

Style-Directed Document Recognition

A. Lawrence Spitz

Document Recognition Technologies, Inc.
616 Ramona Street, Suite 20, Palo Alto, CA 94301 USA
email: spitz@docrec.com phone: +1-650-688-0842 fax: +1-650-688-0841

Abstract

We are developing a document recognition system that can be tunably optimized for performance on documents of specific styles. We interactively generate XML to encode specific knowledge about a class of documents to be input to a recognition system. The encoding includes attributes of document logical structure as well as layout structure constraints. The encoding of document style is used to augment the processes of page segmentation, layout analysis, and character recognition. This paper describes the process of specifying logical structure and effect of imposing stylistic limitations on page layout. The resulting recognition process is more accurate and faster than the unassisted version. Recognition results are enhanced by incorporating into the output representation information on the logical structure.

1. Introduction

Knowledge and encoding of the layout structure of a document is interesting in its own right. Additionally, comparison of detected layout against a stylistic model assists in other aspects of document recognition.

Many, perhaps most, applications of document recognition would be improved if the logical structure of the document were encoded along with the content. Though some progress has been made in the development of logical structure from layout information alone, or from the combination of layout and content information [1][2][3], application of document style models provides a considerable advantage in many of those applications.

The knowledge of document style is used to augment the processes of page segmentation, layout analysis, and character recognition.

We will describe the process of interactively generating style information, the types of information and architecture and effects of considering stylistic limitations during the recognition of page images. Even if the document's logical structure is not of particular interest, application of the model results in enhanced speed and accuracy of content recognition.

Most document recognition systems compromise between accuracy and generality. Documents are basically instances of artistic expression; there are few unbreakable rules of document logical or layout structure that apply across the universe of documents. Development of completely general purpose systems is, therefore, difficult, if not impossible. Some developers have been willing to sacrifice accuracy in order to deal with large ranges of image quality and document complexity. Other systems handle only high quality digitizations of a narrow class of documents.

Traditional document recognition systems start with little or no prior knowledge of the particular documents that are to be recognized. They seek to provide recognition services for a broad range of documents and in failing to constrain the input document set or input image quality, such systems are burdened by the need for generality. The range of documents that can be processed by such systems is limited to those that comply with the rules invisibly embedded in the recognition algorithms. This information is developed for an individual document and usually is not retained for as yet unseen documents.

In contrast, our system is designed for application to a relatively small, set of documents which are readily represented by a style sheet.

2. Style Directed Recognition Processing

A previously described system relied on an encoding of style (logical structure, layout structure and optional examples of required content) to direct the course of recognition[4]. In this system the style information was generated manually.

Style representations are typically used in controlling the synthesis of documents. We have inverted the function of style to provide cues useful in the recognition process.

By reducing the ambiguity at several stages of the process, both speed and accuracy are enhanced. Because this system is designed to provide progressive recognition, it advantageously incorporates knowledge of the style of the document throughout the process. The style encoding represents a standard set for the document. Should the document not comply with that standard, recognition will fail. When these failures occur, they occur early enough in the processing to leave the possibility of backtracking rather than proceeding with incorrect basic assumptions.

Our explanation below will relate to traditional scientific journal layout, but application of the system is not restricted to this narrow range of styles.

2.1. Style Generation

Our current system uses a tool to interactively define the areas of the page and to provide the logical labeling. The user is presented with a representative page image from the publication in question and can use either of two methods of defining rectangular page segments: drawing circumscribing boxes with the mouse that are then shrunk to fit, and selecting a point within a segment from which the region is "grown".

2.2. Style Encoding

The coordinates of each side of each rectangle can have one of three attributes: absolute, relative, and variable. Page elements such as headers are likely to have absolute coordinates on the page. The top edge of a title segment can be represented as having a position relative to the bottom of the header. The bottom edge of the title will be variable since the title may include more than one text line.

2.3. Logical tagging

At the time the segment boundaries are defined the user labels the segment with its logical tag. The logical tag can not only take on traditionally expected values such as author, title, abstract etc., but can also indicate that a particular segment is optional, meaning that it might not be present in every instance where this style sheet is to be used. An example of such an optional segment is author affiliation which some journals have only on some articles.

2.4. Layout Segmentation

The style sheet, though developed interactively, can be used in batch mode to guide the layout segmentation of the page. Our layout segmenter is an extension of those developed by Dias [5] and Kise, *et al.* [6] for the segmentation of text lines. It also incorporates important concepts described in Baird for the segmentation of text blocks [7].

The process of applying style to image segments is one of graph isomorphism, albeit with a very simple graph structure. The graph is simple because there is no accommodation for segments which are nested or partially overlap each other. Also segments are, at this time, restricted to being rectangular.

In determining the internal layout structure of the document image, coordinates expressed in the style encoding are used as starting points in a search for the coordinates in pixel space that indicate the position and size of page segments. Scale factors between specified and measured values are developed at the time of processing and therefore are independent of scan resolution and permit automatic compensation for magnification errors and other affine distortions resulting from photocopying and digitization.

For more detail about the methods of registration of the model to the image see Spitz [4].

Early in the recognition process our system develops knowledge of intractability of a document, whether that intractability is due to lack of document style compliance, or to image quality characteristics. Documents are continuously checked against the style encoding for compliance. At any step in the process when it is impossible to reconcile the instance of the document image against the allowable structure as represented in the style encoding, it is possible to backtrack or to abort processing.

Layout structure is checked against the style encoding only to the depth of that encoding. In other words, structure below the level of that recorded in the style is permitted, but the process of page segmentation is terminated by satisfaction of the style requirements.

3. Style Representation

Our earlier system [4]. used the Standard Generalized Markup Language (SGML) because of its relative simplicity, its extensibility in terms of processing instructions, its compatibility with existing systems and the availability of tools for its creation, verification and manipulation. Two important compatible extensions of

basic SGML were implemented. Flexible style encoding for recognition required inclusion of computed variable values based on mathematical expressions that are functions of measured variables, parametric values, and constants. Since that system was developed, XML has been specified and standardized, and we have decided to adopt it for style representation.

4. Recognition Output

The recognition output described in our earlier system is enhanced in the current system by the addition of logical structure information derived directly from the style encoding. The information includes a set of rectangles, each potentially with a logical tag, coordinates on the page and output from character recognition.

5. Conclusion

The output of our recognition process includes information about the logical structure of the document derived from layout information, content and stylistic models.

This work has been directed at documents with knowable style and therefore is inappropriate for “omni-document” recognition. However, particularly in digital library or document database applications where logical structure information is an extremely valuable asset for information retrieval, style-directed recognition provides a much richer representation of the documents on which to search.

References

1. A. Dengel. “Document Image Analysis - Expectation Driven Text Recognition”, *Syntactic & Structural Pattern Recognition*, Murray Hill, New Jersey, pp 78-87, 1990.
2. Y. Tsuji “Document Image Analysis for Generating Syntactic Structure Description”, *International Conference on Pattern Recognition*, Rome, pp 744-747, 1988.
3. S. Tsujimoto, “Understanding Multi-articled Documents”, *International Conference on Pattern Recognition*, Atlantic City, New Jersey, pp 551-556, 1990.
4. A.L. Spitz, Style Directed Document Recognition. *International Conference on Document Analysis and Recognition*, St. Malo, France, pp 611-619, 1991.
5. A.P. Dias, Minimum Spanning Trees for Text Segmentation, *Symposium on Document Analysis and Recognition*, Las Vegas, pp 51-65, 1995.
6. K. Kise, M. Iwata, K. Matsumoto and A. Dengel, A Computational Geometric Approach to Text-line Extraction from Binary Document Images, *Document Analysis Systems*, Nagano, Japan, pp 346-355, 1998.
7. H. Baird, S. Jones and S. Fortune., “Image Segmentation by Shape Directed Covers”, *International Conference on Pattern Recognition*, Atlantic City, New Jersey, pp 820-825, 1990