitions for the labeling decision tree process.

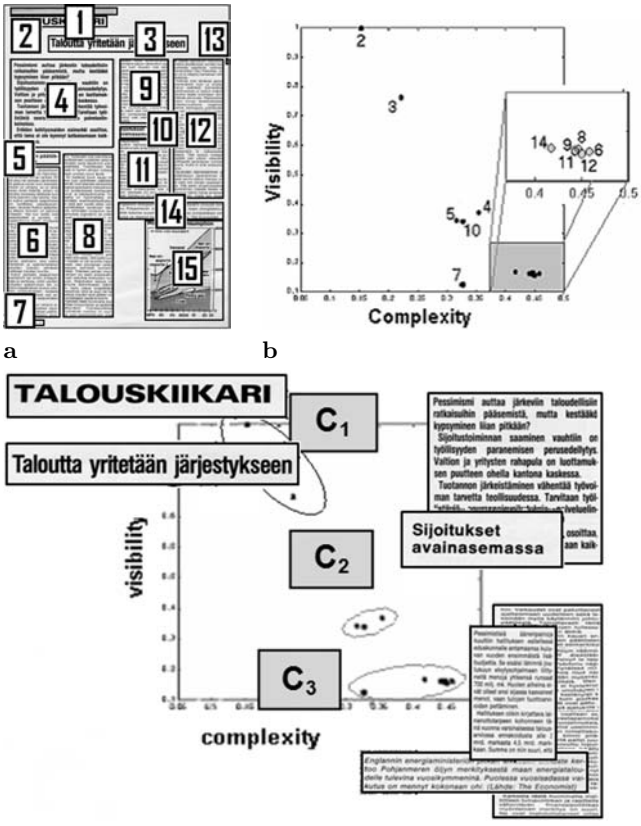
Fig. 15. Illustration of the five clusters with significant text block samples

5.1 Saliency graph and k-means classification

We have chosen to apply a k-means algorithm to detect the pregnant regions of the saliency graph from the location of points that are representative of blocks. The number k of clusters for the k-means classification has been chosen to represent, as best as possible, all possible distributions of points on the page in four corner regions and one central area (Fig. 15). Thus, this value has been fixed at 5. The k-means algorithm is initialized in those five referenced positions that correspond to the four square corners and the central square point, even if there is no point/block in those positions (a more classical use of the k-means algorithm initialized the barycenter positions on already existing points). The convergence of the process leads to a resulting partition of point blocks into at most five clusters. The influence decreases with the eccentricity of the cluster center. In the C_1 cluster, blocks are essentially characterized by low complexity and great visibility. This cluster is representative of *page titles* (with a bold and big typography). In the C_2 cluster, we essentially find section headings, subtitles, and salient text paragraphs (like salient abstracts). In the C_3 cluster, the great complexity is representative of standard text paragraphs. The two other extreme C_4 and C_5 clusters are generally less represented in usual document formats. The C_4 cluster contains all eventual thin main titles (with a thin and large typography), whereas the C_5 cluster contains video inversed texts.

Figure 15 summarizes the *visual* specificity of each cluster with significant text block samples. The k value that we have chosen guarantees a good coherence of results in regard to the great diversity of page layouts and specialized typographical tools.

In a cluster, the blocks have common characteristics but their functional label can be different. In this graph, we have obtained three clusters, C_1, C_2, and C_3. The C_4 and C_5 clusters do not have any representative points on the test page. Figure 16c represents the visual saliency of text blocks in a hierarchy starting with text paragraphs (lower right corner) to main titles (upper left corner). Between these two extremes, we find all blocks belonging to the family of subtitles and salient paragraphs.



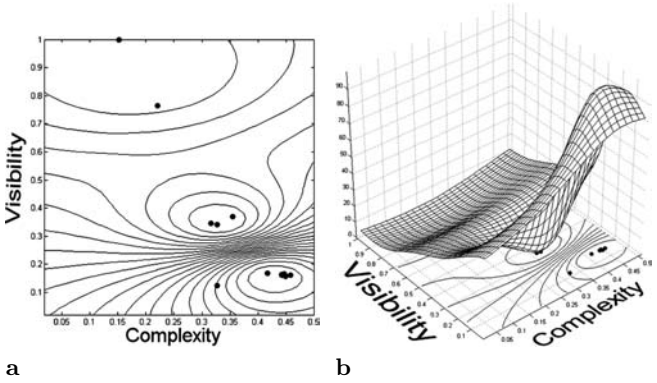
a b

Fig. 16. **a** Composite document of the Oulu MTDB. **b** Saliency graph for the corresponding set of points. **c** K-means cluster decomposition

In Fig. 16, we present the results of the saliency graph that is obtained on a document extracted from the MDTB Oulu database. This page is the *test page* of this paper. Note that blocks 1, 13, and 15 have not been taken into account because they have not been recognized as text blocks in the text block detection step. Block 1 has also been disregarded because it does not contain any text (it is an isolated continuous line).

5.2 Confidence rate

Each cluster contains points that characterize text blocks of the page. Some of these points are near the center of the cluster, others are much further from the center. In practice, the cluster centers are computed as the barycenter of the cluster points (they are inherited from the k-means process). To take this variable distribution into account, we propose to weight each point (each block) with a *confidence* rate that reveals its cluster belonging: a high confidence rate for the points near the center, a much lower one for distant points. This confidence rate will be used in the decision tree process. The closer a point/block P_i is to the barycenter B_k of the cluster C_k, the more we consider that it has been well classified. Conversely, there are many intermediate situations where a point P_i is located on the border between two clusters: in those cases, the initial cluster can be put into doubt and the influence of adjacent clusters must



a b

Fig. 17a,b. Confidence rate representation for Fig. 16 example. **a** Representation by level curves for points in the complexity/visibility plane to belong to cluster C_3. **b** Representation in percent by a surface

Table 1. Confidence rate in percent for some points of Fig. 16 example to belong to each cluster

(%)	6	7	8	9	11	12	14
C_1	0.2	1.5	0.1	0.1	0.1	0.1	0.0
C_2	2.0	17.3	1.1	0.7	0.7	1.1	0.6
C_3	97.8	81.2	98.8	99.2	99.2	98.8	99.4

be taken into account. The confidence rate α_{ik} of the classification of P_i in the cluster C_k is then computed using distances $d_{ij} = dist(P_i, B_j)$ between the points P_i and all the barycenters B_j of existing clusters C_j.

$$\alpha_{ik} = \frac{1}{D} \cdot \frac{1}{(d_{ik} + \varepsilon)^2}\,, \tag{8}$$

with $\alpha_{ik} \epsilon [0,1]$, $\sum_j \alpha_{ij} = 1$,

$$D = \sum_j \frac{1}{(d_{ij} + \varepsilon)^2}$$

, and $d_{ij} = dist(P_i, B_j)$.

In Eq. 8, ε is a constant value arbitrarily small used to avoid computing problems of division by zero. If the point P_i is superposed to B_k, the distance d_{ik} is null and the confidence rate is equal to 1 (or 100% if expressed in percent). Figure 17 shows the evolution of the confidence rate for the test page of Fig. 16. We present here the confidence rate as belonging to the C_3 cluster. Table 1 gives the values (in percent) for some points/blocks of Fig. 16. The classes C_4 and C_5 are not mentioned because no block belongs to them.

The confidence rate is the starting point of the complete labeling process: for each block, the functional label is expressed as a *specialization* of the cluster for which the block has the maximal confidence rate. When the specialization with the higher rate is unsuccessful, a new specialization begins in the cluster corresponding to the second best confidence rate. The process is repeated until the convergence to a specialization or sometimes to a reject. The following section presents the complete method.

5.3 Labeling decision tree LDT

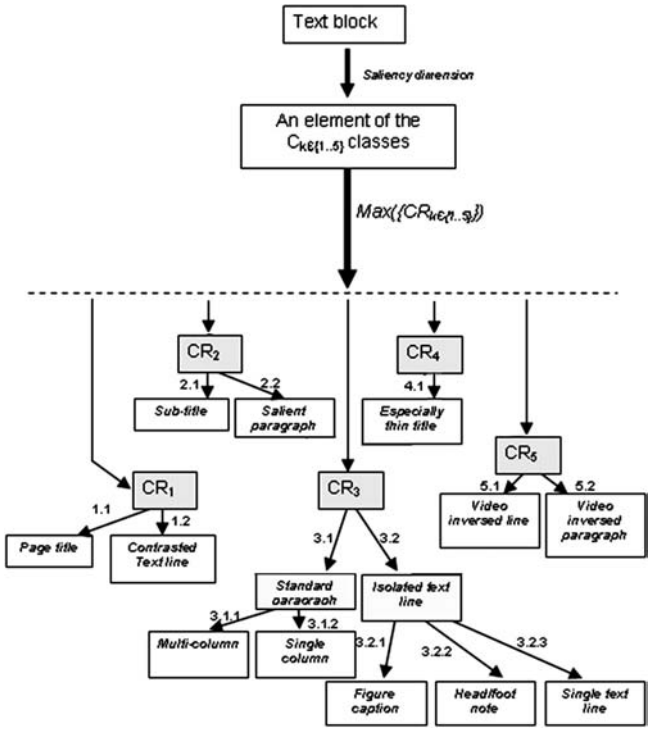
LDT formal specification. The labeling process is based on a knowledge representation model described by a decision tree: it starts from an initial root that is the text block followed by a first link of saliency dimension *(complexity, visibility)*. From the following node corresponding to the class the block belongs to (one of the five visual classes $C_{ii\epsilon\{1..5\}}$ defined in the k-means section), a set of five possible *nodes* can be reached according to the confidence rate CR computed for each class. For each block, we order the confidence rates from the best to the lowest and skim the branch corresponding to the higher rate. Each node is then followed by conditional *specialization* links that lead to label propositions. These specialization links are based on feature combination including the vertical compactness and physical primitives. The decision tree is described in Fig. 18 and the combination features are numbered just above.

When the tree skimming does not lead to any label with the first best confidence rate, we consider the second best rate only if this rate is more than 50% of the initial best confidence rate (this value has been experimentally calculated on the test base). We then test the conditional links corresponding to the second best cluster. If the second rate is not enough, the block is rejected. When the process is unsuccessful after the second confidence rate, we also reject the current block and consider that it cannot be labeled with the proposed method. This situation can be encountered for too small blocks (whose area is inferior to 0.1% of the total image area) and for horizontally oriented images or graphics that have been initially classified as text blocks.

LDT evaluation and stable threshold definition. The considered links are the followings: saliency dimension (*complexity, visibility*), $Max(CR_{ik})_{k\epsilon\{1..5\}}$ correspond to the maximal rate of the ordered list, VCo is the vertical compactness, P corresponds to the list of previous blocks in the page (the P-List), and F is the list for the subsequent blocks (the F-List), W is the block width, and A is the block area. We have also defined some thresholds for conditional links: T_{min} is the maximal VCo of a page title (this value is proportional to the maximum number of lines accepted in a title block and is fixed at 3), W_{max} is the middle width of the analyzed entire page, and A_{min} is the minimum required block area that corresponds to 10% of the total average text block areas on the considered page.

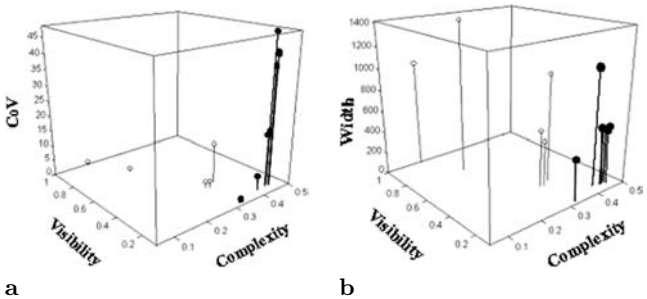
In the decision tree, the possibility of rejection is proposed when the block does not have the required characteristics for its specialization in any of the two best considered classes or when the block area is inferior to the threshold A_{min}.

At the end of the decision tree skimming, we obtain for each block a functional label (or a nonclassification result when the block is rejected). The decision tree can also be visually interpreted with multidimensional feature spaces by considering the saliency graph as the basis of these spaces (Fig. 19).



1.1 $VCo \leq T_{min}$ and P=NIL
1.2 $VCo \leq T_{min}$ and (P or F-List)>1
2.1 $VCo \leq T_{min}$ and P≠Image and P≠NIL and F≠NIL
2.2 $VCo > T_{min}$ and P≠NIL
3.1 $VCo > T_{min}$
  3.1.1 $W \leq \frac{1}{2} W_{max}$
  3.1.2 $W > \frac{1}{2} W_{max}$
3.2 $VCo \leq T_{min}$
  3.2.1 P=Image
  3.2.2 (P or F-List)=NIL
  3.2.3 P≠Image and (P or F-List)≠NIL
4.1 $VCo \leq T_{min}$
5.1 $VCo \leq T_{min}$
5.2 $VCo > T_{min}$

Fig. 18. Principle of functional labeling based on a decision tree



a **b**

Fig. 19a,b. Projection of features in 3D graphs for functional labeling. **a** Illustration with VCo, and **b** W as third dimension

In Figs. 19a and b, we have represented two 3D graphs that are visual representations of block specialization in multicolumn or single-column paragraphs belonging to the C_3 class represented in Fig. 16c. The measures have been computed on all blocks of the test page, but the only ones that are used for the labeling are represented in bold lines in Fig. 19a and b. Note that all points of the C_3 class (except blocks 7 and 14) have a common width that corresponds to the column width. All compactness values (VCo) are high and represent the global number of lines for each paragraph. In this process, the results are not influenced by the order in which blocks are considered. Also note that the proposed thresholds in the decision tree are not dependent on the kind of documents under investigation: the test bases propose a great panel of documents that can be processed with the same approach without changing any threshold value. What is more, our approach is based on a *relativity* notion between blocks: it allows characterizing blocks in regard to all other blocks of the page. The resulting labels express the *relative* hierarchy between textual components.

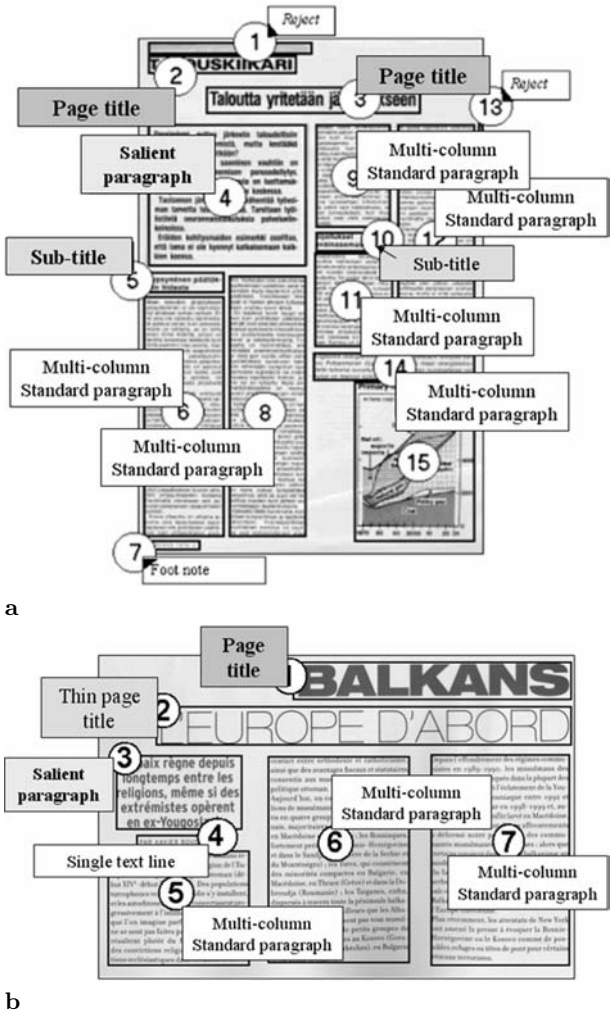


Fig. 20. a Functional labeling results on the test page. **b** Results from our personal corpus

6 Results, discussion, and prospective work

6.1 Labeling results

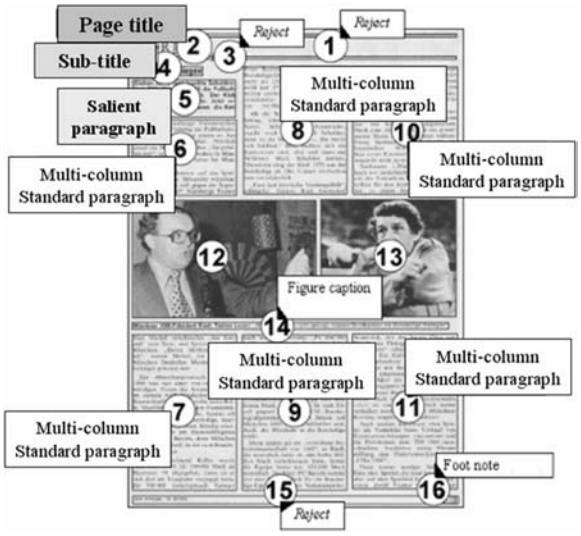
Examples extracted on the test corpus. The system leads to results that are illustrated in six examples that have been extracted from the same newspaper of the MTDB database and from our test base (Figs. 20 and 21). Figure 20a corresponds to the test page. In Fig. 21a, blocks 1 and 3 were rejected during the text block selection step developed in Sect. 4.2. Those blocks are not text blocks, but they contain plenty of continuous separation lines.

Block 15 was also rejected before the decision tree process because it had not been segmented like other homogeneous text blocks on the page: the footnote is surrounded by a large bounding box that recovers the whole page width, so it contains a small line of text and a wide background area.

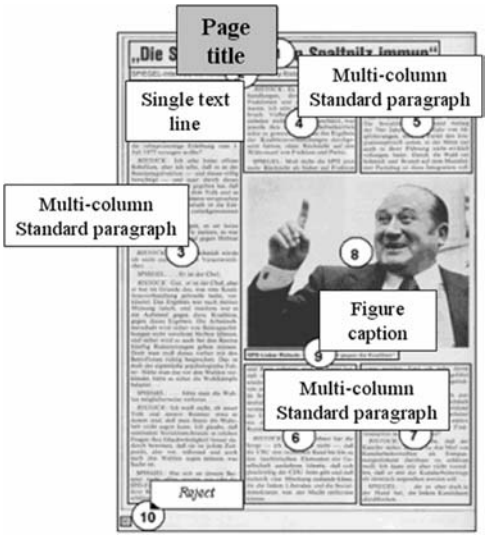
In Fig. 21b, the real ground truth subtitle of the page (block 2) has been labeled "*single text line*" because the visibility of the text is weak compared to the main title of the page. Block 10 has been rejected because the block area is not efficient to compute the complexity and visibility measures. In Fig. 21c, there are two rejects that correspond to a nontextual block (block 12) and a nonhomogeneous text block with a large background area (block 10). The results obtained in the Oulu database are qualitatively similar to those obtained in our personal corpus. The labeling results can be compared to the ground truth labels that are proposed as references in the database. In our corpus, we have applied the same approach with the same referenced labels.

Results analysis and method accuracy. The analysis of the MTDB database and our corpus (both are called test base) leads to the following results, which are reported in Table 2. Table 2 shows the categorization and labeling accuracy of our approach.

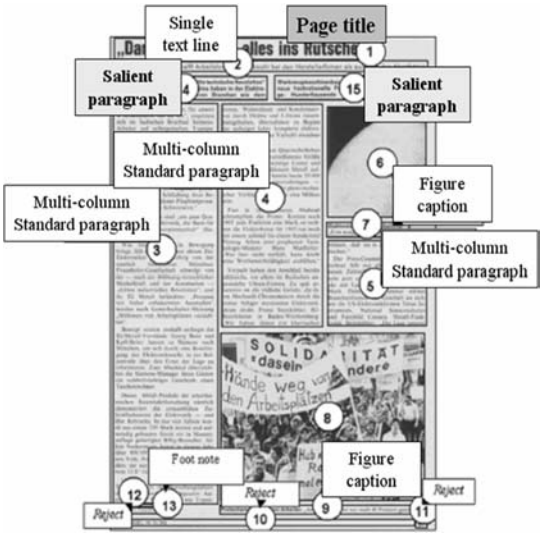
This table must be understood as follows: the diagonal bold values correspond to the real accuracy of the k-means clustering, whereas the horizontal last line values give the real labeling accuracy that is obtained on the basis of the previous results, which is why those last results are very high. The k-means results are not homogeneous for all block types: there is a notable difference between the rates of correct categorization in the different C_i classes. These differences are linked to the page visual presentation. The categorization in the C_2 class is 92.4% correct: this low value is linked to the category of analyzed pages where there are no main titles but only subtitles or body paragraphs (Fig. 22a). In those situations, the hierarchy of visual text elements is different and is translated in the sense where the subtitles are considered as titles not represented on the page. In the same manner, the categorization in the C_3 class is 95.3% correct: the relative great boldness of some text paragraphs leads the analysis to consider them as salient paragraphs (like salient abstracts), whereas they are sim-



a



b



c

Fig. 21a–c. Typical examples of page labeling in newspaper pages extracted from the MTDB database

Table 2. Statistical results of functional labeling on the test database

		Ground truth distribution				
		C_1	C_2	C_3	C_4	C_5
Categorization	C_1	**97.2**	3.8	0.1	4.8	2
after	C_2	1.2	**92.4**	2.6	3.5	2.6
k-means	C_3	0.1	2.5	**95.3**	1.4	2.2
step in	C_4	0.8	0.8	0.6	**90.2**	0.2
class:	C_5	0.7	0.5	1.4	0.1	**94.0**
Final well-labeled blocks (%) among the well-categorized blocks		**98.2**	**97.4**	**96.8**	**98.2**	**97.4**

ple body paragraphs. Conversely, a low relative boldness of a real salient abstract will lead to an erroneous categorization in the C_3 class. The categorization in the C_4 class is only 90.2% correct: this result is linked to the rare situations where a *thin title* is obtained in standard documents. When this situation is encountered, the title is sometimes categorized in the C_3 class.

The final labeling results (the last line of the table) are high because there are only a few situations where an error can be made once the block is correctly categorized in one of the five classes. The definitive labeling accuracy corresponds to the combination between the class categorization rate and the correct labeling percentage. The table does not show the relationship that exists between the number of blocks in the page and labeling accuracy. In fact, there is an increasing error rate that is proportional to the increasing number of blocks contained in a document. Two main parameters influence this phenomenon: the number and the size of blocks on the page. Documents with complex structures very often contain numerous blocks of varying area. In small blocks (like short paragraphs of text or single lines), statistical results are no longer relevant, and the resulting labels are inappropriate because small blocks contain few characters that are not efficient for a statistical analysis. This situation can be encountered in documents containing more than 40 blocks; this situation is rare. In the opposite case, when there are less than ten blocks on a page, our approach becomes less relevant because the determination of the functional classes cannot be based on a too small number of blocks. In this case, we have chosen to analyze several pages together, i.e., we built a unique saliency graph for different pages corresponding to the same journals or magazines. Finally, the best results are obtained for an intermediate category of pages containing less than 40 blocks and more than 10 blocks a page, which corresponds to the majority of pages in our test corpus. In Table 2, we present the average results of our method.

6.2 Limits of the approach

Figure 22 provides different relevant and typical examples of mislabeling linked to occasionally unexpected

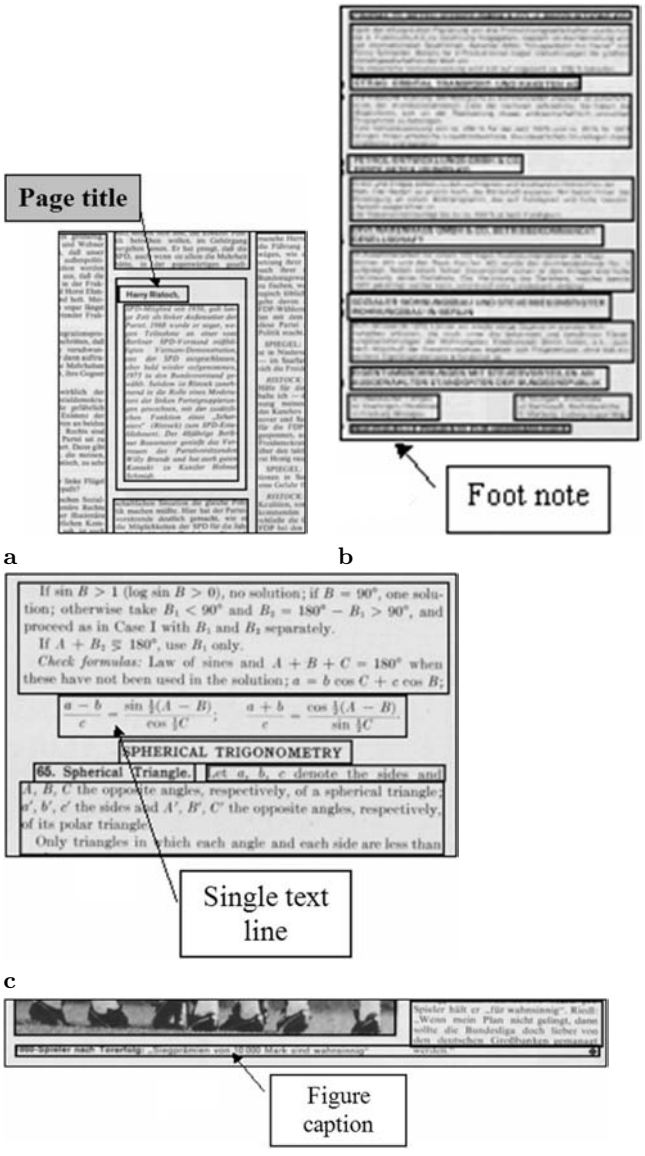


Fig. 22a–d. Typical errors produced by the system

page layouts. The rare errors produced by this labeling system invariably involve unusual document layouts and were found in the following examples: the system proposes the label *page title* for the single bold line in the middle of the page even though it is a contrasted line that presents the author paper (Fig. 22a); in the same way, a title at the bottom of the page with low visibility is labeled footnotes (Fig. 22b).

The first error is due to the great visibility of the block compared with all surrounding text paragraphs. By including a coherence analysis based on the visibility dynamic in the whole page, we are able to avoid such errors. The remaining errors are generally caused by the presence of unusual blocks like formulas in mathematics documents that are not considered in our approach (22c) and by text block shapes linked to the initial segmentation (22d). This last case has been encountered in the MTDB database where blocks having different functional meanings are merged into a single block: in Fig. 22d the last block of the page contains a figure legend, author name, and page number.

6.3 Comparative approach

Region classification and text block labeling have been addressed by other authors with different methods based on an accurate parameterization of document types (Sect. 1). For example, in [21] the authors propose to build a decision tree classifier on the basis of feature vectors and local measures. Their subsequent works, developed in [25], also show that they need a complete training set of feature vectors with true class labels. In our work, the labeling is only based on some assumptions corresponding to the visual hierarchy of text elements on pages, but no precision on typographical features is used. In their work, the authors also used discriminant thresholds to specialize the description of blocks that are computed on the basis of the training set. In our proposition, the thresholds are independent of the kind of documents under investigation: the only hypotheses correspond to the page stability (Sect. 3.1), and no local measures are necessary to determine the functional label of each text block. In comparison with this approach, we do not need any training set to build the decision tree: we only use knowledge about the physical hierarchy of text block entities (that knowledge is gathered in the $\{F_i\}_{i\epsilon\{1..3\}}$ functional family description). We can also mention the work of Liang, who proposes a document zone classification approach by using local sizes of connected components [14]. In [13], the authors have developed an automated labeling system by using generic typesetting knowledge of English text. All those methods suppose a local analysis of text zones and an accurate a priori knowledge about the kinds of documents under investigation. This is not the case with our method. A texture-based work has been proposed by Zhu [28] and is conceptually closer to our labeling approach, but the authors break any visual hierarchy of text components by normalizing all zones and by creating uniform text block in sizes and spaces. Finally, we also note that, in contrast to all the works mentioned above, our labeling system can process several pages of the same document (journal, newspapers, proceedings) in the same process step because the functional hierarchy of text components is preserved.

7 Conclusion

This work is part of a complete project dedicated to printed document structuring where information is retrieved according to its visual saliency, i.e., its perceptual attraction power over the reader's eye. The purpose is to propose a visual and functional labeling of text zones of composite documents having a well-defined and reproducible structure. The visual features that are used to characterize text zones of pages are the complexity, compactness, visibility, and some physical primitives. They

are valuable because they correspond to a reality of visual perception by expressing the visual hierarchy of text zones and their functional properties. By reflecting what attracts the eye in a document, these nonredundant and complementary primitives allow a quick classification of font styles. The final labeling reflects these complementarities. The development of textural primitives is a low-level process, very close to the roots of visual perception, and a generic way to establish a visual and functional hierarchy among all text blocks on one page. This work is a first step toward the text identification that could be associated with a semantic approach. The accuracy of the method is very promising with an average performance of 96% correct labeling.

References

1. Bergler S, Suen CY, Nadal C, Nobile N, Waked B, Bloch A Logical block labeling for diverse types of document images. In: Proceedings of the conference on document layout interpretation and its applications, pp 231–235
2. Bres S (1994) Contributions à la quantification des critères de transparence et d'anisoptropie par une approche globale. Application au contrôle de qualité de matériaux composites. PhD thesis: INSA de Lyon
3. Bruce V, Green PR (193) Visual perception: Physiology, psychology and ecology. Presse universitaire de Grenoble, Grenoble, France
4. Chetverikov D, Liang J, Komuves J, Haralick RM (1996) Zone classification using texture features. In: Proceedings of the 13th international conference on pattern recognition, 3:676–680
5. Doermann D, Rosenfeld A, Rivlin E (1997) The function of documents. In: Proceedings of the 4th international conference on document analysis and recognition, Ulm, Germany, 2:1077–1081
6. Eglin V (1998) Contribution à la structuration fonctionnelle des documents. PhD thesis, INSA de Lyon
7. Eglin V, Bres S, Emptoz H (1998) Printed text featuring using visual criteria of legibility and complexity. In: Proceedings of the 14th international conference on pattern recognition, Brisbane, Australia, August 1998, pp 942–944
8. Jain AK, Zhong Y (1996) Page segmentation using texture analysis. Pattern Recog 29(5):743–770
9. Jain AK, Yu B (1997) Page segmentation using document models. In: Proceedings of the 4th international conference on document analysis and recognition, 1:34–39
10. Jain AK, Bhattacharjee S (1992) Text segmentation using Gabor filters for automatic document processing. Mach Vision Appl 5(3):169–184
11. Julesz B, Bergen JR (1983) Textons, the fundamental elements in preattentive vision and the perception of textures. Bell Sys Tech J 62(6):1619–1645
12. Jung MC, Shin YC, Srihari SN (1999) Multifont classification using typographical attributes. In: Proceedings of the 3rd international conference on document analysis and recognition, pp 353–356
13. Le DX, Kim J, Pearson G, Thom GR (1999) Automated labeling of zones from scanned documents. In: Proceedings of SDIUT'99, pp 219–226
14. Liang J, Haralick R, Phillips I (1996) Document zone classification using sizes of connected components. In: Proceedings of Document Recognition III, SPIE 96, pp 150–157
15. Maderlechner G, Schreyer A, Suda P (1999) Information extraction from document images using attention based layout segmentation. In: Proceedings of the conference on document layout interpretation and its applications, pp 216–219
16. Maderlechner G, Suda P, Brucker T (1997) Classification of documents by form and content. Patt Recog Lett 18:1225–1231
17. Marr D (1982) Vision. Freeman, San Francisco
18. Nagy, G.: Twenty years of Document Image Analysis in PAMI. IEEE Trans Patt Anal Mach Intell 22(1):38–62
19. Randen T, Husoy H (1994) Segmentation of text/image documents using texture approaches. In: Proceedings of NOBIM, pp 60–67
20. Schreyer A, Maderlechner G, Suda P (1998) Font style detection using textons. In: Proceedings of Document Analysis System, pp 99–108
21. Sivaramakrishnan R, Phillips I, Ha J, Subramanium S, Haralick R (1995) Zone classification in a document using the method of feature vector generation. In: Proceedings of the 3rd international conference on document analysis and recognition, pp 541–544
22. Spitz AL (1997) Determination of the script and language content of document images. IEEE Trans Patt Anal Mach Intell 3(19):235–245
23. Strouthopoulos C, Papamarkos N (1998) Text identification for document image analysis using a neural network. Image Vision Comput 16:879–896
24. Suen CY, Bergler S, Nobile N, Waked B, Nadal CP, Bloch A (1998) Categorizing document images into script and language classes. In: Proceedings of the international conference on advances in pattern recognition, pp 297–306
25. Wang Y, Phillips IT, Haralick RM (2002) A method for document zone content classification. In: Proceedings of the international conference on pattern recognition, 3:196–199
26. Wong FWK, Casey R (1982) Block segmentation and text extraction in mixed text/image documents. Comput Graph Image Process 20:375–390
27. Wood S, Yao X, Krishnamurthi K, Dang L (1995) Language identification for printed text independent of segmentation. In: Proceedings of the international conference on image processing, pp 428–431
28. Zhu Y, Tan T, Wang Y (1999) Font recognition based on global texture analysis. In: Proceedings of the 5th international conference on document analysis and recognition, pp 349–352

Véronique Eglin has been assistant professor and researcher on the Pattern Recognition and Vision team in the LIRIS Laboratory since 1998 at the National Institute of Applied Sciences in Lyon (INSA). She is working on document segmentation and analysis by developing methods based on visual perception and multiresolution for information retrieval and characterization.

Stéphane Bres has been assistant professor in the Computer Science Department of the National Institute of Applied Sciences of Lyon (INSA, France) since 1995 and teaches signal processing, numerical analysis, and computer vision. He belongs to the LIRIS Lab of the Pattern Recognition and Vision team (RFV). He has research activities in the field of computer vision and in particular in automatic image and document indexing.