Recursive Logical Document Structure Description Using Textons: a Generic Object-Process Analysis

Dov Dori[1], Ihsin Phillips[2], and Robert M. Haralick[3]
[1]Technion, Israel Institute of Technology, Haifa 32000, Israel
[2]Seattle University, Seattle, WA 98195, USA
[3]University of Washington, Seattle, WA 98195, USA

Abstract

Document understanding has attained a level of maturity that requires migration from ad-hoc experimental systems, each of which employs its own set of assumptions and terms, into a solid, standard frame of reference, with generic definitions that are agreed upon by the document understanding community. This work provides a formal definition of the logical structure of a text-intensive documents. The logical structure conveys semantic information that is beyond its character string contents. To capture this additional semantics, document understanding must relate the document's physical layout to its logical structure. A formal, generic framework for the definition and interpretation of any text-intensive document logical structure, that is not restricted by the size or complexity of the document, is proposed, developed and demonstrated. The logical structure of text-intensive documents is described as a hierarchy of textons. The definition of textons provides a powerful and flexible tool for document logical structure analysis. We also propose a method for determining quantitatively in an objective, reproducible, and unbiased way the complexity of any kind of text-intensive documents.

1 Introduction

Document image understanding encompasses the technology required to make paper documents equivalent to other computer exchange media like floppies, tapes, and CD-ROMs. The physical reader of the paper document is the scanner just like the physical reader of the floppy disk is the disk drive and the physical reader of the tape is the tape cartridge drive, and the physical reader of the CD-ROM is the CD-ROM drive. However, contemporary document image understanding involves

more than just recognizing the character string on a paper document and putting it in a format of our favorite word processing system. This is due to the fact that paper documents convey structural information, expressed by the 2-dimensional arrangement of the text and non-text elements on the page. This arrangement is more complex than just a linear stream of characters.

Comprehensive understanding of paper documents is complicated by the fact that information can also be in a drawing or in graphics. Hence, understanding art line drawings, engineering line drawings, perspective projections, graphs, and special kinds of documents like complex mathematical formulae and music scores, are all part of document image understanding. Furthermore, almost any mix of these and other kinds of documentation is conceivable.

Any document, either in paper or electronic form, has a *logical structure*. There are many ways to store and display the physical appearance of the same document, but the logical structure should remain the same because it reflects semantics that is part of the author's intention but beyond the ASCII stream of characters. Determining the logical structure of paper-based documents therefore constitutes an important aspect of document image understanding.

The geometric, physical structure of a paper document is aimed at reflecting its logical structure. Therefore, determining the logical structure entails finding the geometric, physical page layout first. This is followed by distinguishing between text and non-text areas, optical character recognition (OCR) of the text in the text units. The text-blocks should be semantically ordered by the "reading order" (sequencing) of the text units. Further, each text unit should be assigned a semantic label. For example, in a business letter the sender's address, receiver's address, date, opening salutation, body, closing, and signature can be inferred by their relative location in the paper plane. Likewise, in a technical, text-intensive document (better known as a "paper") the title, author(s), abstract, keywords, sections, displayed equations, tables, graphs, illustrations, footnotes, page numbers, reference list, and other logical components can be deduced by their location and/or sequencing, as well as the font, style and size of the characters that make them up. The resulting recognized text strings are formatted such that their 2-dimensional layout, inferred from the 2-D layout analysis, are recorded along with the text itself.

The resulting complex data structure, if constructed correctly, captures the entire semantics of the original document. However, it is in a much more condensed form, providing for both data compression and noise removal. This data structure enables one to retrieve information through querying and to reproduce the original page document with practically no noise.

Historically, OCR has been more intensively researched. Consequently, it has attained a considerable level of maturity. Document page layout analysis constitutes a subject of intensive research and it, too, has reached a certain level of maturity. Physical layout analysis, combined with OCR, provides for complete

reproduction of documents.

The logical document structure is the hierarchy that conveys the semantics of the document. As noted, the same logical document structure can be formatted in a variety of physical layouts by changing such variables as page and font size, number of columns, etc. Yet, the semantics of the document remains unaltered. Logical, or semantic structure analysis is aimed at determining the document structure and complexity and providing for information retrieval that involves more than string matching. For example, one would like to query all the abstracts of all papers in a database which have some keyword combination and were written within a certain time period. To do this, a complex data structure ought to be precisely defined and agreed upon by the OCR and document understanding community.

In this paper we propose a working framework for the logical structure of text-intensive documents. A key definition is that of *texton*, which provides for recursion and a quantitative definition of document complexity. Embedded in the works is a limited survey of the state-of-the-art in document analysis.

Being a complex system, analysis of document structure and layout requires a sound methodology. We employ the object-process analysis (OPA) methodology (Dori et al., 1995; Dori, 1995). We use object-process diagrams (OPDs), which are the graphic tool used by OPA to express both the structure and behavior of a document analysis system within a coherent, unified frame of reference.

2 Literature survey

The emphasis in the works on logical layout analysis is on developing processes (algorithms) to carry out various tasks related to logical segmentation. Tsujimoto and Asada (1990) assume that each block of the geometric page layout contains exactly one logical class. They organize the geometric page layout as a tree. Each new article in a document such as a newspaper begins with a headline which is in the head block. They find the paragraphs which belong to the head block by rules relating to the order of the geometric page layout tree and are able to assign logical structure labels of *title, abstract, sub-title, paragraph, header, footer, page number*, and *caption*. They worked on 106 document images and correctly determined the logical structure for 94 document images.

Fisher (1991) is an extension of Fisher (1990) and describes a rule based system to identify the geometrical and logical structure of document images. Ingold and Armangil (1991) describe a formal top-down method for determining the logical structure. Each document class has a formal description that includes composition rules and presentation rules. They have utilized the technique on legal documents.

Chenevoy and Belaid (1991) use a blackboard system for a top-down method of logical structure analysis of a document image. The system is defined in a Lisp formalism and has a hypothesis management component using probabilities.

Kreich et. al. (1991) describe a knowledge-based method for determining the logical structure of a document image. To obtain the blocks they search for the largest text blocks because these are the most characteristic elements in the document layout. The search consists of grouping together the connected components which are close enough to each other. Once text blocks are determined, lines are found within each of the text blocks and words within the lines. The determination of document layout structure is based on interpreting documents and their parts as instance of hierarchically organized classes. They have defined over 300 classes for a document image and its parts. No performance results are given.

Derrien-Peden (1991) describes a frame-based system of the determination of structure in a scientific and technical document image. The basis of this system is a macro-typographical analysis. The idea is that in scientific and technical documents, changes of character size or thickness of type, white separating spaces, indentation etc. are used to make visual searching for information easier. The technique therefore searches for such typographical indications in the document and recovers its logical organization without any interpretation of its semantic content. The first step is the determination of the geometric page layout, keeping a *part of* relationship between blocks. The logical structure determination removes running headers and footnotes and searches for the text reading order. Text blocks are then compared to logical models of classes and each text block is then assigned a class. No performance results are given.

Yamashita et. al. (1991) use a model-based method. Character strings, lines, and half-tone images are extracted from the document image. Vertical and horizontal field separators (long white areas or black lines) are detected based on the extracted elements, then appropriate labels are assigned to character strings by a relaxation method. Label classes include *header, title, author, affiliation, abstract, body, page number, column, footnote, block* and *figure.* The technique was applied to 77 front pages of Japanese patent applications. They reported that the logical structure for 59 of these documents was determined perfectly.

Dengel (1993) discusses a technique for automatically determining the logical structure of business letters. He reports that on a test set of 100 letters, the recipient and the letter body could be correctly determined. Saitoh et. al. (1993) determine logical layout with text block labels of *body, header, footer*, and *caption.* They tested the technique on 393 document images of mainly Japanese and some English documents. To characterize performance they measured the average number of times per document image an operator has to correct the results of the automatically produced layout. They report that on the average 2.17 times per image areas not suitable for output have to be discarded, .01 times per image misclassified areas have to be correctly labeled, and 1.09 times per image does a text area have to be reset. With respect to text ordering they report that it required moving connections .47 times per image, on the average, making new connections .11 times per image, and re-assigning type of text .36 times per image.

3 Textons and the Generic Document Logical Structure

Text-intensive documents are documents whose main content is textual. Such documents have a varying number of logical levels that depends on their size and structuredness. We define the term "texton" in order to be able to think and talk generically about a logical document hierarchy that is not restricted by a particular number of levels and associated level names, such as section and chapter.

A *character* is the basic symbol of a text-intensive document.

Reading order is the order in which the characters in a text-intensive document must be read in order for it to make sense.

Reading order makes sense only in the context of text-intensive documents. In graphics-intensive documents, such as maps or engineering drawings, the text is just a supplement and enhancement to the graphics.

The logical, or semantic structure of a document can be viewed as a tree, in which the leaves are the characters, or symbols, and the root is the entire document. This structure is depicted in Figure 1.

A *texton* is a logical unit comprising a document, which consists of one or more textons or characters.

This recursive definition of texton provides for flexibility in describing the document logical structure without being limited by an upper bound on the number of levels nor by specific names for those levels.

As Figure 1 shows, scanning the logical structure from top down, the entire document (encyclopedia, book, paper, business letter, etc.) is the root of the tree. Below the root is a varying number of levels of textons. The black triangle along the paths connecting a whole to its parts in Figure 1 is the aggregation symbol (Dori, 1995).

A *root texton* is a texton which is the entire document. Examples of root textons include book, encyclopedia, concordance, dictionary, journal, newspaper, magazine, report, scientific paper, and business letter.

The definition of texton is recursive, and the halting condition is that the constituent texton is a character. The recursive definition of texton encompasses the entire spectrum of levels in any text-intensive document.

A *phrase* is a meaningful collection of words which is a not a grammatical sentence. Examples of phrases include a title of a document or part thereof, a name of a person or an organization, an address, etc.

A *phrase-list* is a collection of one or more phrases of the same nature. Examples of a phrase-lists include the list of authors of a document and an itemized list of phrases in a section.

A *graphon* is a document element in a text-intensive document whose nature is mainly non-textual, and whose function is to illustrate, explain or demonstrate the text.

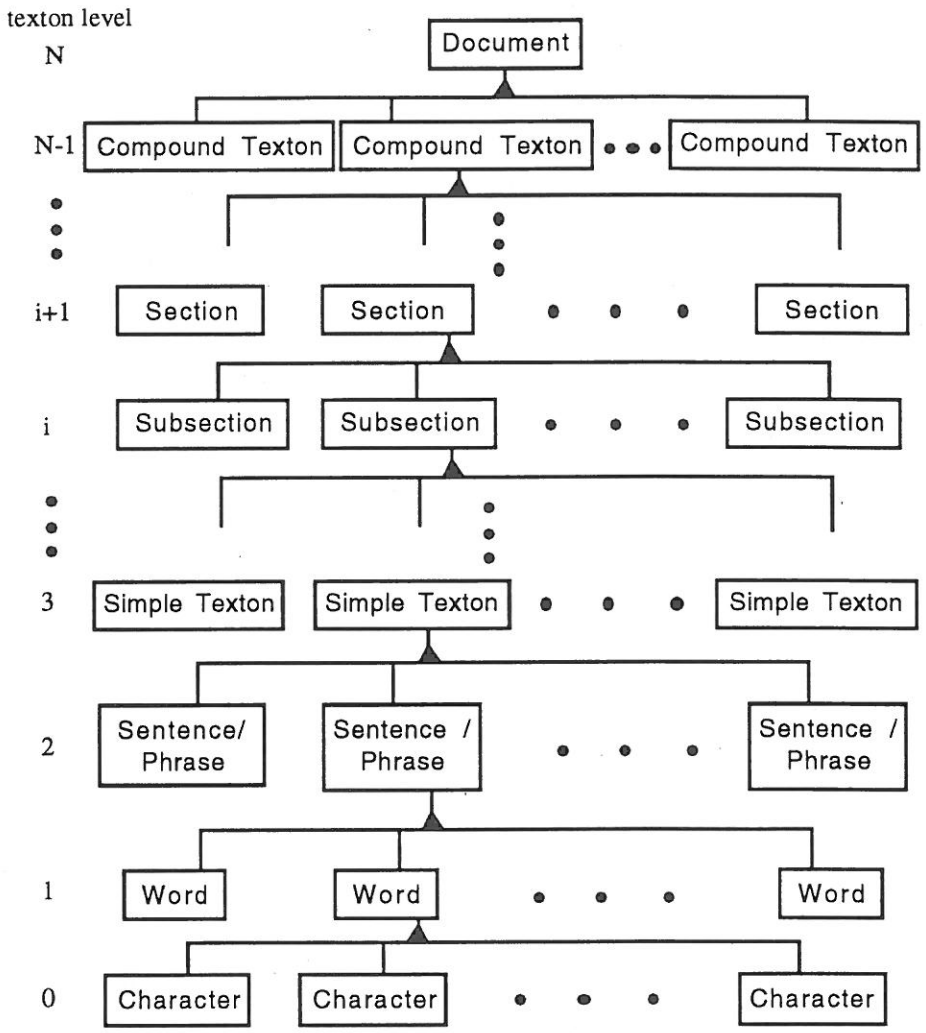

Figure 1: *The logical structure of a document described as a tree*

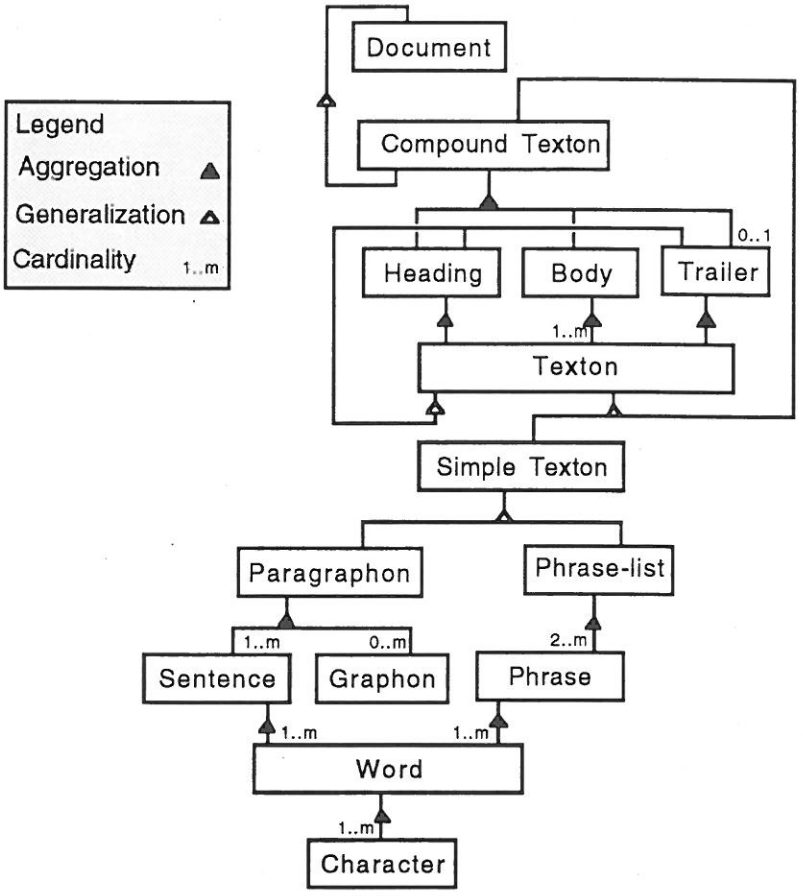

Figure 2: *The logical structure of a document described as a tree*

Examples of graphons are line drawings (engineering drawings or art-line), halftones, photographic (black and white or color) images, maps, diagrams, charts, tables, etc. Although a graphon may contain textual material within, preceding or following the main graphics element, this text explains and enhances the graphics, which, in turn, explains and enhances the text in the text-intensive document.

A table is a boundary case between text and graphon. We classify it as a graphon, because even though it contains text, the text normally does not have a definite linear reading order and it is normally enclosed within graphics—the lines that separate rows and columns.

A grphon in a text-intensive document has a *caption*, which consists of a mandatory *caption header*—the graphon identifier and an optional caption body—the textual title or explanation of the graphon. The caption header is mandatory, because it serves as a reference pointed to by the text.

Since a graphon normally occupies a considerable portion of the page area, as the figures in this document, the physical location of a graphon is allowed to *float* in the neighborhood of where it is referred to in the main text. Logically, however, a graphon is associated with the paragraph in which it is first referred (pointed) to.

A *paragraphon* is a generalization of paragraph, which includes one or more optional graphons in addition to the series of one or more sentences. In the sequel we use the term paragraph to mean paragraphon, unless otherwise stated. For brevity we refer to both paragraphon and phrase-list as the paragraph level.

Complex documents may have textons with names such as subsection, section, chapter, session, part, etc. Hence, the number of texton levels in a document is a finite number, normally not greater than 10, which depends on the nature of the document and indicates its complexity.

Document complexity is the level number of the document's root texton.

Consider, for example, a journal paper, whose body consists of sections. The body of each section is a paragraphon. Assigning the level numbers 0, 1, and 2 to the character, word, and sentence levels, a paragraph is a level 3 texton, and the entire document is a level 4 texton. Hence the complexity of this document is 4. If at least one of the sections is divided into subsections, and no subsection is divided into sub-subsection, then the document complexity is 5, etc.

Although usually there is a relation between the document's size and complexity, these two terms should not be confused. The size can be measured by the number of pages, words or characters. A dictionary, for example, may be a very large document, but its complexity is not necessarily high.

3.1 Simple and compound textons

Having defined textons and their role in the document logical structure, we turn to a more abstract and comprehensive description of the logical document layout than the one given in Figure 1. Figure 2 is an object-process diagram, or shortly, OPD (Dori et al. 1995; Dori, 1995), which describes the structure of a document. The object Document is a specialization of Compound Texton, which is the root of the structure. This is denoted by the generalization symbol—the blank triangle going from Compound Texton to Document.

A Texton is a generalization of Compound Texton and Simple Texton. This is denoted by the blank triangle from Texton to both Compound Texton and Simple Texton in Figure 2. A *simple texton* is a paragraphon or a phrase-list.

A character is defined to be a *level 0 texton. A word* is a level 1 texton, as it consists of one or more characters, and a sentence is a level 2 texton. A simple texton in the main text of the document is therefore a level 3 texton. Below it in the main text reside the sentence or phrase (level 2 texton), the word (level 1 texton), and the character (level 0 texton). As we show below, these level numbers

may vary for side text, such as table of contents in a book.

A *compound texton* is a texton consisting of a distinct header, body, and optional trailer.

Although at the simplest form one may conceive of a primitive document consisting of a single character, perhaps conveying a coded message, a single word document, a single sentence/phrase document or single paragraph/phrase-list document, we refer to the simplest document as a document which is a Compound Texton. Therefore, the minimal complexity of any document is 4. A simple document, such as a standard business letter, is an example of a level 4 document. It has a header—sender and recipient identification and subject, a body—one or more paragraphs (level 3 textons), and a trailer—salutation, signature, etc.

The black triangle between Compound Texton on one hand and Heading, Body, and Trailer on the other hand, is an aggregation (whole-part) relation, expressing the fact that a compound texton consists of these three part. The default cardinality (participation constraint) of the aggregation symbol is 1:1, i.e., one part for one whole.

Consider a texton of level n. The cardinality of the header of this texton is 1, i.e., there is exactly one texton of level $n-1$ which is the header of the level n texton. The cardinality of the texton's body is $1..m$, meaning that there is a number between 1 to many textons of level $n-1$ in the body of the level n texton. Finally, the cardinality of the texton's trailer is 0..1, i.e., there is at most one texton of level $n-1$ functioning as the trailer of the level n texton. The "0..1" next to Trailer denote that Trailer is optional. In other words, a texton has either two or three parts and must have exactly one Heading, one Body and at most one Trailer.

In summary, Heading has exactly one texton, Body has a number of between 1 to many (denoted "$1..m$" in Figure 2) textons, and Trailer, if it exists, has one texton.

For example, a section in a paper is a compound texton. It has a header—the section title, a body, consisting of one or more paragraphons, and no trailer. As another example, a textbook is a compound texton, whose heading is everything from the beginning of the book to the beginning of the first chapter. The body of the book consists of a number of chapters and its trailer is everything from the end of the last chapter to the end of the book (appendices, glossary, index, etc.).

3.2 The recursion in texton definition; the body path

Since Compound Texton is Texton, and Compound Texton has Heading, Body and Trailer, each having at least one Texton, we get a recursive definition. As in any recursion, to avoid infinite looping, a halting condition must exist. The halting condition, as expressed in the object-process diagram of Figure 2, occurs when the textons of Heading, Body and Trailer of the Compound Texton are all

simple. A simple texton implies that it is either a paragraph or a phrase-list. In either case, the recursion stops because from this level on we descend through the Sentence/Phrase level and the Word level down to the character level.

Body path is the path in the tree structure going from the root node—the entire document—through successively decreasing levels of compound textons, all the way down to the simple texton (the paragraph level), such that the path always visits the body of the texton. Since by definition any compound texton has a body, such a path is guaranteed to exist, and it is unique.

The level of a character which is the last node along the body path is defined to be 0. This implies that along the body path the level number of word is 1, the sentence/phrase level is 2 and the level of the paragraphon/phrase-list—the simple texton—is 3. Note that these numbers are not necessarily the same for characters, words, sentences and paragraphs which are not nodes along the body path. As we show in the example below, the level numbers may be higher or lower that the ones along the body path, depending on whether the path from the top texton (the document level) is longer or shorter than the length of the body path, respectively.

The fact that of the three compound texton parts only two are mandatory, gives rise to a 2-3 tree structure, as we demonstrate in the example in the next section.

4 DAS94 Proceedings—a case in point

To demonstrate the use of the concepts and terms presented above, and to show how document complexity is defined, consider the document *Proceedings of DAS94* (Dengel and Spitz, 1994). The structure of the document is described in Figure 3. The structure is detailed down to the Simple Texton level. Since a compound texton may have either three parts (Header, Body, and Trailer) or two parts (Header and Body), the resulting structure in Figure 3 is a 2-3 tree.

As indicated in the legend of Figure 3, the body path is marked by thick line segments. The level numbers are written in parentheses next to the corresponding textons along the path. The body path visits the nodes "Proceedings of DAS94", "Session", "Paper", "Section", "Subsection", and "Paragraph" in this order. Assigning the number 3 to the paragraph level and counting up we get that "Subsection" is at level 4, "Section" is at level 5, "Paper" is at level 6, "Session" is at level 7, and the entire document "Proceedings of DAS94" is at level 8. Hence the complexity of this document is 8.

As a compound texton, "Proceedings of DAS94" has a header, a body and a trailer. The header is a level 7 texton, which, in turn consists of three level 6 texton: a header—"Front Page" and "Copyright note", a body—"Chairmen's Message", and a trailer—"Table of Contents". Chairmen's Message consists of a level 5 header—the title "Chairmen's Message," a level 5 body, consisting of seven

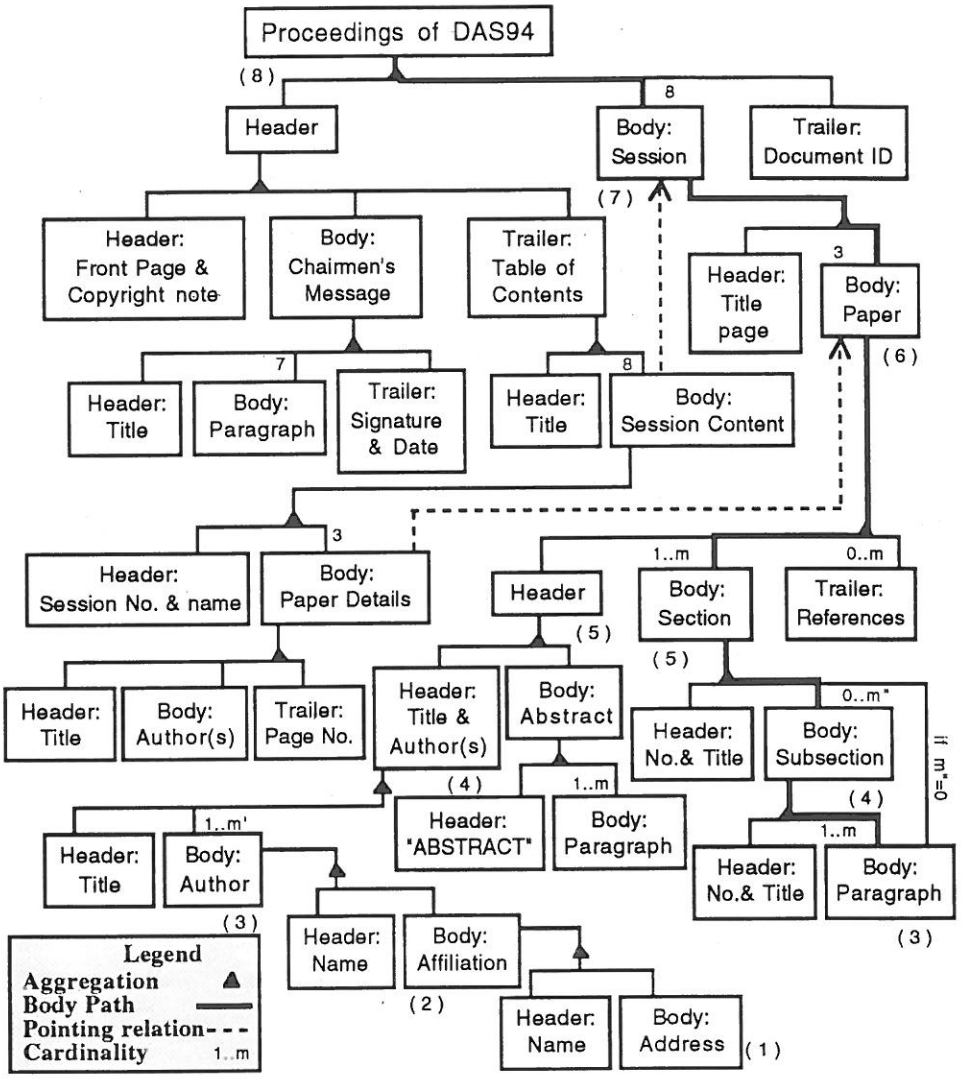

Figure 3: *The logical structure of a document described as a tree*

level 4 paragraphs, and a level 5 trailer, containing two level 4 phrase-lists. The first phrase-list contains the single phrase "Kasierslautern, October 1994," and the second—the names of the two document editors. As we see here, both the paragraphs and the phrase-lists, which are basic textons, are at level 4 rather than 3. The reason is that the path traversed here is not the body path. As noted, a basic texton is guaranteed to be at level 3 exactly only when it is on the body path. In other paths it may be more (as here) or less than 3. The path that ends with "Author", "Affiliation", and "Address", for example, is the longest one. It is longer by two edges than the body path. Therefore "Address", which is a simple texton (a phrase-list) is a level 1 texton in this case, as shown at the bottom of Figure 3.

Table of Contents is a level 6 texton consisting of a header—the title "Table of Contents," a body, and no trailer. The body of the Table of Contents consists of eight items. Each item is a level 5 texton called Session Contents. It has a header—session number and name, a body—an itemized list of three level 4 items, each called Paper Details, and no trailer.

Each Paper Details item is a level 3 texton. It consists of three phrase lists, each containing a single phrase. The first phrase is the paper name, the second phrase is the author name, and the third phrase is the page number.

In most of the papers, Section consists directly of paragraphs, but several papers have subsections (see for example p. 139 in the document). To accommodate this variability, we add the condition "if $m'' = 0$" along the aggregation link from Section to Paragraph in Figure 3, where m'' is the number of subsections in a section. This means that if there are no subsections in the section, then Section consists directly of paragraphs.

5 Summary

The logical structure of a text-intensive document conveys semantic information that is beyond its character string contents. To capture this additional semantics, document understanding must relate the physical layout of the document to the logical structure. This work proposes a formal generic framework for the definition and interpretation of any text-intensive document logical structure, that is not restricted by the size or complexity of the document. The logical structure of text-intensive documents is described as a hierarchy of textons. The definition of textons provides a powerful and flexible tool for document logical structure analysis. We also propose a method for determining quantitatively in an objective, reproducible, and unbiased way the complexity of such documents.

Future research should use the methodology and definitions of in this work to relate the geometric layout to the logical structure. This is a many-to-one relation: The same logical structure can be "disguised" in a variety of shapes and forms, fonts and sizes. However, all of the different forms, if laid out correctly, should

reflect the original intention of the author.

References

[1] Y. Chenevoy and A. Belaid, "Hypothesis Management for Structured Document Recognition," *1st ICDAR*, Saint-Malo, 1991, p121-129.

[2] A. Dengel, "Initial Learning of Document Structure," *2nd ICDAR*, Tsukuba, 1993, p86-90.

[3] D. Derrien-Peden, "Frame-based System for Macro-typographical Structure Analysis in Scientific Papers," *1st ICDAR*, Saint-Malo, 1991, p311-319.

[4] D. Dori, "Object-Process Analysis: Maintaining the Balance Between System Structure and Behaviour," *Journal of Logic and Computation*, 5, 2, April 1995, p1-23 (to appear).

[5] D. Dori, I. Phillips and R.M. Haralick, Incorporating Documentation and Inspection into Computer Integrated Manufacturing: an Object-Process Approach. In *Applications of Object-Oriented Technology in Manufacturing*, S. Adiga (Ed.), Chapman & Hall, London, 1995 (to appear).

[6] J.L. Fisher, "Logical Structure Descriptions of Segmented Document Images," *1st ICDAR*, Saint-Malo, 1991, p302-310.

[7] J.L. Fisher, S.C. Hinds, and D.P. D'Amato, "A Rule-Based System For Document Image-Segmentation," *10th ICPR*, Atlantic City, 1990, p567-572.

[8] R. Ingold and D. Armangil, "A Top-Down Document Analysis Method For Logical Structure Recognition," *1st ICDAR*, Saint-Malo, 1991, p41-49.

[9] J. Kreich, A. Luhn, and G. Maderlechner, "An Experimental Environment for Model Based Document Analysis," *1st ICDAR*, Saint-Malo, 1991, p50-58.

[10] T. Saitoh, M. Tachikawa, T. Yamaai, "Document Image Segmentation and Text Area Ordering," *2nd ICDAR*, Tsukuba, 1993, p323-329.

[11] S. Tsujimoto and H. Asada, "Understanding Multi-articled Documents," *10th ICPR*, Atlantic City, 1990, p551-556.

[12] A. Yamashita, T. Amasno, H. Takahashi, and K. Toyokawa, "A Model Based Layout Understanding Method for the Document Recognition System," *1st ICDAR*, Saint-Malo, 1991, p130-138.