Geometric and semantic aspects for the document layout analysis

Cinzia Marte

Department of Mathematics and Computer Science,
University of Calabria, 87036 Rende, Italy
marte@mat.unical.it

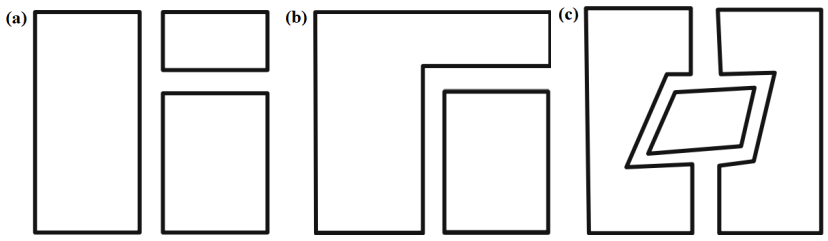
Abstract. The paper presents the research goals for the near feature. I explain the background of the field of document layout analysis, summarize what I have done in my master thesis and outline the research directions for my PhD studies. I work under the supervision of Prof. Marco Manna from the Department of Mathematics and Computer Science at the University of Calabria (ItaIy).

Keywords: Document Layout Analysis, NLP, Information Extraction, Table Recognition, Knowledge Representation and Reasoning

1 State-of-the-art

Document digitalization has been growing in importance for years, and the automatic document processing along with it. Understanding an electronic document is further complicated by the fact that, nowadays, most of the documents are created in various electronic formats, such as PDF. Consequently, the flow of information can be lost, the relations between different parts of the document are more difficult to recognize and, therefore, even documents with a well-outlined structure become difficult to automatically "understand", with respect to other format. This understanding is crucial for tasks such as *Information Extraction* or *Information Retrieval*, etc. Determining what areas are non-text or text, determining the reading order of text blocks and determining information like title, author, etc. constitute important aspects of the *document layout analysis (DLA)* [5, 7], that is defined as the process of identifying the layout structure of a document image.

Documents can have different layouts, that can be grouped in three categories: the simplest represented by *Rectangular layouts*, that is a sub-class of the *Manhattan layout* and the last, more complicated, named *non-Manhattan layouts* (see Fig. 1). There are several ways to analyze a document image [3]. Some algorithms use a bottom-up approach, for example [9, 4, 8], that is a "traditional" approach, that analyzes iteratively a document, starting to link together base elements specified in detail a priori (e.g. pixels, words or text lines) to form larger subsystems, which then in turn are linked, until a complete top-level system is formed. Other use a more recent top-down approach [1, 6], in which a



(a) Rectangular layout (b) Manhattan layout (c) Non-Manhattan layout

Fig. 1. Sub-classes of non-overlapping layout

document is iteratively "cut" into columns and blocks, based on various geometric information and white spaces. There also exits techniques that combine both approaches [2].

The process of analyzing a document image can be *physical*, in the sense that the document is divided in homogeneous areas, or *logical*, when logical roles are assigned among the areas previously extracted. The first is called *geometrical analysis*, that aims to obtain a page segmentation, that is to find maximal homogeneous regions and the spatial relations of these regions (a region is homogeneous if all its area is of one type: text, or figure, etc.). The second is called *logical analysis*, and it aims to assign different semantic labels to the found regions. The transformation of the physical structure of a document into the logical structure is a critical component of document image understanding. Moreover, it is important to notice that logical structure is layout-independent: the contents of a newspaper story and of the various components in the story (text, caption, photograph) as well as the semantic relation between these components, are not affected by the manner in which the story is geometrically arranged on the newspaper page. The standard steps involved in the transformation of a document into its logical description are:

1. physical segmentation of the image into its constituent "blocks";
2. categorization of each of these blocks into blocks types;
3. labeling blocks using domain knowledge;
4. logical grouping of these blocks, using further domain knowledge;
5. determining the reading order of text blocks.

2 Research proposal

Until now, in my master thesis, we focused on the process of identifying "self-consistent" blocks of text from a PDF document. Intuitively, the term self-consistent may refer to a paragraph, a column, a cell of table or even a title. With our algorithm, inspired by the pioneer work of Gorman [8] and a seminal paper by Kieninger [4], we propose a novel approach that seems to unify most of the previous techniques. It is a bottom-up approaches that processes the blocks

starting from the words. Note that, from a physical point of view, any PDF document stores in a proprietary format the positional information of each "atomic" word (as well as any other kind of object, such as lines or images) occurring in it. From a logical viewpoint, each such a *word* can be characterized basically as a tuple

$$w = \langle id, val, x_1, y_1, x_2, y_2, \kappa \rangle,$$

where: (i) its first element id is a positive integer that unequivocally identifies w in the document, and that implicitly induces some total ordering among all the words in the document; (ii) the element val is a string of characters excluding every whitespace, and representing the actual content value of this word; (iii) the pair (x_1, y_1) represents the Cartesian coordinates of the top-left corner of w; (iv) the pair (x_2, y_2) represents the Cartesian coordinates of the bottom-right corner of w; and (v) the number κ is a parameter — depending on the actual size and style of the characters of w — that provides an upper-bound on the width that a standard whitespace wollowing w should have.

Under this assumption, the algorithm is divided in two steps: in the first, the idea is to proceed with the horizontal clustering among words, to form a line block. To this end, we define the notion of "follows", that fixes a binary relation, and so a graph, among the words. In the second step, we apply vertically the same principle as in the horizontal clustering, to obtain blocks of text by grouping line blocks, defining the concept of "connected line", and as above, we define a graph associated to the line blocks. For both kinds of clustering we use some geometrical proprieties of the elements of the document image in PDF format.

Till now, we analyze the document just from a syntactical point of view. However, we notice that a purely syntax analysis fails in some special cases. For example, when two words, or two line blocks are separated by a substantial space, we are not really sure about their connection. For this reason we think to improve our algorithm with some semantic estimation.

First of all, before identifying the final block, we would like to determine a reading flow, because it can change completely the final result of the page segmentation. To this end, we could improve the definition of the relation "follows" and try to estimate how confident we are about this selection. We plan to improve the graph representing words and their connection by adding values on the edges. We envision to estimate this value both from syntactic and semantic point of view. In particular, we can analyze the words' meanings and collocations, grammatical rules, etc. The idea is to identify a text block as a distinct path of the graph that represents the entire document.

We would like to recognize also the logical structure of the document. Notice that, as for the layout of a document, human readers easily understand their logical structure, but it is not trivial how to extract it automatically. The key information used by a human reader is headings in the documents, exploiting the fact that headings appear at the beginning of the corresponding blocks, headings of the same level share the same visual style and, finally, headings of higher levels are given more prominent visual styles. In additional to this features, useful to find the logical structure of a document, we plan to define

a multi-variable function (one variable for each aspect considered - semantical, syntactic, ...), that represents the probability that two words are consecutive. Our final goal is to combine efficiently all this information, exploiting as much semantic information as possible, in order to obtain a realistic page segmentation.

References

1. Baird, H.S., Jones, S.E., Fortune, S.J.: Image segmentation by shape-directed covers. In: Proc. of ICPR. vol. 1, pp. 820–825. IEEE (1990)
2. Cao, H., Prasad, R., Natarajan, P., MacRostie, E.: Robust page segmentation based on smearing and error correction unifying top-down and bottom-up approaches. In: Proc. of ICDAR 2007. vol. 1, pp. 392–396. IEEE (2007)
3. Cattoni, R., Coianiz, T., Messelodi, S., Modena, C.: Geometric layout analysis techniques for document image understanding: a review. IRST, Trento, Italy (1998)
4. Kieninger, T.G.: Table structure recognition based on robust block segmentation. In: Photonics West'98 Electronic Imaging. pp. 22–32. International Society for Optics and Photonics (1998)
5. Mao, S., Rosenfeld, A., Kanungo, T.: Document structure analysis algorithms: a literature survey. DRR 2003, 197–207 (2003)
6. Nagy, G., Seth, S., Viswanathan, M.: A prototype document image analysis system for technical journals. Computer 25(7), 10–22 (1992)
7. Namboodiri, A.M., Jain, A.K.: Document structure and layout analysis. In: Digital Document Processing, pp. 29–48. Springer (2007)
8. O'Gorman, L.: The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1162–1173 (1993)
9. Wong, K.Y., Casey, R.G., Wahl, F.M.: Document analysis system. IBM journal of research and development 26(6), 647–656 (1982)