Intelligent Data Retrieval from Raster Images of Documents

Sargur N. Srihari, Stephen W. Lam, Jonathan J. Hull,
Rohini K. Srihari, and Venugopal Govindaraju

Center of Excellence for Document Analysis and Recognition (CEDAR), State University of New York at Buffalo, UB Commons, 520 Lee Entrance, Suite 202, Amherst, NY 14228-2567
Email: srihari@cedar.buffalo.edu, Voice: (716) 645-6164, Fax: (716) 645-6176

Abstract

Documents on conventional media, such as books, newspapers and microfiche, can be converted into the digital form of raster images by the use of scanners. A digital library is a server on a computer network that can respond to user requests by retrieving relevant data contained within raster image documents as well as in other formatted ASCII documents. The task is to automate the analysis of data contained in raster image documents for the purpose of intelligent information retrieval in a digital library.

The task is to develop computational theories and algorithms, with contributions to the fields of *document understanding* and *intelligent interactive information retrieval*. The limitations of a technology necessary to convert books to text-searchable form, *viz*., optical character recognition, will be addressed. Specific research agenda items: adaptive document understanding, robust page layout analysis, table understanding, intelligent text recognition, graphics analysis, topic categorization, content-based retrieval of captioned pictures and query processing for information retrieval.

Keywords: Document image understanding, OCR, pattern recognition, artificial intelligence, page layout analysis, document scanning.

Introduction

Advances in computer technology during the past decade have resulted in electronic means for generating, storing, retrieving and using data of all types. Yet there remain large volumes of documents that exist only in paper form, *e.g.*, books, or in the form of images of paper documents, *e.g.*, microfiche. It is now straightforward to convert such paper documents and their images into digital form by a process of scanning that yields a raster image of each page. In a raster image the reflected light intensities of discrete points on the page are stored as pixel values. Such images not only consume enormous amounts of storage compared to their formatted ASCII counterparts, but technology is seriously lacking for searching such documents for content (*e.g.*, keywords). In the future, many documents will continue to be available only in paper/raster form for reasons which include convenience (paper is easy to carry around), non-standardization of document preparation formats, and the global and contextual cues that images of paper documents provide to humans that are absent in pure ASCII text presentation.

This paper describes the project of automating the task of analyzing data contained in raster image documents for the purpose of intelligent information retrieval from a digital library [DL]. A DL is envisioned as a server connected to a computer network of users where the server has the ability to respond to user requests by retrieving relevant information from document data stored on an electronic storage medium. The specific research topics are determined by the architecture of the DL, which is shown in Figure 1. It consists of an off-line process wherein document data is captured and prepared for storage, and an on-line process where queries are received and processed for information retrieval. There are three major computational processes shared between the off-line and on-line processes of the DL: document data capture, data integration/indexing and information retrieval.

Three areas mentioned above are described more detail below.

Document Data Capture

Conversion of raster images of text into text-searchable form is the technology popularly known as optical character recognition [OCR]. AT present, although there are a half-dozen commercial software packages available, the technology is well-known still to be lacking [1]. Recognition problems occur when there are a variety of page layouts, fonts, print qualities and contextual differences. Such variations are commonly found in a typical library archive. A recent study suggested that having such capability in reading machines [2] is still decades away. Several

Proc. of Digital Libraries '94, June 19-21, 1994, College Station, TX, pp. 34-40.

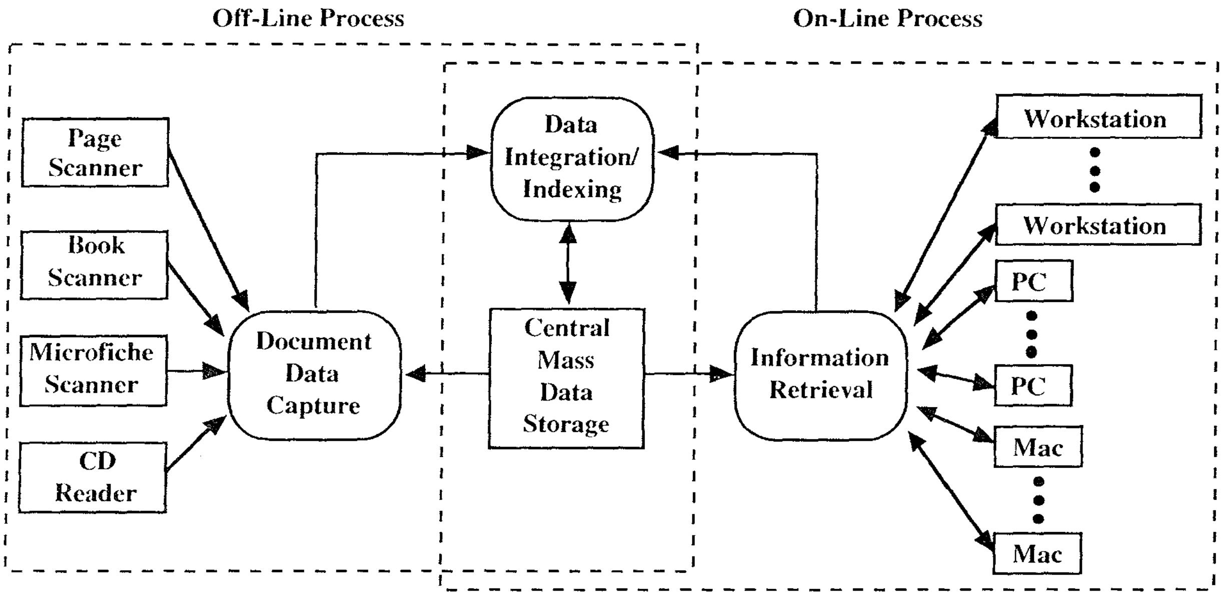


Figure 1. Major components of the digital library.

promising methods have been recently developed at CEDAR. This research will focus on improving the methods for processing large volumes of documents.

The algorithms necessary for conversion of raster image data into a form that supports intelligent retrieval can be categorized as performing one of three functions: analysis, recognition or understanding. Analysis algorithms map images to other images. Recognition algorithms map images to symbolic classes. Understanding algorithms map images to symbol structures. The following five topics are of importance for DLs: adaptive document understanding, robust page layout analysis, intelligent text recognition, tabulated data interpretation, and graphics recognition. Success in these tasks will allow a majority of documents to be processed within the DL.

Adaptive Document Understanding

A document understanding [DU] system begins with a raster image representation and obtains a high-level "useful" representation of document content. DU systems can be said to be either "closed" or "open," depending on whether or not they are designed for a particular document type. DU systems in use today are all closed systems, *i.e.*, designed for a particular document domain (such as forms, specific journals and business letters). Closed systems employ knowledge specific to the document type to be processed. They cannot be adopted easily to other types of documents.

An open system architecture has been developed at CEDAR. It is designed for multi-page documents, i.*e.*, it generates a logical interpretation of a document by combining information on different pages of the document. The architecture (see Figure 1) consists of three components: control, knowledge base and tool box. The *Control* possesses some general purpose document analysis strategies and the *Tool Box* contains a set of generic document image processing tools that are applicable to different documents. The system has no prior knowledge about document domain. The use of strategies and tools relies solely on the knowledge of the document of interest defined in the *Knowledge Base*. Since document-specific knowledge is not part of the system, it can be viewed as input (in addition to the document images) to the system. A prototype system based on this architecture has been developed at CEDAR to process a variety of documents such as forms, IEEE journals and postal mailpieces [3, 4]. The research focus is on how to use knowledge to adjust the system reading strategies for handling different types of library documents.

Page Layout Analysis

Layout analysis starts with block segmentation which decomposes the digital image of a document page into regions. The process locates large streams of white space (background analysis) running horizontally or vertically. Streams that are considered region boundaries are used to partition the image into regions. A region must be bound by two horizontal and vertical background boundaries. Criteria used when considering a white stream as region boundary are derived from the analysis of all the white streams in the page image. Non-boundary white streams, which are usually smaller in size, separate elements of the same region such as those between text lines. This global statistical approach does not need to predefine the size of the white streams, which varies from page to page. This local adaptive method increases the robustness of the segmentation process. Figure 2 shows the regions of a document after block segmentation.

The next step is to classify a region into one of the structural categories such as text, line drawings, tables and

Proc. of Digital Libraries '94, June 19-21, 1994, College Station, TX, pp. 34-40.

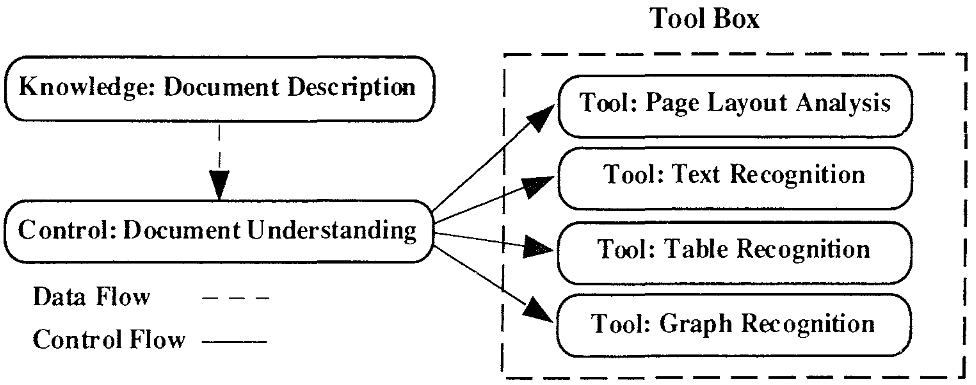


Figure 2. Components of an adaptive document understanding system.

photographs. The classification of a region is performed by matching a set of features extracted from the region against the predefined reference features of a category. This approach allows the addition of new categories only if they contain distinctive features. It is very flexible and is able to handle various types of documents.

Text Recognition Using IR Technology

The recognition of word images is a solution to text recognition by which images of text are transformed into their ASCII equivalent. Word recognition algorithms are an alternative to traditional character recognition techniques that rely on the segmentation of a word into characters. This is sometimes followed by a postprocessing step that uses a dictionary of legal words to select the correct choices.

Errors in the output of a word recognition system can be caused by several sources. When a noisy document image is input, the top choice of a word recognition system may be correct only a relatively low percentage. However, the ranking of the dictionary may include the correct choice among its top N guesses (N=10, for example) in nearly 100% of the cases.

Solutions to improving the performance of a text recognition system have utilized the context of the language in which the document was written. An observation about context beyond the individual word level that is used here concerns the vocabulary of a document.

A methodology is proposed to predict the vocabulary of a document from its word recognition decisions. The N best recognition choices for each word are used in a probabilistic model for information retrieval to locate a set of similar document in a database. The vocabulary of those documents is then used to select the recognition decisions from the word recognition system that have a high probability of correctness. Those words could then be used as "islands" to drive other processing that would recognize the remainder of the text. A useful side effect of matching word recognition results to documents from a database is that the topic of the input document is indicated by the ti-

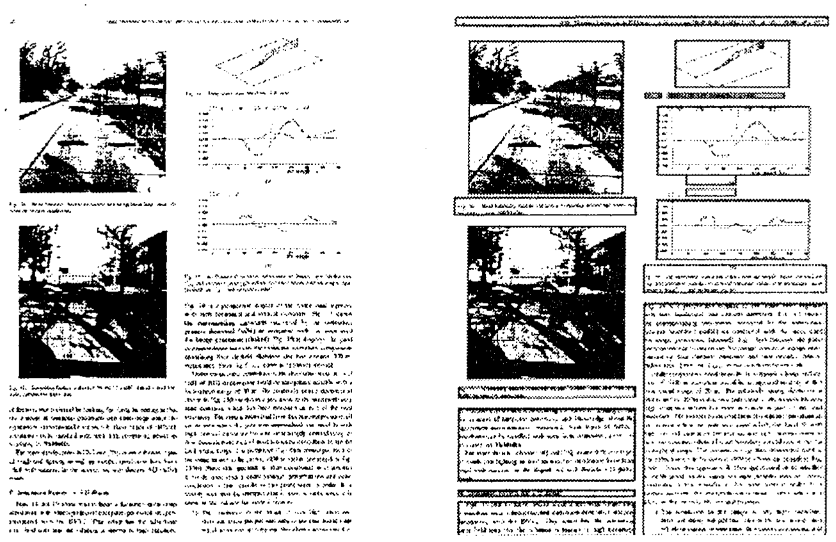


Figure 3. Result of block segmentation. *(a)* Original document image. *(b)* Regions located by the block segmenter.

Proc. of Digital Libraries '94, June 19-21, 1994, College Station, TX, pp. 34-40.

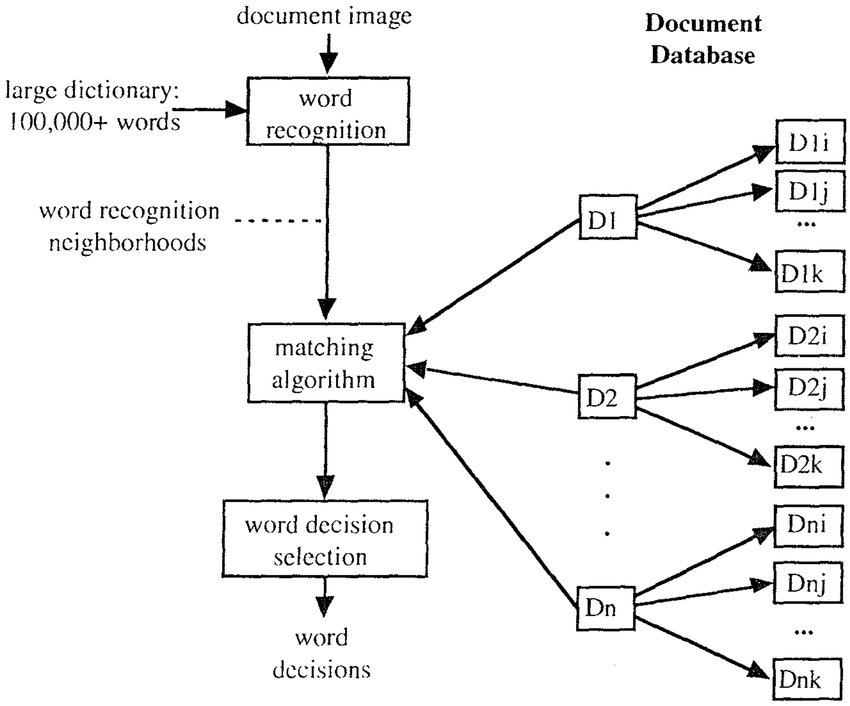


Figure 4. Word matching algorithm.

tles of the matching documents from the database. The algorithmic framework discussed in this paper is presented in Figure 3. Word images from a document are input. Those images are passed to a word recognition algorithm that matches them to entries in a large dictionary. *Neighborhoods* or groups of words from the dictionary are computed for each input image. The neighborhoods contain words that are *visually* similar to the input word images.

Tabulated Data Recognition and Interpretation

Tables are used in documents as an effective means of communicating attribute-value information pertaining to several data items (keys). The spatial layout of these items communicates the desired associations. In many cases, tables are intentionally composed so as to allocate the logical items and the relationships among items into the physical layout structures.

The goal of table understanding is to recognize the data items in the table and to derive a logical interpretation of the table based on contextual information of the data items and the physical layout of these items. The task of table understanding can be distributed over three tightly coupled modules: (i) alignment detection, (ii) layout understanding, and (iii) data entry recognition.

Graph/Chart Recognition

Input to this process is a graphic block and text from the caption. The objective is two-fold: (i) classify the graphic block as one of several graphic types: mechanical drawings, schematics, flow diagrams, intensity images, x-y plots, bar charts, pie charts, etc. (ii) decompose the graphic block into a set of primitives and capture the relationships between them (vectorization). Classification of graphics is necessary to keep the graphics processing streamlined and domain specific [5]. This knowledge can be obtained from collateral textual sources like the caption or derived from preliminary image processing as described below in brief.

Graphics classification can be effectively performed using shape based heuristics [6]. A histogram of greyscale values can discriminate between intensity images and binary line drawings. A pair of distinct peaks indicates an intensity image. Presence of circles, boxes, and diamonds indicates a flow chart or a schematic diagram. Further discrimination between classes can be accomplished by detecting presence of connecting lines between symbols. Global properties like location, length, and number of all the lines are also useful indicators.

Multimodal Data Integration/Linking

The task of multimodal data integration/indexing is to remove the physical barrier imposed on the document content. It creates a logical interpretation of the documents in the DL database by linking spatially and contextually correlated document elements together. The hierarchy can be constructed by following the hierarchical structure of the document description defined for document understanding. Two objects will form a linkage when they satisfy certain spatial or contextual constraints. For example, a spatial link is formed between a photograph and the text below, and a contextual link is formed between a paragraph of text which has an explicit figure callout (such as "see Figure 1") and the photograph and caption which are

Proc. of Digital Libraries '94, June 19-21, 1994, College Station, TX, pp. 34-40.

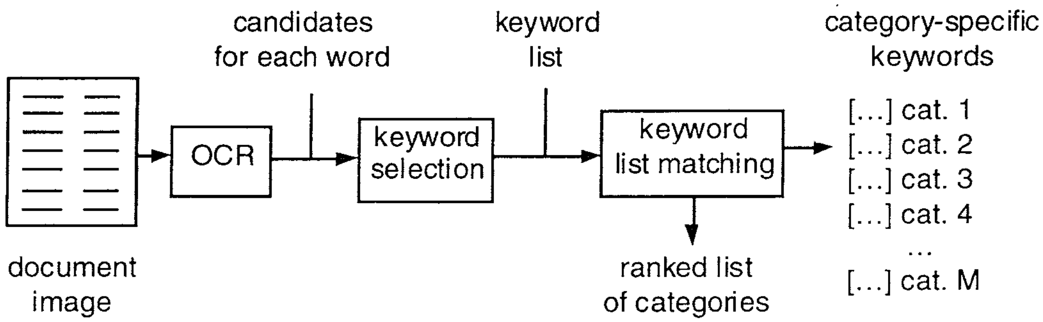


Figure 5. Topic categorization from OCR results.

indexed by that figure number. There are two types of links: (i) generic links — the links that apply to all documents and (ii) specific links — the links that are only applicable to a particular type of document. Specific links must be predefined at the knowledge acquisition stage.

Research will focus on developing techniques to automate the creation and organization of linkages in a way that will facilitate effective information retrieval. Research subtopics are: (i) topic categorization, (ii) picture-text integration, (iii) linkage detection.

Topic Categorization

Categorization of the topic of a passage of text is an important part of entering it in the database of the DL system. Given an input document image and a fixed number of categories, the objective is to assign the document to the category that it best matches. This problem is compounded by the imprecision that is present in OCR results (word decisions offered by the OCR may not be correct).

The following solution is proposed (see Figure 4). Top *N* decisions for each word are computed by postprocessing the OCR results. These words are then input to a keyword selection algorithm that chooses a fixed number of words that characterize the topic of the text [7]. Those keywords are then matched to lists of keywords that have been precompiled for each of the categories. The output is a ranked list of a small number of categories that contain the input document.

Research issues are choosing the keywords from the OCR results and matching the derived keywords to the category-specific keyword lists. Current work has shown that a methodology based on the Salton method for keyword selection [8] provides reliable performance in the presence of noise. This technique will be extended and other algorithms for keyword selection will be tested.

Content-based Retrieval of Captioned Photographs

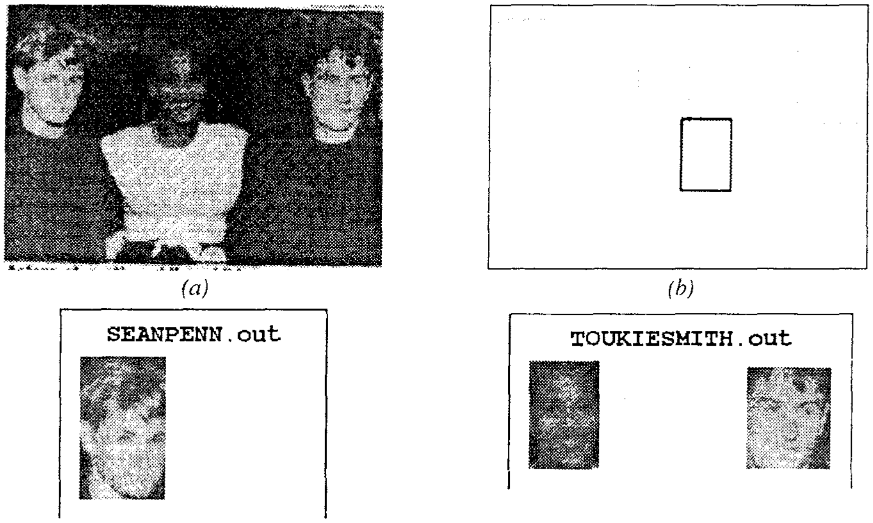
This research explores the interaction of textual and photographic information in document understanding. Information contained in both pictures and captions enhances overall understanding of the accompanying text, and often contributes additional information not contained specifically in the text. The research described here is designed to be applied at document processing time permitting more intelligent retrieval of photographs than is currently available. Specifically, the research has two main objectives: (i) natural language analysis of the caption at document processing time designed to extract significant attributes of the picture which can be used to automatically "tag" the pictures, and (ii) content-based retrieval of photographs.

PICTION: A Caption-Based Face Identification System: A prototypical system, PICTION [9,10], which understands photographs based on collateral text has been developed at CEDAR. More specifically, when given a text file corresponding to a newspaper caption and a digitized version of the associated photograph, the system is able to locate, label, and give information about objects which are relevant to the communicative unit. A common representation for the information content from both the picture and the caption is employed.

A key module in PICTION is the face locator [11], which locates human faces in arbitrarily complex photographs. PICTION is noteworthy since it provides a computationally less expensive alternative to traditional methods of face recognition in situations where pictures are accompanied by descriptive text. Traditional methods [12] employ model-matching techniques and thus require that face models be present for all people to be identified by the system; our system does not require this. In PICTION, spatial constraints obtained from the caption, along with some key discriminating visual characteristics (e.g., male/female, color of hair) are sufficient for eliminating false face candidates and correctly identifying faces. Figure 5 is an example of a digitized newspaper photograph and accompanying caption that PICTION can handle.

Linkage Detection

A critical aspect of the DL project research is data linking — the linking together of items within a document or between several different documents. Links will be used by the system to facilitate user navigation through the documents and to support system reasoning to retrieve items from the database. Important research issues include: (i) what should be linked? (ii) using the information from DU module, caption understanding module, and topic categorization module, how can link creation be automated as much as possible? (iii) how can the links be grouped into



(c) (d)

Figure 6. (a) photograph with caption "Actors Sean Penn, left, and Robert DeNiro pose with Toukie Smith, sister of the late fashion designer Willi Smith, at a New York celebrity auction Sunday in memory of Smith" (The Buffalo News, Feb. 27, 1989); (b) output of face locator; (c,d) output of PICTION.

concepts? and (iv) what relationships should be maintained between concepts?

The system cannot create and explicitly store all possible links between all low level items (for example, all numbers in a table). The storage required would be immense, many of the links would be rarely used, and processing time in following useful links will be slower. Instead, the system will use knowledge of structural units and a hierarchy of concepts to extract detailed linking information.

Three types of links have been identified: structural, temporal, and conceptual. All three will be stored external to the document. *Structural links* reflect structure in the original document and will be created from the output of the DU module. A structural link may point from the ASCII document text to a graphic located at that location in the original document. Links will be created from the table of contents, section indices, section headings, figures, figure references, figure caption, bibliographic references, bibliographic references, footnotes, and internal page references. A *temporal link* connects items by time. For example, all documents in a series produced over a five year period will be connected by a temporal link. Temporal links between whole documents are simple to implement once the time relationship is known. A *conceptual link* will connect items containing related information. Links may be internal to a document or may be between documents. Conceptual links arise from understanding of the content (caption understanding and topic characterization) and not exclusively from position in a document.

Information Retrieval

The IR module is the mechanism by which a user accesses the DL. Intelligence of the IR module is defined by its ability to retrieve correct information based on imprecise and insufficient user descriptions as well as imperfect OCR results in the document database. Retrieval systems can be evaluated by two quality measures: *precision*, the proportion of retrieved documents which are relevant, and *recall*, the proportion of relevant documents which are retrieved. Research in IR will focus on two areas: (i) NLP techniques for query analysis, and (ii) intelligent retrieval methods. Both of these will be incorporated into a graphical, interactive user interface which assists the user in accessing the DL. Experienced users will have the choice of entering query phrases directly (consisting of search terms and boolean operators). Other users will be able to express their queries in terms of English sentences. Much of the research in intelligent IR assumes a text only document database. The proposed research will consider the additional requirements posed by a multimodal (text, tables, graphics, photographs) database. Statistical methods such as vector-based models are proposed for use in the retrieval system. To date, very little improvement has been demonstrated by using NLP methods during the retrieval stage.

Conclusions

This paper presented three major processes which help to build a DL. Several components of the system are shown to be useful for creating a DL database. The design of these components stresses the importance of robustness and ease of adaptation to the processing of different documents. An adaptive approach to document understanding was presented in this paper. Its robustness was shown to be crucial to the success in processing varied library documents. This paper also presented an adaptation of the vector space model for information retrieval to improve

Proc. of Digital Libraries '94, June 19-21, 1994, College Station, TX, pp. 34-40.

the performance of the text recognition module. Output generated by the document data capture and data integration/linking modules facilitates the IR mechanism to produce intelligent response to user queries.

References

[1] S. N. Srihari. "From Pixels to Paragraphs: The Use of Contextual Models in Text Recognition". *Proc. Second International Conference on Document Analysis and Recognition*. Tsukuba Science City, Japan. October, 1993.

[2] G. Nagy. "What Does a Machine Need to Know to Read a Document?" Proc. Symposium on Document Analysis and Information Retrieval. Las Vegas, NV. April, 1992.

[3] S. W. Lam and S. N. Srihari. "Multi-domain Document Layout Understanding." Proc. First International Conference on Document Analysis and Recognition. Saint-Malo, France, September 1991.

[4] S. W. Lam and S. N. Srihari. "Frame-based Knowledge Representation for Multi-domain Document Layout Analysis," Proc. IEEE International Conference on Systems, Man, and Cybernetics, Charlottesville, VA, October 1991.

[5] S. N. Srihari. "Document Image Understanding." Proc. IEEE Fall Joint Computer Conference. Dallas, TX. 1986.

[6] R. Kasturi, R. Raman, C. Chenubhotla and L. O'Gorman. "Document Image Analysis: An Overview of Techniques for Graphic Recognition." Proc. IAPR Workshop on Syntactic and Structural Pattern Recognition. Murray Hill, NJ. July, 1990.

[7] J. J. Hull and Y. Li. "Interpreting Word Recognition Decisions with a Document Database Graph ." Proc. Second International Conference on Document Analysis and Recognition. Tsukuba Science City, Japan. October, 1993.

[8] G. Salton. Automatic Text Processing. Addison Wesley. 1988.

[9] R. K. Srihari. "PICTION: A System that Uses Captions to Label Human Faces in Newspaper Photographs." Proc. 9th National Conference on Artificial Intelligence. Anaheim, CA. 1991.

[10] R. K. Srihari. "Intelligent Document Understanding: Understanding Photos with Captions." Proc. Second International Conference on Document Analysis and Recognition. Tsukuba Science City, Japan. October, 1993.

[11] V. Govindaraju, S. N. Srihari and D. B. Sher. "A Computational Model for Face Location based on Cognitive Principles." Proc. 10th National Conference on Artificial Intelligence. San Jose, CA. 1992.

[12] R. Weiss, L. Kitchen and J. Tuttle. "Identification of Human Faces Using Data-driven Segmentation, Rule-based Hypothesis Formation and Iterative Model-based Hypothesis Verification." University of Massachusetts at Amherst, COINS Technical Report 86-53. 1986.

Proc. of Digital Libraries '94, June 19-21, 1994, College Station, TX, pp. 34-40.