Bidimensional Relations for Reading Order Detection

Marco Aiello
DIT, University of Trento
Via Sommarive, 14
38100 Trento, Italy
aiellom@dit.unitn.it

Arnold M.W. Smeulders
ISIS, University of Amsterdam
Kruislaan 403
1098 SJ Amsterdam, The Netherlands
smeulders@science.uva.nl

Abstract

We use a propositional language of qualitative rectangle relations to detect the reading order from document images. Document encoding rules are introduced and, expressed in the propositional language of rectangles, are used to build a reading order detector for document images. Results of testing the framework on a collection of heterogeneous document images are reviewed.

Introduction

We are witnessing a constant growth in the proliferation of documents in electronic format. However, in our everyday life we still deal with paper documents. Often, one needs an electronic version of a paper document. A raw image of the document is a way of achieving this, but it has a number of drawbacks: it is usually of big size, it can hardly be edited, transmitted and enjoyed in different modes. An understanding of the document image is necessary to perform a number of tasks, such as document reproduction, digital libraries, information retrieval, office automation, and text-to-speech.

One may have different goals when performing document understanding. For instance, one may be in interested in the reconstruction of the reading order of a document from its image. One way to achieve this is by performing the following intermediate steps. First, one identifies the basic components of the document (*document objects*). Second, one identifies the logical function of the document objects within the document such as title, page number, caption (*logical labeling*). Last, one infers the order in which the user is to read the document objects. This phase is called the *reading order detection*. In the process, one moves from basic geometric information of the document composition, the *layout structure*, to semantic information, the *logical structure*.

In (Aiello et al., 2002), we have presented a logical structure detection architecture. Departing from a pre-processed document image the goal of such an architecture is that of logically labeling the document objects and subsequently identifying the reading order. Here we focus on the reading order detection component of the architecture. In particular, we argue that a qualitative language of rectangles is suitable for capturing the basic reading rules that govern documents. The system exhibits flexibility in that it can analyze heterogeneous collection of documents. The methodology is implemented in a prototype system named SpaRe (Spatial Reasoning component). To ground our argument we review a number of experiments performed on the prototype.

Related work The first document image analysis systems were built to process documents of a specific class, e.g., forms for telegraph input. One of the recent trends is to build systems as flexible as possible, capable of treating the widest variety of documents. This has led to categorize the knowledge used in a document image analysis system into: class specific and general knowledge (e.g., Cesarini et al., 1999). In addition, such knowledge can be explicitly available or implicitly hard-coded in the system.

Lee et al. (2000) present a system to analyze technical journals of one kind (PAMI) based on explicit knowledge of the specific journal. The goal is that of region segmentation and identification (logical labeling). The knowledge is formalized in "IF-THEN" rules applied directly to part of the document image and "IF-THEN" meta rules. Though the idea of encoding the class specific knowledge of a document is promising, it is not clear whether the proposed approach is scalable and flexible. Given the specific form of the IF-THEN rules, the impression is that the system is not suited for the analysis of documents different from PAMI.

There are a number of problems related to the rule based approaches found in the literature. The most prominent is the high specificity of the rules. The specificity makes it hard or impossible to extend such systems to documents of a class different from the one for which the system was originally designed. Another problem is the lack of proof of correctness or termination. Rule-based approaches for layout and logical structure detection are presented in (Klink and Kieninger, 2001; Lee et al., 2000; Niyogi and Srihari, 1996; Tsujimoto and Asada, 1992).

Given the difficulty in designing appropriate rules for the analysis of documents, approaches based on learning are interesting. The document classification components of the WISDOM++ system (Altamura et al., 2001) are based on first-order learning algorithms (Esposito et al., 2000). Another advantage of such systems is their flexibility compared to the non-learning based systems. By training the system on a different class of documents with similar layout, it should be possible to reuse the same architecture. On the negative side, the rules learned are not human readable or intuitive.

Most often, these rules are impossible to modularize for further use on different document classes.

As we are investigating practical applications of spatial reasoning, it is relevant to review approaches using these kind of formalisms. In particular, we consider bidimensional extensions of Allen's interval relations, that is, rectangular relations. To the best of our knowledge, bidimensional Allen relations have been used in document image analysis in three cases (Klink et al., 2000; Singh et al., 1999; Walischewski, 1997). In all these approaches, bidimensional Allen relations are used as geometric features descriptors, at times as labels for graphs and at other times as layout relations among document objects. Thus, the use of Allen relations is relegated to mere feature comparison, while we propose an inference mechanism based on these relations were documents are seen as formal spatial models and reading orders are the output of a constraint satisfaction process.

Methodology

In Figure 1 the overview of the data and knowledge flow of the methodology we propose is sketched. First, the generic document knowledge in the form of document encoding rules may have different origins (from an expert, from previous learning or are given directly by the document author.). Second, the spatial reasoning module `SpaRe` is actually composed of two sub-modules. The first one, which performs inference on the spatial relations of the layout and on the document encoding rules, is based on constraint satisfaction techniques. The second one is a module to sort graphs, that is, directed transitive cyclic ones. In the following sections, we analyze each of these items.

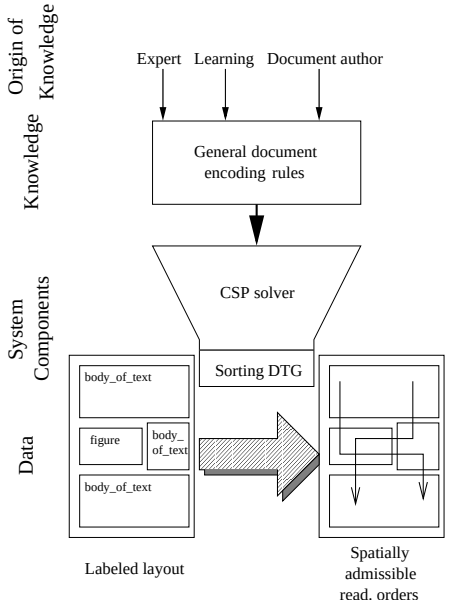

Figure 1: The flow of knowledge and data in the proposed methodology.

Document encoding rules A *document encoding rule* is a principle followed by the author of a document to convey an intent of the author by layout details. Document encoding rules can be one of two types: general or class specific. Document encoding rules can be expressed in a informal or in a formal manner. Informal rules are proposed in natural language or by sketch. Examples are found in books such as (Reynold, 1979). Formal rules may be expressed in a number of ways which include LaTeX, WYSIWYG (e.g., MS Word), SGML languages, or abstract languages. For a discussion of how the above formalisms may express document encoding rules we refer to (Aiello and Smeulders, 2002).

An example of a general document encoding rule stated informally in natural language is:

> "in the Western culture, documents are usually read top-bottom and left-right."

One can immediately spot a problem of stating rules in natural language, that is, ambiguity. In fact, we do not know if one should interpret the "and" as commutative or not. Should one first go top-bottom and *then* left-right? Or, should one apply any of the two interchangeably? It is not possible to say from the rule stated in natural language.

Considering relations adequate for documents and their components, requires a preliminary formalization step. This consists of regarding a document as a formal model. At this level of abstraction a document is a tuple $\langle D, R, l\rangle$ of document objects D, a binary relation R, and a labeling function l. Each document object $d \in D$ consists of the coordinates of its bounding box (defined as the smallest rectangle containing all elements of that object)

$$D = \{d \mid d = \langle id, x_1, y_1, x_2, y_2\rangle\}$$

where id is an identifier of the document object and (x_1, y_1) (x_2, y_2) represent the upper-left corner and the lower-right corner of the bounding box. In addition, we consider the logical labeling information. Given a set of labels L, logical labeling is a function l, typically injective, from document objects to labels $l : D \rightarrow L$. In the following, we consider an instance of such a model where the set of relations R is the set of bidimensional Allen relations and where the set of labels L is {title, body_of_text, figure, caption, footer, header, page_number, graphics}. We refer to this model as a *spatial [bidimensional Allen] model.* Bidimensional Allen relations consist of 13×13 relations: the product of Allen's 13 interval relations (Allen, 1983) on two orthogonal axes. Each relation $r \in A$ is a tuple of Allen interval relations of the form: $precedes$, $meets$, $overlaps$, $starts$, $during$, $finishes$, $equals$, and $precedes\_i$, $meets\_i$, $overlaps\_i$, $starts\_i$, $during\_i$, $finishes\_i$. We refer to the set of Allen bidimensional relations simply as A and to the propositional language over bidimensional Allen relations as $\mathcal{L}$ the remainder of the paper. Since Allen relations are jointly exhaustive and pairwise disjoint, so is A. This implies that given any two document objects there is one and only one A relation holding among them.

Document objects are represented by their bounding boxes and the relative position of these objects plays a key role in the interpretation of the meaning of the document.

For example, if a document object is above another one it is likely that it should be read before. If a document object is contained in another one, it is likely that the contained one is a 'part' of the containing one, for instance the title of a remark inside a frame. Document objects can be also overlapping. This last feature is more common when the document has different colors and colored text runs over pictures, logos and drawings.

Document encoding rules with $\mathcal{L}$ The language $\mathcal{L}$ is adequate to express mereotopological and ordering relations among rectangles. Here, we show how to use this power to express formal unambiguous document encoding rules.

Take an informal document encoding rule expressed in natural language which says that documents are organized in columns that go from top to bottom and from left to right. A rule to encode this behavior is expressible with $\mathcal{L}$ in the following way:

$\text{before\_in\_reading}^{\text{col}}(d_1, d_2)$ iff

$$
\begin{aligned}
&precedes\_x(d_1,d_2) \vee meets\_x(d_1,d_2) \vee \\
&(overlaps\_x(d_1,d_2) \wedge \\
&(precedes\_y(d_1,d_2) \vee meets\_y(d_1,d_2) \vee overlaps\_y(d_1,d_2))) \vee \\
&((precedes\_y(d_1,d_2) \vee meets\_y(d_1,d_2) \vee overlaps\_y(d_1,d_2)) \wedge \\
&(precedes\_x(d_1,d_2) \vee meets\_x(d_1,d_2) \vee overlaps\_x(d_1,d_2) \vee \\
&starts\_x(d_1,d_2) \vee finishes\_i\_x(d_1,d_2) \vee equals\_x(d_1,d_2) \vee \\
&during\_x(d_1,d_2) \vee during\_i\_x(d_1,d_2) \vee finishes\_x(d_1,d_2) \vee \\
&starts\_i\_x(d_1,d_2) \vee overlaps\_i\_x(d_1,d_2))) \qquad (1)
\end{aligned}
$$

An analogous *row-wise* rule is obtained by inverting the axes in (1).

Inference Equipped with a qualitative spatial language for document objects $\mathcal{L}$, with document encoding rules and the layout and logical labeling information, we are now in the position to perform inference in order to achieve 'understanding' of a document. Following is the definition of document understanding in this context.

First, we define the notion of an admissible transition between document objects. Given a pair of document objects d_1 and d_2, a document model $\langle D, R, l\rangle$ and a set of document encoding rules S, we say that (d_1, d_2) is an *admissible transition* with respect to R iff the bidimensional Allen relation $(d_1, d_2) \in R$ is consistent with S.

A *spatially admissible reading order* with respect to a document model $\langle D, R, l\rangle$ and a set of document encoding rules S is a total ordering of document objects in D with respect to the admissible transitions.

The *understanding* of the document with respect to a document model $\langle D, R, l\rangle$ and a set of document encoding rules S is the set of spatially admissible reading orders.

Following the above definitions, we see that inference is performed by two following steps. The first one is a constraint satisfaction step in which instances of bidimensional Allen relations are matched against document encoding rules expressed in $\mathcal{L}$. The second one is a graph sorting procedure similar to topological sorting.

Algorithmic details are presented in (Aiello and Smeulders, 2002) while the relevant Eclipse[1] source code is in the appendix of (Aiello, 2002).

Complexity The number of textual document objects that can be present in a page can vary greatly. It may go from a few in a simple one column page to more than 20 in complex multi-column pages and, in extreme cases such as big-format newspapers, to over 50. This generates a concern about the complexity of the methodology proposed here. It turns out that the methodology has polynomial complexity in the number of document objects present in a page as shown both formally or empirically in (Aiello and Smeulders, 2002).

Evaluation

The methodology proposed has been implemented in a prototype system: `SpaRe`. The core of `SpaRe` is implemented in the declarative programming language Eclipse, making use of the finite domain constraint satisfaction libraries.

To test `SpaRe`, we used the 171 pages collection Media Team Data-Base (MTDB) from the University of Oulu, (Sauvola and Kauniskangas, 2000) and the 624 pages of the University of Washington dataset UW-II (Phillips and Haralick, 1997). The first data set consists of scanned documents of various types: technical journals, newspapers, magazines, and one-page commercials; while the second consists of scientific journal papers.

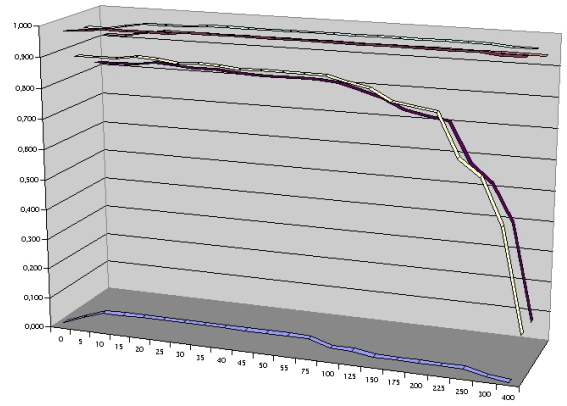

Figure 2: Comparing precision and recall for three different document encoding rules with respect to increasing boundary thickness. From foreground to background, the precision rates for a general rule and two column/row rules, then the respective recall rates for the same rules.

The goal of the experimentation was to evaluate whether `SpaRe` is effective in the detection of the reading order given the layout information. As subtasks, we were interested in evaluating the performance with different document encoding rules, the effectiveness of introducing a notion of thickness—see Aiello and Smeulders (2002) for the

[1] `http://www-icparc.doc.ic.ac.uk/eclipse`.

definition—in the interpretation of bidimensional Allen relations (Figure 2), and the time performance of such a system (Figure 3). Extensive presentation and discussion can be found in (Aiello et al., 2002; Aiello and Smeulders, 2002), here we provide a brief summary.

SpaRe has shown high recall rates (up to 100%). As for precision, SpaRe has shown high rates when using the rule 1 (up to 89%), while exhibiting lower recall rates when using more generic rules.

The introduction of a notion of thickness in the interpretation of Allen bidimensional relations has proved to be essential in avoiding brittleness coming from the raw document images, improving overall performance of 13% to 16%.

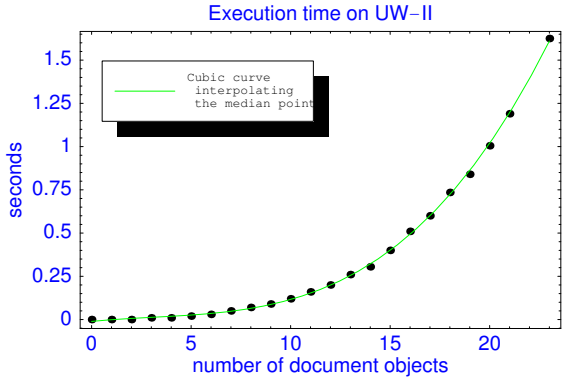

Figure 3: Execution time in seconds with respect to the number of document objects in the pages from the UW-II dataset.

The execution time needed by SpaRe to analyze a document is a cubic function of the number of document objects. In Figure 3, the median execution time for SpaRe on the UW-II collection is shown.

Concluding remarks

We have shown the feasibility, and efficacy, of applying a symbolic approach to logical structure detection in the context of document image analysis and understanding. The approach is based on a spatial language of rectangles and basic mereotopological rectangle relations (bidimensional Allen relations). Inference is achieved via constraint satisfaction.

Two notable features of the presented symbolic approach are its flexibility and modularity. SpaRe is flexible enough to treat a wide variety of documents, including scientific articles, newspapers, magazines and commercial hand-outs, in a single run.

References

Aiello, M. (2002). *Spatial Resoning: Theory and Practice*. PhD thesis, University of Amsterdam. DS-2002-02.

Aiello, M., Monz, C., Todoran, L., and Worring, M. (2002). Document Understanding for a Broad Class of Documents. *International Journal on Document Analysis and Recognition*, 5(1):1–16.

Aiello, M. and Smeulders, A. (2002). Thick 2D relations for document understanding. Technical Report DIT-02-63, DIT, Univ. of Trento.

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843.

Altamura, O., Esposito, F., and Malerba, D. (2001). Transforming paper documents into XML format with WISDOM++. *International Journal of Document Analysis and Recognition*, 4(1):2–17.

Cesarini, F., Francesconi, E., Gori, M., and Soda, G. (1999). A two level knowledge approach for understanding documents of a multi-class domain. In Hull et al. (1999), pages 135–138.

Esposito, F., Malerba, D., and Lisi, F. (2000). Machine learning for intelligent processing of printed documents. *Journal of Intelligent Information Systems*, 14(2/3):175–198.

Hull, J., Lee, S., and Tombre, K., editors (1999). ICDAR, IEEE.

Klink, S., Dengel, A., and Kieninger, T. (2000). Document structure analysis based on layout and textual features. In Murshed, N. and Amin, A., editors, *Proc. of International Workshop on Document Analysis Systems, DAS2000*, pages 99–111. IAPR.

Klink, S. and Kieninger, T. (2001). Rule-based document structure understanding with a fuzzy combination of layout and textual features. *International Journal of Document Analysis and Recognition*, 4(1):18–26.

Lee, K., Choy, Y., and Cho, S. (2000). Geometric structure analsysis of document images: A knowledge approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(11):1224–1240.

Niyogi, D. and Srihari, S. (1996). Using domain knowledge to derive logical structure of documents. In Alferov, Z., Gulyaev, I., and Pape, D., editors, *Proc. Document Recognition and Retrieval III.*, pages 114–125. SPIE.

Phillips, I. and Haralick, R. (1997). Uw-ii english/japanese document image database. CD-Rom.

Reynold, L. (1979). *The Thames and Hudson Manual of Typography*. Thames & Hudson.

Sauvola, J. and Kauniskangas, H. (2000). MediaTeam Document Database II. CD-ROM collection of document images, University of Oulu, Finland. http://www.mediateam.oulu.fi/MTDB/index.html.

Schiirmann, J., editor (1997). ICDAR, IEEE.

Singh, R., Lahoti, A., and Mukerjee, A. (1999). Interval-algebra based block layout analysis and document template generation. In *"DLIA'99"*.

Tsujimoto, S. and Asada, H. (1992). Major components of a complete text reading system. *Proc. of the IEEE*, 80(7):1133–1149.

Walischewski, H. (1997). Automatic knowledge acquisition for spatial document interpretation. In Schiirmann (1997), pages 243–247.