

RASADE : Automatic Recognition of structured document using typography and spatial inference

LEBOURGEOIS Frank, SOUAFI-BENSAFI souad
Equipe de Reconnaissance de Formes et Vision
I.N.S.A. de LYON - Bât 403
20 Avenue A. Einstein 69621 Villeurbanne Cedex FRANCE
Tél : (+33) 72 43 80 93 Fax : (+33) 72 43 80 97
E-mail : flebourg@rfv.insa-lyon.fr

Abstract

This paper describes the overall scheme and features of an industrial project which aim to index automatically books and magazines by reading their content tables. After a short presentation of this project and its objectives, the paper describes our general approach, the model we currently use. Then we describe the physical and typographical layout extraction and logical layout retrieval.

1. Presentation

1.1 RASADE project

The goal of RASADE* project (Automatic Recognition of Structured Documents) consist to index automatically a large variety of printed documents especially periodicals, such as newspapers, magazines, and journals, in a database by reading their contents tables. We currently meet two problems :

- The variability of their logical layout from a document to another or even periodically for the same journal,
- The complexity of the physical layout because of great variety of fonts of characters print on various background in many colors. Content tables try to be more and more attractive by using pictures, color backgrounds and complex layout.

Logical items we extract from content tables are the name of magazines, and for each subject, titles, authors, summaries and references. At the end of the project, the software will become a component of an electronic management system of documents already installed in several libraries and documentation centers.

1.2 Proposition

Logical contents are visually recognizable by a specific typography (character sizes, styles, colors) organized in a same manner. As position of logical items in the document vary from a document to another, we base our system on typography as main features combined with information about document organization. In a future version, we may use other information like frames, images location, but we meet

some difficulties to differentiate pictures from color frames where text are printed. On several images, we notice that typography is stable for periodicals because the same journal keep a common style and change characters typography very rarely. Compare to logical layout analysis based only on the text content obtained by OCR [9], our proposition is different but can be considered as a complementary approach.

We propose to detect the recurrence of physical features combining information about text blocks location and typography. To process a wide variety of documents, we use a supervised recognition based on a logical layout model built for each journal. The training process is reduced to the labeling of recurrent text zones which are organized in the same manner. Text blocks chains are defined by both spatial information (blocks location compare to other blocks) and typography (colors, fonts, character style, character spaces, spaces between lines..). This model provides enough information to retrieve the document contents (titles, authors, summary, references). The training process asks the user additional information like marks location to enable the automatic recognition of the journal in order to apply directly the appropriate model. The application area of this project is not limited and requires to process images in gray level to retrieve information about colors which is also important for the layout recognition.

2. The Model of logical layout

For each periodical we separate information about the journal which authenticate the document origin from information about its logical layout. General information about the journal help to find automatically the right model to apply and give precious information about scan area, number of page to scan. For journal identification retrieval, we use simple robust features like the position and picture of word "contents" and the specific logo or journal name. We use also the journal caption which must legally be printed on the same page. We store the localization of the main table of contents in order to save the computation by reducing the typography analysis to this area only. Our model take into account the number of page required by the table contents since periodicals use one or two pages for the presentation. Model of the logical layout is based on text blocs obtain by grouping words having the same logical

label. Each block contain two categories of information : physical information found by a bottom-up analysis from the image document to a physical and typographical layout, and logical information computed iteratively by the system. The physical and typographical layout (*local fixed features*) is computed only once by a bottom-up approach, and these information do not change during the recognition of the logical layout. In the opposite information about the logical layout (*external variable features*) change during the final recognition to retrieve the logical content according to logical blocks organization. The model is represented by following record :

□ **Information about the journal for its recognition**

- **Document size (scan area, width and height)**
- **Position and pictures of word “Content” and logo or marks defined by the user**
- **Position of journal name**
- **Journal captions (ISSN, number, volume, date)**

□ **Information about the logical layout for the indexation of its content**

- **Number of page containing the contents table**
- **Position on each page of the main content table**
- **Prototype of binary Patterns of characters**
- **Physical information (*local fixed features*)**
 - ◆ *Block Position*
in absolute within the document and relatively to text line, paragraph, and column.
 - ◆ *Typographical family recognized by the system*
 - ◆ *Typographical information extracted*
 - ⇒ Font of characters, size, style (bold, italic..), color of character and background.
 - ⇒ distances between characters, words, lines
 - ⇒ Justification fixed or variable, alignment.
- **Logical information (*external variable features*)**
 - ◆ *Temporary decision of the block logical content*
 - ◆ *Confidence rate of this logical label computed*
 - ◆ *Assumptions arrays containing relevance score for each possible logical content*
 - ◆ *Information about logical content of blocks found in the neighborhood*
 - ⇒ Logical content of left, right, up, down blocks
 - ◆ *eventually pointers to adjacent blocks making a recurrent chain list of blocks.*

We can deduce from this model a lot of hidden information like spaces between lines and blocks, average and standard deviation of text block position, text density in each block or column, relative importance of text for each areas. Moreover, the system deduces the relative variability of feature localization for each journal. This information is very important, because the recognition stage takes this variability into account during contents retrieval. But the model assumes that a logical block contain only one typography and we cannot process document which have several typographies related to a same logical block. As we never meet such cases, this assumption requires a correct recognition of

typography. In the opposite, we can process document which use the same typography to describe several logical contents. The system builds automatically spatial relations between typographical regions with a nearest neighbor rule (right, left, bottom, up neighbors) and tries to detect a logical recurrence between typographical regions. If the document does not use different typography for different contents objects, then we only use spatial relations between text blocks and eventually characters delimiters which are commonly printed to separate information [1].

3. Realization

We recall briefly works already achieved and published for the segmentation of the physical layout in gray-level document images. Since our approach is mainly based on typographical information, we need to take care about typography recognition. As OFR (Optical Font Recognition) still remain a serious problem [8] that we do not want to solve, we avoid this problem by introducing a robust method to both extract typography and match these information for other documents. The typographical layout is extracted by analyzing all different patterns of characters within the document. We extract typographic families of characters by grouping recursively words containing characters having same pattern. This robust method does not provide fonts names but recognizes all different fonts, style and color of characters within a document without substitution.

3.1 Physical Layout extraction

Journals and magazines are frequently printed in color. Information about colors can be used to recognize the document's logical layout. For example, an identical color indicates that characters belong to the same class of semantic contents (page number, author, title of article, summary, separation). From a large variety of documents printed in color, we have highlighted that analyzing image in gray level is sufficient to retrieve information about colors used for frames, backgrounds and characters and reduces the computational effort and storage. We have already developed a robust and simple method which localizes text lines without background and lighting constraints [7][4]. We keep gray-level information of both background and character color for the analysis of document layout. We extract directly from gray level images text lines and simultaneously separate text from graphics. But this process can make some false detection of text in difficult graphics areas. Graphic regions considered as text lines are eliminated by studying the coherence of connected components alignment within a text line. A graphic is a region which contains more randomly distributed connected components compared to well ordered connected components for text lines.



Figure 1: Physical Layout extraction by segmentation of text lines, binarization and bottom-up growing process

3.2 Typographical layout extraction

Typographical information helps readers to find the document's logical layout and increases readability. Generally, authors use typographical information with a minimum of coherence. Font, style, size and color of characters contribute different information. The size of characters is related to text hierarchy (title, subtitle, regular text, footnotes, etc.), bold style emphasizes text importance, italics are used to make a separation, and color provide additional information which generally helps readers to find the logical structure. We have already developed a recognition of typographical families of characters by using printed character pattern redundancy [6]. Character pattern redundancy computation mainly uses pattern matching to compare similarity between two binary shapes [2]. In order to increase comparison speed, we have organized partial tests in a decision tree using template matching [3]. The introduction of new character patterns decreases progressively after several readings of documents of the same type [5]. It can be noted that redundant binary shapes can refer to characters printed in different colors on various backgrounds independently of character position within the document. If we assume that all words contain characters from the same font and style, and if each different character printed with a particular font provide a unique binary pattern, then we can establish a relation between words which share at least one character having the same binary pattern. This relation makes a new set of words printed with the same style and font. But our method gave more typographical families than were actually found. Some isolated families cannot be classified in other families using our approach alone. It can also be noted that numerals make up a particular family, provided the text does not contain numerals and letters associated in a word. If families are sorted by decreasing order of corresponding words, we automatically find the main typography as well as minor

typographies used occasionally for a precise logical meaning. Very Small typographical families represent words which occur very rarely in the text (text in graphics, titles, authors' names, etc.); thus another process is needed to classify them within the other main families. Finally, the system merge consecutive words of same typography into typographical regions. All pattern of characters are stored in the model and are used to match typographical families in other documents from the same origin. We project to reduce the number of typographical families by using information store in the model about typographical information so as to preserve a coherence between record *Typographical family* and record *Typographical information*.

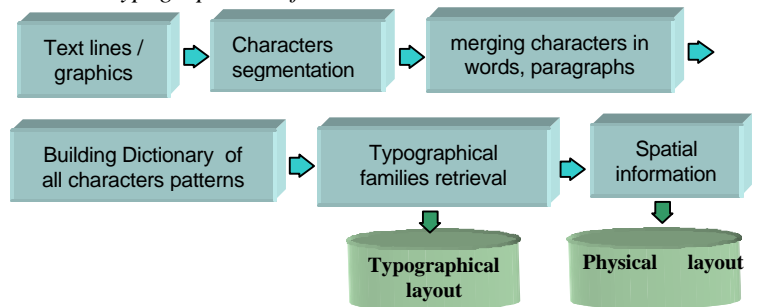


Figure 2: physical and typographical segmentation scheme

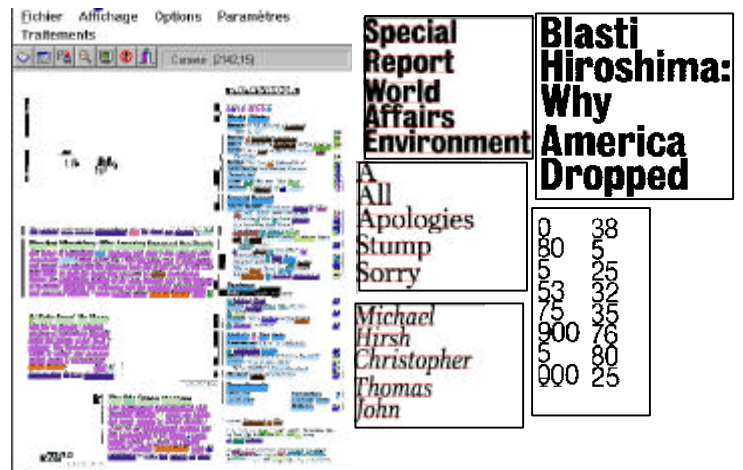


Figure 3: typographical layout and typographical families

3.3 logical layout recognition

For a unknown document, the system tries to recognize the name of the journal by matching specific marks previously defined in the model. We use pattern matching of gray levels picture of logos or word "contents" by direct correlation computed on a reduced size of document image. If no match is found, the system ask the user the right name. This step can be aborted by user who can give directly the name of the document. Then the recognition is achieved in three different steps. In a first step, we match position of frames, columns and text blocks. We use the variability of these blocks to achieve this task. For fix blocks, we find all text lines which can belong to these blocks. For more variable blocks, we build new blocks from text lines found in the new document which have the best match with the model. This step provide the localization of the main region of interest bounding the table of contents. During

a second step, we automatically match all binary patterns of characters found in the contents table, with prototypes stored in the model. As each pattern of character from the dictionary store in the model is linked to its typographical family, we deduce immediately the correspondence between typographical families found in the model and the unknown document. This operation is necessary to get compatible labels of typographical families between the model and the input document. If too many characters patterns are found different, then we exit the recognition process and ask for a model updating. This process give a structural and typographical layout which is compatible with the model. The results of this first analysis is stored in a temporary model associated with the new document. We save the localization of the main content table, binary patterns of characters found in the new document, information about typography, labels of typographical families compatible with the model. All these information are stored in the physical part of the new model which never change during the next step.

In a last step, we use both spatial relation between blocks and their logical content computed iteratively by using external features of the new model. At the beginning, we consider all words as a single block. According to the physical and typographical layout, we initialize the assumptions arrays with a relevance score computed for each possible logical content. For each iteration, we label temporarily each block with the logical class given by the maximum relevance score in the assumption array. Iteratively ambiguous blocks modify their logical classification according to logical content of adjacent blocks. During this step, adjacent blocks which are already classified with a high relevance score are merged in a single block. The iterative process stop when no modification have been found or if all blocks are classified with a high relevance score. For ambiguous blocks which are not completely classified, we ask the user to manually label them. The recognition stage show the main region of interest and logical blocks which have a match with semantic object of the model. Contents of each logical block are read by OCR and index in an object database.

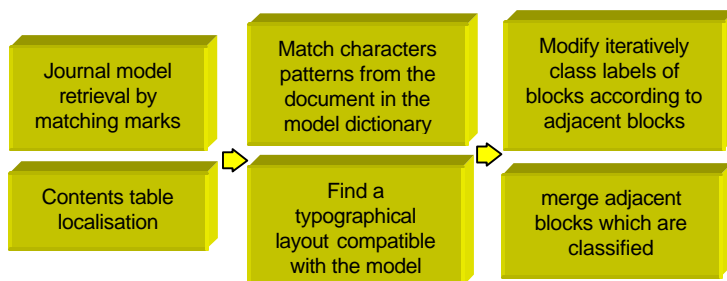


Figure 4: logical segmentation scheme in three steps

During the model updating, add in the model the new alternative positions blocks and update the dictionary of binary patterns of characters, the typographical labels of new families.

*RASADE project is granted by EVER company www.ever.fr and also supported by ANVAR www.anvar.fr



Figure 6: logical labeling using typography and adjacency relation

4. Conclusion and perspective

Our method can be applied to every type of document, especially complex and typographically rich documents. We aim to reduce the complexity of the training and simplify the operation of labeling by improving the ergonomic of the software. The processing time for the physical and typographical layout segmentation and for the logical layout recognition is inferior to one minute per document. For instance we have tested our method only on few different journals selected for their logical layout complexity. We are going to experiment this method on a real document database of hundreds of journals to measure the real gain of productivity. We will evaluate the time lost by users for correction compare to a direct manual indexation of contents. This study must objectively answer if such system is really interesting to increase productivity of documents indexation.

5. References

- [1] Baird H.S., Bunke H., Yamamoto K. "Structured document image analysis". Springer Verlag 1991, 582 P.
- [2] Wahl F, Casey G, Wong K. "Block segmentation and text extraction in mixed text/image documents". Computer graphics image processing, 1982, n°20 p375-390
- [3] Lebourgeois F., Henry J.L. "An OCR System for Printed Documents". The Proceedings of the IAPR Workshop on MVA, Tokyo, December 7-9,1992.p.83-86.
- [4] LeBourgeois F. "Robust multifold OCR system from gray level images" fourth ICDAR, International Conference on Document Analysis and Recognition, Ulm, 1997, p. 1-5.
- [5] LeBourgeois F., Henry J.L. "A Contextual Processing for an OCR System, Based on Pattern Learning". Proceedings of the 2nd International Conference on Document Analysis, Tsukuba Science City, October 20-22, 1993. p. 862-865.
- [6] Duffy L., Lebourgeois F., Emptoz H. "the improve of logical structure analysis by typographic characteristics extraction". Proceedings of 9th ICIAP'97, Firenze, September 17-19, 1997. p. 639-646.
- [7] LeBourgeois F., Emptoz H. "segmentation des documents composites en niveaux de gris" 1^{er} Colloque International Francophone sur l'Écrit et le Document CIFED 98, quebec, may 11-13 1998, p82-91
- [8] F. Bapst, R. Ingold, Using Typography in Document Image Analysis, RIDT'98 Raster Imaging & Digital Typography to be published by Springer-Verlag in LNCS.
- [9] A. Belaïd, Y. Chenevoy, Constraints Propagation vs Syntactical Analysis for the logical structure Recognition of library References, Lecture notes in computer science 1339, BSDIA'97, springer, pp 153-164, Curitiba, nov 2-5, 1997