From Exchange Format to High Level Document Description

N. Vincent, M. Crucianu, R. El Ayadi
LI/E3I – Université de Tours
64, avenue J. Portalis
37200 Tours – France
{vincent,crucianu}@univ-tours.fr

C. Faure, M. Giba
ENST
46, rue Barrault
75634 Paris – France
cfaure@lsi.enst.fr

M. Venet
Sharing Technologies
Tour Franklin, La Défense 8
92042 Paris La Défense – France
venet@sharing.com

Abstract

We are presenting here the main steps needed to build a high-level description for any shared document when the common exchange format employed relies on a low-level, graphical representation of the electronic document. We insist on the similarity between the methods put to work in this case and those encountered in the field of document image processing. Our approach strongly relies on perceptual grouping methods, but we modify them according to the elementary components to which they are applied. The role of the logical structure of a document is highlighted; we then show why the logical labeling and the physical labeling are complementary.

1. Introduction

The nets, either Intranet within the companies or Internet, give communication facilities to capture, to share and to use information supported by electronic documents. Solutions have been proposed to facilitate the collective use of documents by a group of people. Papirus is one solution developed by Sharing Technologies. This software enables the members of a community of practice to collaborate by sharing the same data and knowledge. Figure 1 shows the global organization of the system. An original document is "printed" to a file to be encoded into PGF (Papyrus Graphical Format). Compared to the pixel level, this format is a high-level description that includes both text format information and graphical information such as lines or rectangles. It also supports images. Under this format, documents can be visualized in a WYSIWYG mode and directly manipulated regardless of the applications used to create the documents. As far as our work is concerned, the PGF file is a set of low-level components (textual, graphical, bitmap) that are defined according to a "printer logic" approach. Figure 2 displays in (a) the image of an original document build using both a text editor and a graphic editor, in (b) the set of associated PGF elements, in (c) two of the physical text structures which must be extracted from the PGF file, namely text lines and paragraphs, and in (d) some elements of the logical structure.

Figure 1. General process to be improved.

The goal of this project is to build from the PGF a high-level document description in terms of components and relationships that are meaningful for the reader. Such a representation is considered essential for improving the usability of the HCI and for enabling new functionality. We only mention here document indexing, or the
The automatic construction of a table of contents together with the associated links. Besides, connections to existing applications and tools become possible if the new representation is expressed into a standard such as XML.

The detection of the "reader structure" from the PGF document is a problem that has a lot in common with document image analysis. Other studies, such as going from Postscript to HTML [POB97], are centered on similar questions.

During the early stages of this project, we concentrate on the three important issues below:
- perceptual grouping,
- distinction between physical and logical structures,
- document description.

2. Perceptual grouping

Perceptual grouping refers to the operations leading to the segmentation of a document into homogeneous regions. The meaning of "homogeneous region" is dependent on the description features that are used. The major criteria involved are the spatial relationships (alignment, proximity, inclusion, intersection) between boxes, the white areas or free spaces and, if available, the typographic characteristics (size, style, ...). A structural graph can be defined according to these relationships and associated with the document.

The basic concepts of perceptual grouping in document image analysis can be reused when processing the PGF. The "image boxes" are then the bounding boxes of the connected components (before and/or after RLSA filtering). In fact, when processing a PGF document, the elementary boxes correspond to the bounding boxes of the PGF components that are available in the code. Then, the usual algorithms have to be modified, since we are no longer concerned with the pixels, but instead with more global boxes that can be significant at various levels of interpretation.

For the first stages of our processing, our main concern is to rebuild a visual representation of the document, i.e. using elements that are perceived by humans. While the basic concepts of perceptual grouping can be reused when processing the PGF, specific problems arise. In the image of a document, only the visible information is represented, but in a PGF file invisible information is also encoded, such as blank characters, transparent surfaces or hidden objects produced during the editing phase. Therefore, a rather complex filtering step must be performed in order to detect the "visual" bounding boxes, analogous to the image boxes. Moreover, several PGF components may correspond to the same "visual" component. For example, a text in shadowed characters may be represented by two text components containing the same symbols and located in slightly different positions.
classified, according to studies in interval and rectangle algebra. It appears that such a syntactic level of classification leads to ambiguities. In choosing the interpretation and the action to be performed, one must rely on the label of the components (text, graphic, image) or even finer visual attributes or symbolic content, and very often on the spatial context.

3. Physical versus logical structures

This last remark highlights the need for a strong integration of several kinds of information to detect the structure of a document when no (or a poor) prior model of the document is available. Actually, it is difficult to constrain the document analysis process to follow a two-step process, a first one for detecting the physical structure (without using the symbolic content of the textual parts) and a second one for labeling the physical components with logical labels. Examples given in [FAU99] show the need for perception and symbol reading in order to detect the reading order, the captions and the list of headings in a scientific paper. Other examples show that physical breaks do not necessarily engender logical breaks (e.g. end of page, end of line), and logical breaks do not necessarily rely on physical breaks (e.g. inline formulae, inline enumerations, inline headings). Moreover, visual appearance remains ambiguous in the absence of symbolic and/or contextual information (e.g. inline heading in bold vs. a group of words in bold in a same line).

In fact, if the distinction between the physical structure and the logical structure may be valuable in the final description of a document, any attempt to separately extract these two levels of description should clearly be avoided.

4. Document representation

The PGF encoding has the advantage of providing reliable symbolic information, unlike in document image analysis where OCR errors are a source of problems. In our application, this advantage is counterbalanced by the great variability in document layout. The documents we must process may come from various sources and cannot all be classified in any previously described class. The documents are edited for several purposes (reports, oral presentations, quantitative studies,...), using text editors, graphic editors or table editors.

The intended document description should be general enough to capture any kind of layout (see Web documents for examples of layout diversity). The document image analysis community put forward document descriptions that are influenced by the underlying structures handled by text editors. The layout components for which a style

---

(c) Physical text structures to be extracted

(d) Elements of the logical structure

Figure 2. One page of a sample document.
may be defined during the editing phase are all explicit in
the description: page, column, paragraph,... Tables,
graphics and bitmaps are usually regarded as blocks and
are not broken down. Furthermore, several proposals were
made for the description of composite documents
[WAN99] and the description of tables [KIE98].

The notion of "reading unit" is introduced in order to
cope with the problem of document layouts which do not
follow the text editor model or the table model. A reading
unit has its own reading order, and can be read
independently of the others. Reading units often appear as
floating parts or may be the annotations, the footers or the
headers. A few such examples can be seen in the
document shown in Figure 2, where one finds a header, an
outline of the document and a special note enclosed in a
rectangle. The frames in Web documents may also be
considered as reading units. In the case of a document
produced with a text editor, the header, footer or even the
list of references in a scientific paper may again be
regarded as reading units. This new type of component
provides further flexibility when the document layout is to
be interpreted or modified. However, the representation
becomes more difficult as two incompatible aspects are
concerned, physical layout and logical structure. The
flexibility is important when transforming the original
document into hypertext, when the system must decide
how the user should navigate and how each reading unit
should be visualized. For example, the list of references is
no longer the last section of the paper but an autonomous
reading unit that can be visited by direct access or by
activating an internal link in the document.

The automatic detection and description of document
structure is a stage in the life of the document that is not
performed independently of the other stages. It is
influenced by the editing process, the intention of the
writer, the kind of editing tool and the kinds of
information that are needed in the future to transform the
document into hypertext or to index it for knowledge
management or data mining.

5. Conclusion

We attempt to build, from a Papirus Graphical Format
(PGF) file, a high-level document description that is
meaningful for the reader and enables new functionalities.

Our work shows that the basic concepts of perceptual
grouping in document image analysis could be reused,
with several modifications, when processing the PGF. We
take advantage of the fact that the PGF encoding provides
reliable symbolic information, unlike in document image
analysis. But, in our application, this advantage is
counterbalanced by the great variability in general
document layouts and by the fact in the PGF files
invisible information is also present.

We are developing a general representation of
documents that takes into account both physical and
logical structure. We found that these two aspects cannot
be dissociated when reconstructing a document equivalent
to the initial one. Nevertheless, these two aspects may
lead to conflicting hypotheses that have to be solved.

The improvement that our study will bring can be
evaluated by the degree of satisfaction of Papirus clients
when reusing the shared documents.

6. Acknowledgment

This work is performed within the RNTL project
Kenobi, supported by the French Ministère de l'Industrie.

7. References

[1] C. FAURE. Prattective Reading and Selective Attention
for Document Image Analysis, Conf. on Document Analysis and
Recognition (ICDAR'99), Bangalore, India.

table converting system. Document Analysis systems (DAS98),
Nagano, Japan.

[3] B. POIRIER, M. D'AGENAIS. An interactive System to
Extract Structured Text from a Geometrical Representation,
Conf. on Document Analysis and Recognition (ICDAR'97),
Ulm, Germany.

to SGML/XML Representation: One Method, Document Layout
Interpretation and its Applications, (DLIA99), Bangalore, India.