# The Function of Documents

D. Doermann, A. Rosenfeld
Language and Media Processing Lab
University of Maryland
College Park, MD 20742

E. Rivlin
Department of Computer Science
Technion Institute of Technology
Haifa, Israel 32000

## Abstract

*The purpose of a document is to facilitate the transfer of information from its author to its readers. It is the author's job to design the document so that the information it contains can be interpreted accurately and efficiently. To do this, the author can make use of a set of stylistic tools. In this paper we introduce the concept of document functionality, which attempts to describe the roles of documents and their components in the process of transferring information. A functional description of a document provides insight into the type of the document, into its intended uses, and into strategies for automatic document interpretation and retrieval.*

*To demonstrate these ideas, we define a taxonomy of functional document components and show how functional descriptions can be used to reverse-engineer the intentions of the author, to navigate in document space, and to provide important contextual information to aid in interpretation.*

## 1 Documents as Message Conveyors

The general purpose or "function" of a document is to store data produced by a sender in a symbolic form to facilitate transfer to a receiver. Traditionally, the data takes the form of a set of markings on a page, with the sender corresponding to the "author", and the receiver to the "reader". We limit ourselves to the understanding and interpretation of these "traditional" 2D documents which the reader receives visually.

When documents are regarded as message conveyers, we can classify them according to the type of message that is conveyed: differentiate between three classes of messages: informational ( report, dictionary, newspaper, novel, catalogue), instructional (recipe book, a do-it-yourself manual, road sign), and identificational (a street sign, a car license plate, a name tag).

The types of messages describe above are formulated from the author's point of view. The reader, the receiver of the document, may have different goals, and may abstract the document's contents at many different levels. Readers can become quite skilled at abstracting task-dependent information from a document and using this information to establish a context for further interpretation. For example, when looking for documents created on a specific date, an experienced reader can rapidly locate the dates of documents such as business letters and forms without reading them entirely. If it is then decided to "read" the document, the context helps with its correct interpretation and provides a framework in which to proceed through it in an orderly fashion. We can distinguish three basic ways of doing this: **Reading** - which usually involves examining the document from beginning to end (letters, articles, and many types of books); **Browsing** - which involves examining only selected parts of the document to determine if more in-depth examination of these parts is required ( newspapers, magazines, and journals); **Searching** (or referencing) - which involves looking for a specific piece of information in the document (dictionaries, encyclopedias, directories, manuals, handbooks, catalogs, etc).

These modes of interaction with a document apply not only to text-intensive documents; they can also apply to documents which are primarily representational, such as maps and drawings. However, the processes used to read, browse, or search a document depend on the document type. For example, browsing a newspaper and browsing a map have the same basic goal of examining only selected parts, but the methods which are used to accomplish this are quite different. Similarly, searching a phone book and searching a map both require "navigating" and making decisions based on partial information, but they involve different processes. For phone books, one uses index terms and alphabetical relationships; for maps, one uses symbols or landmarks and spatial relationships.

A great deal of work has been done on the analysis of document structure. Almost all of this work, however, has involved models for specific classes of documents. We believe that significant progress in the automated analysis of general classes of documents depends on the development of a general framework for describing document structure. This paper attempts to develop a such a framework.

## 2 Document Structure

In document understanding, documents have traditionally been viewed according to their geometric and semantic organizations[1]. Both organizations have a common *content* which represents a base level of data (typically text, but also possibly including graphics or images). The content's *geometric* nature refers to how it is presented on the page (for example, typeface and

[1] This is the view taken in the ODA standard [6].

0-8186-7898-4/97 $10.00 © 1997 IEEE

| Structure | Example | Use |
|---|---|---|
| header | centered | relative importance, focal point |
| list | enumerated<br>itemized | conveys temporal sequence<br>suggests similar level of descriptiveness |
| separator | white space or rule line | physical and possibly semantic dis-association |
| attachment | footnote<br>boxed text<br>sidebar | supplemental information under some semantic hierarchy |
| illustration | table<br>figure | supplemental - preserves 2D associations<br>graphical representation of information |

Figure 1: Some structures and their uses

font size, for text; line widths and symbols, for graphics), and its *semantic* nature refers to its meaning.

Similarly, a document has both geometric and semantic *structure*. The *layout* structure corresponds to the organization of the document into geometric groupings such as characters, lines, blocks, columns, etc. It describes the relationships among these components and the relationships of the individual components to the entire page. The *logical* structure, on the other hand, organizes the content according to the interpretation of the reader, and also provides global relationships such as reading order. The logical structure corresponds to the document's semantic or conceptual organization.

We claim that there is a level of document organization, which can be regarded as intermediate between the geometric and semantic levels, that relates to the efficiency with which the document transfers its information to the reader. We refer to this level as the *functional* level.

A document obeys conventions such as the use of an alphabet and a language common to the author and reader, and the use of standard presentation rules such as word and line spacing, punctuation, etc. As the information content of the document becomes more complex, these conventions may no longer be adequate for efficient information transfer. Appropriate structures can be used to enhance efficient transfer of information and reduce its ambiguity. For example, an author may use page or section headers to "summarize" content; ordered lists to enumerate or itemize information; separators to "punctuate"; attachments (such as footnotes and sidebars) to subordinate; tables or graphs to present numeric data; maps to present spatial data and their interrelationships. (Note that graphs and maps involve augmenting the basic language with more expressive constructs.) Figure 1 shows some examples of such structures.

As an illustration of the relationship between the geometric, functional, and semantic organizations of a document, consider a **block** of text at the top of a page. Its dimensions and location on the page, as well as properties of its components, are geometric or layout attributes. The fact that we have grouped the components together to form the block is based on geometric proximity. We can use the block's attributes (position, size, etc.) in a class-independent manner to conclude that the block is a **header**; this describes it functionally. If we make a class-dependent identification of the block as a **title**, we have given it a semantic description. Note that a similar block could be a running head or a letterhead in a different context.

The functional description of a document is often independent of document type and can be derived from geometric considerations. Headers, footers, lists, tables, and graphics are examples of generic structures which can be common to many types of documents. Such functional structures will be referred to as class-independent.

If the type of the document is known (for example, business letters or memos, forms, advertisements, or technical articles), a component can have functionality with respect to the documents of that type. For example, in a letter, functional components may include the sender, receiver, date, and salutation. Such functional components will be referred to as class-dependent. The formats used in documents of specific types, such as business letters or journal articles, also serve to enhance information transfer by helping to organize and prioritize the information.

Within a document, structures such as those shown in Figure 1 can be used as aids in the organization of information. The author of a document can take advantage of these principles to design the document so that the reader can use it effectively by using combinations of layout and emphasis to convey an intended organization, or to assign priorities to specific components.

## 3 Exploiting Function

In order to effectively process a document, most document image understanding systems rely on relatively specific information about a restricted domain in order to accurately model the expected document class(es). This allows the system to richly interpret the document, and extract detailed information about its content. For example, in the domain of business letters, a great deal of work has been done on both their structural and logical interpretation ([1], [2], [3], [7], [8], [9], [10]). Unfortunately, for less homogeneous environments this approach cannot be effectively applied. As the set or stream of documents becomes more diverse (both intra-class and inter-class), the formulation of models becomes more difficult. Functional interpretation of documents can greatly facilitate tasks associated with their classification and use. In the following paragraphs we give three examples of tasks which can be addressed by identifying functionally meaningful constructs in documents.

**Use Classification:** In Section 1, we identified three major ways in which a reader can use a document:

reading, browsing, and searching. Documents designed for these purposes can be grossly characterized by the size and organization of their information units, which can be identified by repetitive patterns in the document. For example, reading documents such as journal articles tend to have a single read-order and large information units; browsing documents, such as newspapers or popular magazines, tend to have multiple head-body structures, since their designer's goal is to give the reader quick access to the contents with "handles"; and searching documents tend to have many small information units such as the entries in an index or phone book. An instructional document intended for modification by the reader, such as a form, is characterized by small, blank information units such as horizontal line segments or boxes (including small check boxes).

**Type Classification:** Functional features such as head/body pairs and the locations of handwritten regions allow us to distinguish between document types such as letters and memos. Using functional features, we can achieve a gross categorization of the documents in a database. Given a large heterogeneous database of documents, this allows us to provide groups of documents which are likely to contain some piece of requested information, even if we cannot provide the specific information.

**Functional Enhancement:** We can use the functional organization of a document to help decide which portions of it should be presented to a user and which can be ignored or considered as lower priority. The extraction of functional constructs allows this to be done without the need for content-level reasoning. In fact, many of the relationships which are explicit in the structure cannot be found at the content level; examples are the ordinal relationship between items in a list, or the spatial relationships between columns in a table. Based on these ideas, techniques can be developed to present document images to users who want to browse collections of documents. Such techniques, make it possible to provide documents to a user in a way which is consistent with how the documents were intended to be used, or which is consistent with the goals of the reader.

# 4 Experiments

In this section we describe some experiments on document use and type classification, and briefly outline some methods of functional enhancement. These tasks rely heavily on the identification of information units, information structures and their properties. The first step, therefore, is a segmentation of the document into appropriate information unit primitives whose properties can be used for classification or enhancement.

## 4.1 Extracting Units and Structures

In our experiments, we will consider characters, graphics blocks, and image blocks to be the basic information units. We assume that the document has been separated into text, graphics and image regions, and we then further decompose the text regions [5]. The extraction of information units is related to the Gestalt principles, as discussed briefly in Section 2, and we rely on this in our approach to text segmentation. Proximity grouping of text is performed bottom-up to obtain a component hierarchy, and similarity grouping (boldface, italics and text size) and "good continuation" segmentation are then computed top-down. A description of our text segmentation scheme is given in [4].

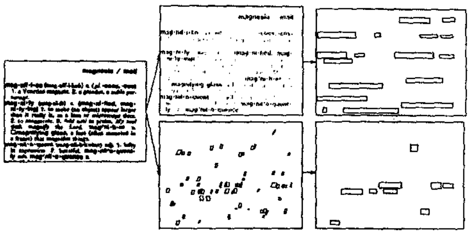


Figure 2: Boldface (top) and italic (bottom) word detection.

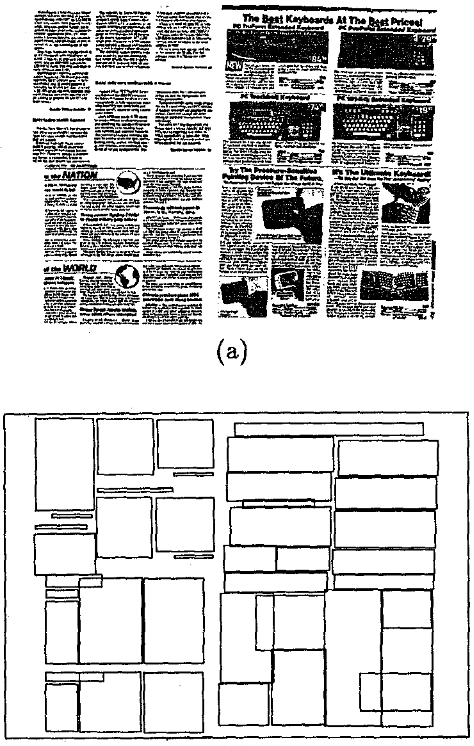
From the segmented text, a next level of characterization is based on information unit properties. First, a gross characterization of the text height is made for each block. The height of each line's bounding box is computed, and the average height of all the lines in all multi-line blocks is computed as the average text height, based on the assumption that multi-line text blocks are a good indication of the standard "body" text of a document. Text blocks are then characterized as large or small when they vary by more then 25% from the average.

Words are also identified as italic or boldface. Italic words are identified by the following algorithm. The minimum upright bounding parallelogram (i.e., a parallelogram with horizontal base and top) is constructed for each component and the slant measured relative to the vertical axis. Since it is difficult to make an accurate determination of the angle from short characters, symbols taller then the average are weighted more heavily. Words in which 50% of the characters have slants greater than $\delta$ degrees are classified as italic (Figure 2). We have used $\delta = 11$ in our experiments.

Boldface is also identified at the word level, but using a morphological approach applied to individual blocks (Figure 2). An opening transform is applied in an attempt to eliminate or severely distort non-boldface text. An erosion transform is applied until more than 80% of the pixels have been eliminated, at which point a dilation is applied for an equal number of steps. When the resulting image is compared to the original image, words which are not in boldface have very limited similarity to the original while boldface characters tend to remain intact.

## 4.2 Use Classification

As suggested in Section 3, the population of text blocks and their descriptions can be used to classify a



(a)

(b)

Figure 3: Browsing document segmentation

document into the usage categories of reading, browsing, and searching (and modifying).

The following heuristics can be used to identify these classes:

**Reading documents** are characterized by a relatively small number of large text blocks on each page. The majority of the document is composed of text that has a single point size.

**Browsing documents** tend to have medium to large text blocks, and small text blocks of a larger point size which act as focal points for the reader. Although readable documents have similar handles, browsable documents typically have many such handles.

**Searching documents** are characterized by small, repetitive text blocks.

Figure 3 illustrates the use of these criteria in the block-level segmentation of a browsing document. Some of the specific properties which can be used include: the number of text blocks, the distribution of the geometrical sizes of the blocks, the number of words and lines per text block, the geometrical arrangement of the blocks, the existence of multiple point sizes, and the existence of graphic and image components

Using a set of very simple criteria, based on a subset of the above properties, we were able to classify approximately 80% of a 100-document database correctly, with approximately 5% being unclassified. The criteria used were as follows:

- In a searching document, no more than 25% of the text blocks should have more than five lines. There should be no image components, and few or no graphic components.
- A browsing document must have at least three head/body pairs. A head is in an emphasized font (boldface, italics, or a large font) and has no more then two lines. A body is standard text with more then two lines.
- A reading document must follow a strict (one- or two-column) column structure and must have large text blocks, primarily of a standard point size.

These criteria will not perform well on very complex structures. One of the difficulties is that many documents belong to more than one use class. Consider, for example, the "yellow pages" of a telephone book. The individual line listings are clearly designed for searching, but they are intermixed with "advertisements" which have browsing characteristics. Similarly, a journal article's bibliography exhibits both reading and searching characteristics.

### 4.3 Type Classification

Type classification is a refinement of use classification; the type of a document refers to a more specific document-level characterization such as journal article or newspaper article, or a page-level characterization such as title or contents page.

We can use function-based analysis as a basis for type classification. As an example of how to perform classification at this level, we attempt to classify individual journal pages as being title, reference or body.

A set of 59 journal page images from the University of Washington English Document Image Database-I was used for training and testing. This database contains images of pages as well as page- and zone-level ground truth for each page. Each description includes general characteristics of the page and characteristics of each zone on the page. The page characteristics include, for example, "dominant-font-size", "dominant-font-style", and "number-of-columns", while the zone characteristics include, for example, "type", "location", "text-alignment", and "dominant-font-style". The classification of pages into the three categories was not provided in the ground truth, and was performed manually.

The complete database was converted to Document Interchange Format (DIF). In this format, each page is described by specifying general information about the page, and a list of zone descriptions.

To classify the pages, we used a small set of attributes of the zones. The most discriminatory attributes turned out to be the number of vertically

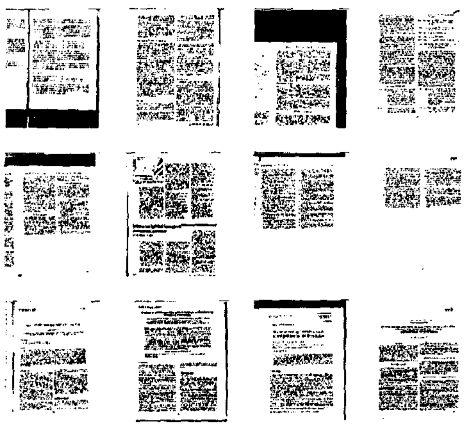


Figure 4: Pages classified as body (top), reference (middle) and title (bottom).

neighboring zones with consistent height and the average size of the zones.

Using rules based on these attributes we were able to classify journal page images with an accuracy of over 90%. The rules are intuitively plausible and highly consistent with our functional principles. The number and average size of the information units (zones) play major roles in the rules.

Examples of documents that were classified into each class are shown in Figure 4. Note that the second example of a reference page is also a title page.

### 4.4 Functional Enhancement

If we can decompose a document into functional components, we can use its functional organization to help decide which portions of it should be presented to the user and which can be ignored or considered as lower priority. The extraction of functional constructs allows this to be done without the need for content-level reasoning. Using these ideas,we can present document images to users in accordance with their goals. If a user wants, for example, to browse collections of documents, we can provide only the upper-level headers, and give the user the option to retrieve full information when needed. Examples are given in [4].

The pieces of a document which we choose to present are based on the observation that there appears to be a close analogy between the three modes of document usage and three methods of traversal of a tree structure. Reading a document corresponds to a depth-first search of the tree. We expand each node in turn and traverse the tree depth-first. Browsing resembles a pruned depth-first search; the reader identifies nodes at higher levels which are of interest, and prunes those which are not. Searching can be implemented by treating the tree as a decision tree; a node or set of nodes is explored at each level, until the one which contains the appropriate

## 5 Discussion and Conclusions

Document functionality relates to how the document conveys information to its user. In this paper, we have provided a basis for understanding the functional aspects of document design and usage. Authors use layout and emphasis to make it easier to extract information from documents. Traditional document understanding and conversion techniques have ignored the intended functionality of the document, especially its class-independent functional structure. An important advantage of our approach is that it provides an ability to organize documents without understanding their content.

We plan to extend our work to provide a more complete taxonomy of functional primitives, and to implement a full-scale system for functional typing and document classification.

### Acknowledgments

This paper has resulted from research supported in part by the US Department of Defense under contract number MDA 9049-6C-1250.

## References

[1] H.S. Baird. Anatomy of a versatile page reader. *Proceedings of the IEEE*, 80:1059–1065, 1992.

[2] H.S. Baird, H. Bunke, and K. Yamamoto. *Structured Document Image Analysis*. Springer, 1992.

[3] A. Dengel, R. Bleisinger, F. Fein, R. Hoch, F. Hones, and M. Malburg. Officemaid - a system for office mail analysis, interpretation and delivery. In *International Workshop on Document Analysis Systems*, pages 253 – 276, 1994.

[4] D. Doermann, E. Rivlin, and A. Rosenfeld. The function of documents. Technical Report CAR-TR-841, University of Maryland, 1996. To appear in IJCV.

[5] K. Etemad, D. Doermann, and R. Chellappa. Multi-scale document page segmentation using soft decision integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):92–96, 1997.

[6] International Standards Organization. *Text and Office Systems—Office Document Architecture (ODA) and Interchange Format*, 1989. International Standard 8613.

[7] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):737–747, July 1993.

[8] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1162 – 1173, November 1993.

[9] S. Liebowitz Taylor. Information-based document analysis systems in a distributed environment. In *International Workshop on Document Analysis Systems*, pages 93 – 108, 1994.

[10] T. Watanabe, Q. Luo, and N.Sugie. Structure recognition methods for various types of documents. *Machine Vision and Applications*, 6(2–3):163–176, 1993.