Interval-Algebra Based Block Layout Analysis and Document Template Generation

Rohit Singh, Ankur Lahoti and Amitabha Mukerjee
Center for Robotics
Indian Institute of Technology, Kanpur

{rohitsi,lahoti,amit}@iitk.ac.in

Abstract

Interval algebra provides a qualitative ordering for intervals in one dimension, which can be extended to orthogonal multi-dimensional spaces by using interval projection vectors. These are well suited to capture block layout descriptions in a qualitative structure.

In this work, we present an interval algebra based approach for representing block layout templates as context free grammar involving relative block sizes and positions. This grammar can then be used to identify the layout structure of a document.

Given a set of documents adhering to a certain layout standard, it is often desirable to identify the general structure or rules underlying the formatting. With this algebra, it is possible to identify the individual descriptions for each document and to generalize them into a grammar for the entire class. We present a mechanism to generate these descriptions from individual document images and then unify each of these descriptions to generate a qualitative description of the entire class of documents. We also discuss techniques for simplification of such descriptions.

1 Qualitative Modeling of Block-Structured Images

Given a set of documents, can we learn a template that describes all of them? This is the goal of this paper. We use the tools of interval algebra to obtain a basic description which is then simplified. Many artificial objects have a rectangular block structure which can be interpreted using relative positions of the blocks. In many instances, the exact location of a block is not as important as the relative position between nearby blocks [7]. This concept can be captured using the notion of Qualitative Position [5, 1]. Qualitative Spatial Reasoning [4, 3] extends this notion to that of higher dimensional spaces. A context-free grammar is derived from a set of document images, which can then be used to determine if a new document belongs to this class or not.

The advantage of qualitative models is that unlike *ad hoc* abstractions of quantitative data, the set of relations is **complete** within the assumption that relative positions are unimportant unless involving tangency. This ensures that the description language has sufficient power to be used for discovering relations that exist but otherwise may not be evident from the data. Figure 1 shows an example of a 2-D interval relation based on the relationships of the endpoints with respect to an interval. There can be five spatial relations between a point and an interval (along 1-D). Thus, "**-**","b"',i","f","+" represents the endpoint being behind (**-**), at the back (**b**), inside (**i**), at the front (**f**), and ahead (+) of the interval. The logical extension of this is the spatial relationship between two intervals which can be of thirteen different types. For example, "**-**i" is contained, and "bf" is equals. Thus "A (bf) B" means that A and B are congruent. Note that many important aspects of document layout (such as "Block A is near Block B") cannot be captured directly in interval algebra, which has no notion of proximity. Qualitative models have also been extended to include fuzzy notions for proximity and relative position ("A is to the left of B") but in this work we restrict ourselves to a purely interval algebra and such notions are not considered.

However, the model for document analysis frequently uses certain composite operations which are convenient to define in terms of the basic interval relations. For example, the notion of equality of interval sizes can be constructed from the basic relations using a flush operator which aligns one endpoint of two interval. If the resulting relation is **bf** [2] the two intervals are equal. This can also be extended to the following *centered* operator:

$$SameSize((\mathcal{I}_\infty),(\mathcal{I}_\in)) \Leftrightarrow Flush_{(\mathcal{I}_\in)}((\mathcal{I}_\infty))\{\texttt{ bf }\}(\mathcal{I}_\in)$$

$$CenteredSmaller((\mathcal{A}),(\mathcal{B})) \Rightarrow (\exists(\mathcal{I}_\infty))(\exists(\mathcal{I}_\in))SameSize((\mathcal{I}_\infty),(\mathcal{I}_\in)) \bigwedge((\mathcal{I}_\infty)\ \texttt{bi}\ (\mathcal{B})) \bigwedge((\mathcal{I}_\infty)\ \texttt{-b}\ (\mathcal{A})) \bigwedge((\mathcal{I}_\in)\ \texttt{if}\ (\mathcal{B})) \bigwedge((\mathcal{I}_\in)\ \texttt{f+}\ (\mathcal{A}))$$

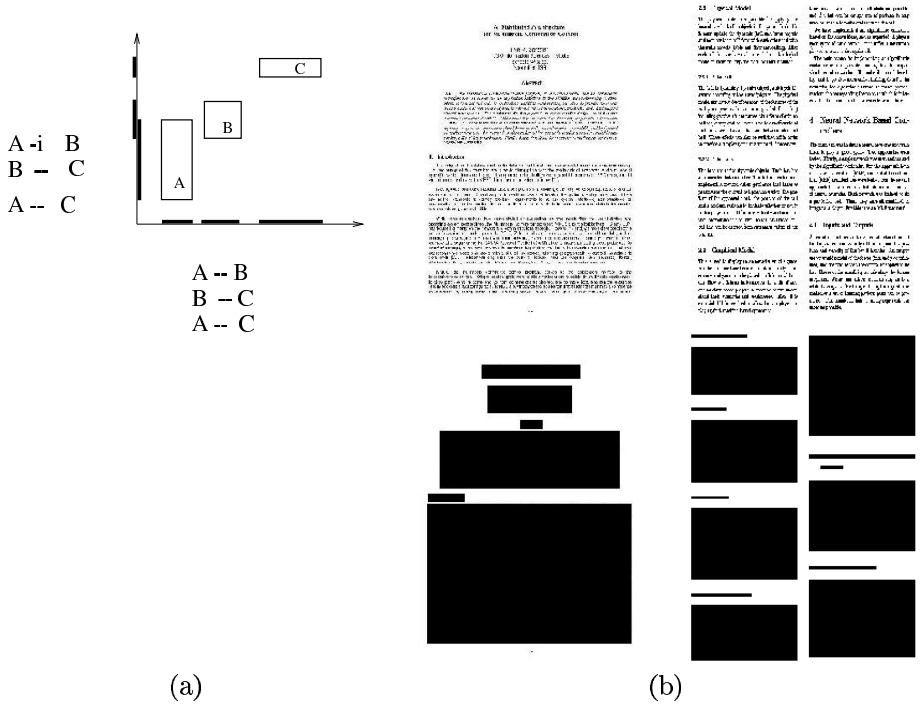

Figure 1: **(a)***Interval Algebra in 2-Dimensions.* For example, along the y-axis, the relation between (A) and (B) is `-i`, i.e. with respect to (B) the start point of (A) is '-' and its endpoint is 'i'.**(b)***A Sample Set of 2 Input Images and their corresponding Block Structures*

where the flush-translation operator $Flush_{(\mathcal{B})}()$ is defined such that $Flush_{(\mathcal{B})}(\mathcal{A})$ returns an interval $(\mathcal{X})$ of the same length as $(\mathcal{A})$ but such that the back endpoints of $(\mathcal{X})$ and $(\mathcal{B})$ are aligned.
An important observation is that the various interval algebra relations follow certain rules of transitivity [2]. This observation is critical in simplifying document descriptions.

2 Block Layout Analysis

The first task is to identify the block structure of a document given the document image. We combine the well-known RLSA [9] technique with histogram analysis to obtain a robust block characterization of a document image. The quantitative block positions can be matched with a known format - i.e. a set of qualitative interval relations. These rules constitute a grammar specifying the the block structure. At this point, a certain tolerance of error in the block positioning is required because it is unlikely that the exact typesetting information will be reproduced accurately.

In matching the rules to the quantitative block structure or vice-versa, it is assumed that the blocks are listed from the top right to bottom left, and that the relations are also ordered in the same fashion. Thus, the title relations are usually the first seen in a title page. This assumption is not a necessary one but it greatly simplifies the matching of a document template to its specified grammar.

3 Document Template Learning

Given certain documents of similar structure, can we identify a common underlying structure? In this work, we observe that a description of a document instance is a conjunction of relations between the constituent blocks. This permits us to use the concept of Version Spaces [6] which is very simple and works well with conjunctive descriptions. In this model, for each positive instance of the document template, the hypothesis corresponding to the block-layout grammar is generalized. Since the input models used are only positive examples, there is at present no need for specialization. However this can

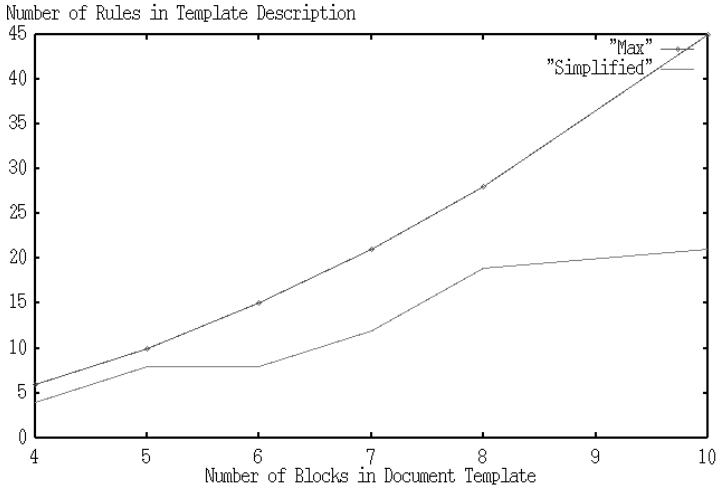

Figure 2: *Grammar Simplification with Increasing Template Complexity.*If the document template had $n$blocks, the initial template , denoted by Max, (generated by simple disjunctions) had $\binom{n}{2}$ rules.

be incorporated, for example, to prune undesirable formatting elements.

The objective is to develop a set of unified grammar rules consistent with all the input document images. Since the relative positions may differ, we will have to take the disjunction of two or more rules and try to find the minimal expression for this disjunction. The simplification process for the minimising this disjunction uses the transitivity rules to determine subsumption of one or more of the rules in the disjunction from other rules of the disjunction. In some cases, we may look for a more general relation expressing all the sub-relations. Following is the simplified grammar that we obtain for the one-column document shown in Fig.1:

$R1$: Title $(^{cs}_{ii})$ Frame
$R2$: Authors $(^{cs}_{ii})$ Frame
$R3$: Authors $(^{cs}_{--})$ Title
$R4$: AbstractTitle $(^{cs}_{--})$ Authors
$R5$: Abstract $(^{c}_{--})$ Authors
$R6$: AbstractTitle $(^{cs}_{++})$ Abstract
$R7$: Heading $(^{cs}_{--})$Abstract
$R8$: Para $(^{cs}_{--})$ Heading

3.1 Grammar Simplification

Combining the different document templates using disjunctions leads to a large set of rules. However, for ease of understanding, it is important that the set of rules (generated grammar) be as compact as possible. So this set of rules has to be simplified to achieve a more compact representation. Ideally, the result of this simplification process must be the prime implicant [8] which logically implies all the positive document structures.
Currently we are using a relatively naive approach in which we choose rules from this original set and add it to a minimal set (which was empty to start with). Then all such rules, in the original set, which can be subsumed by the rules in the minimal set are removed. This process is iterated till the original set is empty. A variety of factors are important while using the transitivity rules to infer subsumptions. We implemented the subsumption test in Prolog which made it somewhat slow. A important factor in this test was the search depth. **Search Depth** for any subsumption is defined as one more than the maximum number of times the transitivity rules are used to infer the subsumption. Atleast in some cases, setting the search depth to a larger value resulted in greater compaction of the rule-set. Fig. 2 shows how the algorithm performed for different number of blocks in a document template.

4 Conclusion

The task of extracting block layout grammars is an important one for a number of domains such as video processing, visual language compilation, VLSI layouts, etc. The methodology of Qualitative Spatial Reasoning provides a powerful tool for abstracting much of this information. A number of approaches are suggested in the literature for handling non-text objects like images and special cases like lists and equations. However, our primary objective was to investigate the block relations. Hence such mechanisms have not been incorporated in the image processing part of the current implementation. Also, a number of difficulties arise owing to complications in the image processing stage which makes it difficult to detect alignments without unnecessarily high tolerances.
More improvements are needed in the template learning part. It is important that the generated grammar be short enough for it to be comprehensible. The notion of Prime Implicates has provided the basic motivation in our approach. Currently, we use subsumption based techniques for grammar compaction. Even in this approach, there is a significant scope of improvement. A factor which affected the size of the minimal set was the choice of the next rule to be included in it. A heuristic based choice function might be able to provide much greater simplification. However, subsumption is not the ideal technique for our purposes [8]. Alternative approaches might yield much better results. Based on the concept of Version Spaces some negative examples can be incorporated in our grammar to generate unified descriptions. Such grammars would conform only to the documents that belong only to a certain class.

References

[1] James F. Allen. Maintaining knowledge about temporal intervals. *CACM, November 1983. Also in "Readings in knowledge representation", ed. Ronald J. Brachman and Hector J. Levesque, Morgan Kaufman, 1985*, 26(11):832–843, 1983.

[2] Hiroko Fujihara and Amitabha Mukerjee. Qualitative reasoning about document structures. In *Symposium on Document Analysis and Information Retrieval, Las Vegas, March 16-18, 1992*, 1992.

[3] Daniel Hernandez. *Qualitative Representation of Spatial Knowledge.* Springer Verlag Lecture Notes in Artificial Intelligence, vol.LNCS804, 1994.

[4] Gene Joe and Amitabha Mukerjee. Qualitative spatial representation based on tangency and alignments. Technical Report Texas A&M University Technical Report 90-014, July 1990, 54 pages, TexasA&M-CS, 1990.

[5] Benjamin Kuipers. MIT Press, Cambridge, MA, Artificial Intelligence Series, 1994,452 pp, 1994.

[6] Tom M Mitchell. Version space: a candidate elimination approach to rule learning. In *Proceedings of the seventh international joint conference of Artificial Intelligence, IJCAI-81, August 1981*, pages 29–37, 1981.

[7] Mukerjee, Amitabha; 1998 Neat vs Scruffy: A review of Computational Models for Spatial Expression In *Representation and Processing of Spatial Expressions, ed. P. Olivier and K-P. Gapp. Lawrence Erlbaum Associates, Mahwah,NJ* 1998, p.1-36.

[8] Ramesh, Anavai; and Neil V. Murray; 1994 Avoiding tests for subsumption *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle WA*, July 1994, p.175-180

[9] Wong K. Y., Casey R. G. and Wahl F. M. Document Analysis System. In *IBM J. of Res. Develop.* 26(6):647-656, 1982.