LAMP-TR-0013
CAR-TR-878
CS-TR-3876

MDA 9049-6C-1250
February 1998

The Indexing and Retrieval of Document Images: A Survey

David Doermann

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 1998		**00-02-1998 to 00-02-1998**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Indexing and Retrieval of Document Images: A Survey	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**39**	

The Indexing and Retrieval of Document Images: A Survey

David Doermann

Language and Media Processing Laboratory
Center for Automation Research
University of Maryland
College Park, MD 20742-3275

Abstract

The economic feasibility of maintaining large databases of document images has created a tremendous demand for robust ways to access and manipulate the information these images contain. In an attempt to move toward a paper-less office, large quantities of printed documents are often scanned and archived as images, without adequate index information.

One way to provide traditional database indexing and retrieval capabilities is to fully convert the document to an electronic representation which can be indexed automatically. Unfortunately, there are many factors which prohibit complete conversion including high cost, low document quality, and the fact that many non-text components cannot be adequately represented in a converted form. In such cases, it can be advantageous to maintain a copy of and use the document in image form.

In this paper, we provide a survey of methods developed by researchers to access and manipulate document images without the need for complete and accurate conversion. We briefly discuss traditional text indexing techniques on imperfect data and the retrieval of partially converted documents. This is followed by a more comprehensive review of techniques for the direct characterization, manipulation and retrieval, of images of documents containing text, graphics and scene images.

The support of this research by the Department of Defense under contract MDA 9049-6C-1250 is gratefully acknowledged.

1 Introduction

The advance of office automation and onset of the information age is fundamentally grounded in the ability to create, manipulate, store, retrieve and transmit electronic documents. Word processing applications allow us to easily create and edit electronic documents, compression of the documents' electronic representation allows efficient storage and transmission, and the text analysis community has focused on the processing, indexing and retrieval of text from large repositories. For documents whose entire life cycle is electronic, these tasks are relatively straightforward, and users have therefore come to view documents as manipulable, searchable entities. Unfortunately, for those documents which were created manually, or for those documents whose electronic representation is simply not available, systems must deal, at least initially, with scanned images of the hard copy representations.

Although some work has been reported on the use and manipulation of individual documents in image form [4, 5], in the absence of complete conversion, there is far more potential in the topics related to document image databases including information extraction and indexing and retrieval.

1.1 Document Databases

The purpose of a database is to provide a wrapper through which one can retrieve information which is relevant to a given query. Depending on the nature of the data to be accessed (fielded, full text or image, for example) different techniques must be used for creating indexes, formulating queries and retrieving records. For traditional databases, data is maintained in a structured way and the contents of specific fields can be accessed directly. Queries can be formulated and records retrieved by indexing fields, indexing the results of calculations on fields or indexing relations among fields. For such databases, index information is typically derived directly from the data, but unfortunately, queries are often limited to accessing the specific fields and relations

which are defined a priori. For tabulated data in a well-defined, known environment, this may be acceptable.

Maintaining and accessing a database of full-text documents is a more challenging problem. For text document databases, the incoming data is typically less structured so it is difficult, and perhaps even impossible, to populate fields a priori. Nevertheless, similar retrieval goals can be defined. For example, given a large heterogeneous database of text, one may ask "How many votes did Bill Clinton receive in the state of Florida during the 1996 presidential election?". The ability to provide the user with an appropriate *answer* to this query is difficult, and it is quite different from the ability to simply *retrieve* documents which most likely contain the desired information. In general, answering such a question requires some basic level of comprehension of the content of both the query and the documents being stored in the database.

A realistic expectation for today's text document database systems is not to provide answers, but rather to provide mechanisms to retrieve documents which are most relevant to the formulated query, in ranked order. For example, in the above query, we would expect a system to return a subset of the collection of documents which contain election returns, and to mention Clinton and Florida. More desirably, we may hope to have documents weighted more heavily if the terms "votes" and "Florida" are used in close proximity. The text retrieval community has made significant progress in Information Retrieval (IR) and addressing related information processing problems such as topic clustering and information filtering [48]. Longer-term goals are of course to be able to answer such queries by returning a quantitative result when appropriate. This is clearly a research issue, however, and will require a more complete processing of the natural language than is necessary for retrieval. Although indexing and retrieval of text documents is a general problem, it is arguably much more difficult than indexing and retrieval from traditional databases. This is due to the fact that for full text databases, the text is rarely organized into meaningful fields and the (natural) language is far less constrained. On the positive side, however, the content

of the document is often represented directly by the text. When documents contain non-text regions to be indexed, or when the document does not originate in electronic form, the problem becomes more complex and a new set of issues must be addressed.

1.2 Document Image Databases

It has not been until recently that it has become economically feasible to deal with documents by maintaining large databases of document images. In an attempt to move toward a paper-less office, large quantities of printed documents are digitized and stored as images in databases, but often without adequate index information. One option for automatically generating index information is to attempt to convert the document to an indexable representation. Although text can be represented very efficiently in electronic form, low document image quality can inhibit accurate symbol recognition (OCR) and make it difficult to preserve stylistic features such as font and layout. Furthermore, it may not be possible to provide a suitable representation for certain scanned graphic and picture regions, other than as a digital image, primarily because there is no simple decomposition into concise units that can be easily recognized.

Scanning and storing documents as images clearly has advantages over storing hard copy, and unlike converted documents, digital document images can provide a precise, high-quality representation of the original document, including graphics and images. Having the ability to capture an accurate representation of both the content and structure is especially important considering the complex, multi-modal documents that can be created with today's desktop publishing systems.

Given that it is beneficial to maintain at least a copy of most documents in image form, the main issue then becomes the need for robust ways to access or "index" the information these images contain. With electronic text documents, the index information can be extracted directly from the text, but we do not have that luxury with document images. The "content" of document images is not directly available

since the internal representation is simply a set of pixels, so it is difficult to perform automatic indexing. For some archival applications a simple document ID, such as a case number, may be sufficient for indexing, and can be provided manually. Unfortunately, the rich content of many documents prohibits comprehensive manual entry of index information, especially for graphic and image regions and for document structure.

In this survey, we review techniques for content-based, automatic indexing and retrieval of document images, and highlight several topics related to the creation and maintenance of document image databases. There is clearly a spectrum of possibilities for indexing from partial conversion and recognition of text, graphics and images to the indexing of the pixels directly. Some approaches have been developed to avoid problems with low-accuracy conversion, while others have been developed because appropriate conversion is not possible. Our goal is to focus on an in-depth discussion of techniques which do not try to perform complete and accurate conversion and high-level indexing, but rather which attempt to characterize, index, and retrieve documents based directly on image features.

For completeness, in Section 2 we provide a brief discussion of techniques for classical indexing and retrieval of electronic text, as well as techniques for IR which operate on OCR'd text and work performed on the enhancement of OCR'd text for IR. In Section 3 we review some research which considers methods of characterizing, indexing and retrieving images of text without conversion to ASCII. In Section 4, we highlight methods for indexing at the document level, including retrieval based on physical and logical structure. In Section 5, we consider retrieval-driven indexing of graphics and in Section 6 we consider several specific applications related to document image database management, indexing and retrieval.

2 Analysis of Electronic Text

Accessing collections of document text is a problem that has been addressed by the information retrieval (IR) community for many years [47, 48]. For applications such as the analysis of newswire articles or the analysis of other sources generated electronically, significant progress has been made. For much of that time, however, it has been assumed that the systems would deal exclusively with clean and accurate data. Not until recently have techniques been developed to deal with noisy data, such as text transcribed from speech or text recognized from document images. The general consensus of the community had been that given sufficient computational resources, the text in document images could be recognized and converted so that standard retrieval techniques could be applied. For certain domains this is true, but in general, the lack of structure in recognized or converted documents, combined with the often sub-standard accuracy of the conversion process, makes converted documents difficult to index. Nevertheless, many of the lessons learned from classical IR will influence content-based document image IR as the field matures.

2.1 Classic Indexing and Retrieval

Many documents which are created electronically have both structured and unstructured components. For example, electronic mail (EMail) delineates the sender, receiver, date and subject, in addition to the message. It is useful to distinguish between document indexes which rely on *objective*, structured identifiers, such as authors' names, titles and publishers, and *non-objective* identifiers which are extracted directly from the text content [47]. If the document analysis front end provides objective identifiers, standard database operations can be used to query document databases. In the absence of objective identifiers, methods for characterizing the full text content of the converted documents must be developed. The latter is a more challenging problem, however, and where most research is being carried out.

Research in IR has led to the use of a wide variety of techniques for automatic indexing. Historically, experts have been called upon to manually provide a concise index of content descriptors for each document. More recently, techniques such as inverted indexes, term weighting and its variations, and the extraction of relationships between terms to preserve content, have evolved to provide advanced automatic indexing. For retrieval, vector space, probabilistic and boolean retrieval models provide a foundation for document similarity.

The basic idea behind most approaches to text indexing and retrieval is to provide the ability to characterize the text corpus in a meaningful way, to allow users to provide a query as a set of terms, and to provide mechanisms to retrieve, in ranked order, the most relevant documents for that query.[1] One common way of characterizing a document's content is to consider the full text, filter out common "stop" words which have a negligible effect on the content, and then represent the document by a term vector consisting of the frequencies of meaningful terms. Furthermore, suffixes can be stripped (stemming), low-frequency words replaced with thesaurus equivalences, and high- frequency words replaced with phrases, in an attempt to reduce inappropriate variability in the text.

Trends in the documents' content can be compensated for by weighing the term frequencies by the inverse document frequencies (tf-idf). That is, for a document i and a weight w_{ij} for term j, the term frequency tf_{ij} is multiplied by the inverse of the number of times the term appears in the entire collection. This is given by

$$w_{ij} = tf_{ij} \log \frac{n}{df_j} \tag{1}$$

In this way, a collection which contains the term "president" in every document, for example, would not be assigned a very high weight to that term even if it it occurs often in a given document. Since the term occurs in a large number of documents,

[1] Of course more complex approaches which consider the natural language are of great interest, but beyond the scope of this paper.

it is not likely to provide significant discrimination power. On the other hand, if the term appears often, but in only one document, it would receive a much higher weight. In document image databases, similar techniques must be developed to filter out irrelevant index information and weight features appropriately.

Once the documents are indexed, the resulting index vectors can be considered as signatures and used for retrieval. To query the collection, a simple measure can be used to compute the "distance" between the query vector and the document vector. A straightforward choice is cosine or inner product similarity, but many others exist [47].

The basic idea of providing a characterization and a similarity measure is essential for all retrieval problems, including the retrieval of document images.

2.2 Processing Converted Text

As alluded to previously, a desirable goal of text retrieval is to be able to operate effectively on noisy data, such as the output of an Optical Character Recognition (OCR) system. Toward this end, the Text Retrieval Conference (TREC), co-sponsored by NIST and DARPA, provides a forum for the evaluation of how well retrieval systems perform on recognized text. In 1996, the conference held a confusion track where the test data was obtained by printing, scanning and recognizing (via OCR) data from the Federal Registry [27]. Given ground truth for the documents, the organizers were able to characterize the character-level error rate at approximately 20%. Systems which participated in the evaluation were provided with the original and converted text, a set of queries and a set of associated relevance judgments, allowing the calculation of the effective recall and precision measures. They were asked to run their algorithms on both corrupted and uncorrupted data. Overall there appears to have been a significant drop in accuracy between the results from the corrupted and uncorrupted data. More detailed tabulation of the results is not currently available, but will be reported at a later date.

A number of attempts have been made to bring the document analysis and information retrieval communities together to work on the retrieval of noisy data. From 1992 through 1996, the University of Nevada, Las Vegas, held an annual Symposium on Document Analysis and Information Retrieval [1]. The goal of the Symposium was to offer a strong program of basic research in the two complementary fields of information retrieval and document analysis, and did indeed promote work on problems associated with both areas. In 1994, Croft et al. [16] carried out IR tests on simulated OCR data, and found that for high-quality conversion, there is little effect on retrieval performance. For low-quality and short documents, however, a significant degradation in retrieval performance can result. Similar results were reported by Myka and Guntzer [44].

In more recent work in the same area [40], Lopresti and Zhou evaluated the performance of several classical and enhanced IR models using simulated OCR data. To enhance traditional IR models to deal with the imperfect data, they used approximate string matching and fuzzy logic. In general, they were able to show that the new methods are more robust to noisy data than the original methods, suggesting that simple enhancements can be used to improve performance.

Ohta et al. [45] described a system for full text search in which they augment three probabilistic text retrieval methods with knowledge about expected OCR errors. The approach used confusion information for specific characters, along with bi-gram probabilities of character occurrences to create multiple possible search terms for each initial search term. After performing the search with each new term, the validity of returned documents is based on the confusion and bi-gram occurrence probabilities. The results claim increases from 2-3% in recall with decreases or 4-5% in precision. Fujisawa and Marukawa used a similar approach in which they use confusion statistics to generate an enhanced finite state machine for query terms in Japanese text [25].

An alternative approach to attempting to modify the query to deal with poor quality is to modify the matching algorithm, as described by Takasu in [63]. To

obtain speed, the approach uses a two-stage algorithm where the first stage uses a fast string matching algorithm to generate match candidates, and the second stage uses a more appropriate similarity distance measure, such as the Levenshtein distance. As with [45] this approach shows slight improvements in recall, with slight decreases in precision.

All of these techniques improve recall marginally, but are only applicable for text recognition rates above 80%. For rates lower then 80%, recall and precision fall drastically. For such cases, systems are forced to either clean up the data, or resort to other retrieval techniques.

One system for processing and enhancing noisy data is presented by Taghva et al. [62], who describe an expert system for automatically correcting OCR errors by post-processing the text, prior to subsequent retrieval by an IR system. Rather than trying to identify all misspelled words in the text (which is typically attempted by the OCR system anyway), the system focuses only on words that will likely be used for retrieval, under the assumption that "significant" mis-recognized words will be used elsewhere in the same document. The system combines special recognizers to identify acronyms, proper nouns and garbage strings. It uses stop lists, followed by pre-clustering to determine indexed terms and insure their correct spelling using a dictionary. Overall the system claims to perform well in "cleaning" OCR data, primarily because it focuses on areas where the advanced IR systems tend to be sensitive.

In 1995, Doermann and Yao [23] presented work on a system for modeling errors in the output of OCR systems. They described a set of symbol and page models which are used to degrade an ideal text by introducing errors which typically occur during scanning, decomposition and recognition. The advantage of their system is that performance of text analysis systems can be evaluated under a wide variety of controlled conditions at relatively low cost, simply by varying parameters of the model.

On the application side, several interesting approaches have been proposed for retrieving document images, using OCR'd text. Most notably, Cornell University was faced with the challenge of how to get all of their technical reports on-line. It was decided that the cost of manually correcting automated OCR results was prohibitive, while the quality of the resulting transcription was too poor to be presented to the end users. They decided to use the imperfect OCR results to create indexes to the image, and have searches applied directly to the recognized text, transparent to the user [39]. The approach was motivated by the observation that statistical methods in information retrieval do not need perfectly clean data to work well. The retrieved documents were then presented to the user in image form. Given a retrieval mechanism which is robust to OCR errors, this represents a reasonable compromise with complete conversion for content-only retrieval.

Of the few other results reported, Tsuda et al. [67] developed and evaluated an approach to document clustering for browsing document collections. After performing tests on *ideal* data, they performed the same experiments on corrupt data. The experiments measured the level at which "term loss" (i.e. terms which were discarded since they were not valid English words) would affect the results of document clustering. The term loss was computed for various level of OCR accuracy as 22.7% for 85% character accuracy, 14.4% for 90% character accuracy, 4.3% for 95% character accuracy and 0.5% for 99% character accuracy. Using these datasets, they found that clustering could be done effectively down to approximately 85% character accuracy.

In summary, the community finds that there is a direct correlation between accuracy and retrieval performance metrics. For OCR accuracy in the range of 70-80% n-gram techniques have been shown to perform well. From 80-95% accuracy, enhanced IR techniques work well and above 95%, most IR techniques are completely unaffected by errors.

For the remainder of this paper, we will concentrate on reviewing literature associated with direct access to the document image, citing other work where appropriate.

3 Indexing Images of Text

In order to perform retrieval on document images which contain text, there must be a way to characterize the document content in a meaningful way. Researchers have addressed a number of problems ranging from attempting to identify proper nouns (Section 3.1) to automatic image abstracting (Section 3.3). Such approaches are especially appropriate for indexing low-quality documents where the recognition results are expected to be poor, or in filtering scenarios such as copiers or fax machines where the image data may not be kept internally.

3.1 Characterizing Text

DeSilva and Hull [53] have addressed the problem of detecting proper nouns in document images. Proper nouns tend to correspond to the names of people, places and specific objects and are valuable for indexing. It was noted that it is very expensive to run contextual post-processing on the recognized data which may have to identify over half a million surnames, not to mention the problems associated with word recognition itself. In their approach, DeSilva and Hull segment the document image into words and attempt to filter proper nouns by examining properties of the word image and its relationship to its neighbors. The post-processing would then be done on the resulting, much smaller, set of words.

Using a large text corpus, they developed and tested a set of seven features including capitalization, location in a sentence, length of word, length of the previous word, length of the following word, syntactic category of the previous word and syntactic category of the following word for proper noun identification. For capitalization, it was noted that over 95.5% of proper nouns are capitalized, and that over 35% of capitalized words are proper nouns. Furthermore, of the capitalized words which occur at the beginning of a sentence, only 10% are proper nouns and proper nouns tend to be longer than other words on average by over one character. Examining the word

lengths of neighbors shows that over 85% of capitalized words following a one-letter word (70% for two-letter words) are proper nouns. Further observations were made that proper nouns tend to follow prepositions, tend to be followed by short words, and tend to be followed by nouns.

To extract the necessary features, algorithms are presented for detecting capitalized words using baseline and character shape information, and classifying them. In experimental results, nearly 90% of proper nouns are consistently correctly identified. By performing this pre-classification, it becomes possible to add information from the image which is not available in the recognized text, and greatly reduce the complexity of the post-processing steps.

This work is of particular interest because it demonstrates that features are indeed present in image-based representations which may not be available in converted or electronic text.

3.2 Keyword Spotting

A second problem which has its roots in traditional IR is that of keyword spotting (or keyword searching). Some interesting work has been done on the problem of searching for keywords in document images using only image properties. Keywords are valuable indexing tools and if they can be identified at the image level, extensive computation during recognition will be avoided. All of the approaches described below rely to some extent on properties of a word's shape that can be stable across fonts, styles, and ranges of quality.

The approach of Chen et al. [10] is segmentation- and recognition- free, and has applications in the information retrieval domain when boolean models are used, as well as for information filtering. Chen first identifies candidate lines of text using morphology and extracts shape information from normalized lines of text. The upper and lower contours of each word are identified and used, along with the autocorrelation between columns of pixels, as input to a Hidden Markov Model (HMM). A keyword

HMM is created from a series of appropriate character HMMs and a non-keyword HMM is based on context-dependent sub-character models. Viterbi decoding of the HMM is used to identify the keywords and keyword variations in the line. Experiments show detection rates of over 95% on a restricted domain (lower-case letters) with a false alarm rate of approximately 1%. A subsequent paper [11] describes a simplification of this process which uses vertical character alignment information and a single model for each character. Related components of the system are described in more detail in [12], [13], [14] and [38].

DeCurtins and Chen [18] use word shape information and a voting technique to perform matching of keywords, also without segmentation. The approach is based on features including blanks, horizontal strokes, vertical strokes, ovals and bowls extracted from a contiguous line of text. A model is constructed a priori that specifies which words should contain which patterns. At runtime, the same features are extracted from a line of text and a voting scheme provides keyword hypotheses. The approach is based on the observation that the strokes of touching and corrupted characters often remain relatively unaffected by degradation, so it is especially well suited for noisy documents.

In [66], Trenkle and Vogt describe a preliminary experiment on word level image matching. In their approach, a query term is expanded to include its variations, and an image is generated in each of several fonts and with both lower- and upper-case characters. For each keyword image, the number of characters is estimated and features are extracted for the baseline, concavities, line segments, junctions, dots and stroke directions. The retrieval phase consisted of two steps, a filtering which eliminates candidates based on string length, and a matching which computes a distance metric for each keyword/candidate pair. Several experiments were performed which showed recall rates near 90% for low-resolution data.

In 1988, Tanaka and Torii [65] described a Transmedia Machine and its ability to search images of text. The system identifies word level components from the image,

and codes them using two bits per character corresponding to whether or not the character is an ascender or descender and whether or not it crosses the mid-line exactly once. When a query is presented as a set of keywords, a similar encoding can be done and matched against the codes of the original document. The approach is clearly sensitive to font, and does not provide sufficiently unique identification of words. They extended the original code to treat ascenders and descenders separately, and experienced marginal improvements in uniqueness. This appears to be the first use of what is now commonly referred to as shape coding.

In later work, Spitz [57] used similar yet more robust character-level shape information to map individual symbols in the image to a set of character shape codes which is much smaller than the original ASCII set. These shape codes form word tokens which can be indexed by the mapping of the desired keyword into the same space. Since the shape information and reduced dimensionality mapping can presumably be obtained with greater accuracy and at a lower cost than OCR, it has advantages over conversion, especially for some applications dealing with degraded images. At query time, the query text is mapped to a shape token using the coding obtained directly from the ASCII via a lookup table. The resulting token is mapped to the shape-code representation of the text. Inherent conflicts in the word tokens do not tend to affect the overall precision and recall rates dramatically, as there is sufficient redundancy in the English language.

In other work Spitz further exploits the English-language redundancy to perform OCR based on these robust shape codes, which can in turn be used to refine and enhance retrieval [56]. In [55], he demonstrates the use of shape coding in a working filing system for scientific papers.

In [43], Manmatha applies the word spotting paradigm to the indexing of images of handwritten archival manuscripts. The claim is that for a given document, the variation in word style is small since it was produced by a single writer. The approach segments the page into words and calculates equivalence classes from matches be-

tween words. The most difficult aspect of the problem is, of course, matching. Two matching algorithms are explored, one based on a translation-invariant Euclidean distance mapping and the other based on an affine-invariant word transformation. Preliminary results produced precision rates between 80% and 90%.

At SRI, some work was done by DeCurtins on evaluating the performance of OCR vs. Word Shape recognition for keyword spotting. Their system, Scribble, uses shape coding as described previously. The basic problem was to classify pages as being "interesting" or "non-interesting" based on occurrences of the keywords. When comparing based on overall keyword spotting accuracy and accuracy as a function of image degradation, Scribble outperformed the commercial OCR system consistently by more then 5%.

Each of these techniques provides some level of robustness to noise and the inability to OCR the documents reliably. Furthermore, these techniques are applicable because of the of tremendous redundancy embedded in words of the English language.

3.3 Automatic Abstracting of Images

A third topic of interest is that of automatic image-based abstracting. Although the problem of automatic creation of abstracts has been studied for many years [42], it had not previously been attempted on document images. Unlike keyword spotting, abstracting must develop a model for internal consistencies within the document, and extract relevant text which accurately summarizes the content.

Chen et al. [9, 8] provide a system for selecting portions (phrases or sentences) of an imaged document for presentation as a summary. The work is modeled after abstracting techniques used on ASCII text, but performing the task in the image domain, presents several interesting challenges. For example, in text abstracting, stop words are typically ignored, but this is difficult to do so without recognition. A second problem is word equivalences and how to identify them. Font size and imaging conditions make identifying different instances of the same word difficult.

The general approach is to extract word images and cluster them independently of their meaning. Using statistical characterization of the words and their locations in the document, stop words are identified and summary sentences and key phrases are extracted. Since word images cannot be compared to a stop list, words which behave like stop words are identified. This is accomplished by using a combination of the word width (stop words tend to be short), frequency (stop words tend to occur often) and location in the sentence. Sentences which contain a significant number of frequently-occurring keywords are chosen for the summary. Overall, the work presents a new and very interesting approach to image-based summarization.

Related work is described by Doermann et al. [20], who motivate the use of functional attributes derived from a document image's physical properties to perform classification and to facilitate browsing of document images. First, they identify salient regions which presumably correspond to, for example, titles, abstracts, and index keys, based only on image structure and zone properties. The document is then classified as intended for reading, browsing or searching using the distribution of functional units. Reading documents tend to have few titles and larger content blocks; browsing documents tend have a larger number of head/body pairings; and searching documents tend to have a larger number of smaller, similar-sized blocks. Based on this classification, an approach to browsing the document images is suggested by presenting the most salient regions in a limited hierarchical manner. They claim that if the document is presented as a logical tree, reading a document is similar to a depth-first search, browsing is similar to a pruned depth-first search, and searching a document is similar to a decision tree. The reader can then traverse the document structure at will. This work may lead to progress on a significant problem of document image databases, namely the inability to effectively browse or skim multi-page documents.

Both of these efforts make unique contributions because of the way they view document images as entities which are to ultimately be interpreted by humans. The

ability to organize and/or reduce the amount of information presented to the user in essential for navigating large collections.

Moving away from direct access to text content, we find that a signficant amount of work has been done in which relies heavily on the document's structure.

4 Indexing Document Structure

At the structural level a majority of the work which has been done relies on the output of a document analysis system. This is not surprising considering the relatively free-form nature of today's structured documents. There are essentially two ways to index a heterogeneous document collection based on structure without text content, using physical (layout) structure or using logical (semantic) structure.

In a typical document decomposition process, the first step is to perform a physical segmentation of the document. The physical segmentation provides only information about the physical characteristics of the markings on the page. This is typically followed by functional and logical labeling of constituent components. The functional labeling provides important insight into the general use of specific physical constructs, and the logical labeling provides a description of the document's semantic components based on a priori models for the document's class. A few selected articles are cited, primarily based on their ultimate goal of retrieval without transcription.

Section 4.1 addresses approaches which rely at least in part on segmentation, and Section 4.2 describes two approaches to categorization and retrieval using texture measures.

4.1 Segmentation-Based Approaches

One example of structural indexing is found in work by Herrmann and Schlageter [28]. Noting the lack of non-text retrieval mechanisms, they use traditional document analysis techniques to populate a relational database and propose a layout editor to

form queries. They are motivated primarily by the fact that personnel filing systems are often accessed using layout knowledge. The retrieval components are based on layout and are intended to supplement full-text search.

Takasu et al. [64] describe a method using model-based segmentation for incorporation of specific journal references into an on-line database. They process images of the table of contents and use a decision tree classifier based on physical features of the segmented blocks to label tokens. The process can be used to identify regions of the image for subsequent indexing.

Jaisimha et al. present an overview of text and graphics retrieval systems in [35]. Their system allows keyword searches on raw OCR results but does not provide any mechanism to deal with the highly degraded documents the system claims to handle. After manual segmentation, the system does allow image-based matching of similar logos and suggests that similar capabilities are available for signatures. On the interface side, it allows users to construct queries visually, and provides mechanisms for visual feedback.

In other work, Doermann et al. [22, 52] provide a framework for general document image retrieval using partial results from document analysis including zone type and zone attributes. In particular, the system introduces the concept of layout similarity and provides an approach to defining structural similarity between the physical components of documents. The system is based on the physical locations of zones and a high-level characterization of zone properties. The approach uses a weighted area overlap measure between regions on the page in an attempt to retrieve based on visual consistency between pages. Examples of its use are shown for the retrieval of documents which have similar column structure and similar layout of graphics regions. They also use the similarity measures to identify "title-like" pages by image example.

4.2 The Use of Texture

Texture features have been used extensively for page segmentation and zone classification [34, 68, 33, 15, 24] and have recently been used for image categorization and retrieval.

Soffer [54] attempted to extend the concept of n-grams (used extensively for noisy text data) to images by extracting NxM-grams of image intensity patterns. The technique essentially codes any given image as a set of small feature vectors (3-3), and uses a histogram of vectors to match against a database. While the approach is efficient, the simplicity of the feature vector suggests that it will only be able to distinguish between very gross texture features. In the document domain, the approach was demonstrated on music scores and various classes of English and Hebrew writing.

Cullen et al. [17], briefly describe a system which uses texture features based on the distribution of feature points extracted using the Moravec operator. The premise is that queries should be supplemented with information about how the user perceives the layout.

Both of these are examples of techniques that may be useful for a first-pass filtering of the database.

5 Indexing and Recognizing Graphics

For graphical documents, almost all of the work relies on at least partial recognition. The reason for this is that even for well-defined classes of engineering drawings, it is difficult to break the drawing up into unique constituent components that can be indexed robustly.

In this section we divide our survey between indexing drawings and maps (Section 5.1), which tend to be less constrained, and recognizing and indexing logos (Section 5.2), which can often be treated as a classical pattern recognition problem.

5.1 Drawings and Maps

Lorenz and Monagan [41] describe a system which indexes both line-level and textual features to retrieve machine part drawings. The text is recognized and traditional IR retrieval techniques are used. The drawing features extracted include parallel lines, junctions and adjacent lines. *Feature frequency weighting* is used (similar to the document term weighting described in Section 2.1) for indexing and ranked retrieval. Given a query by example, the approach can be used to index across heterogeneous collections of documents.

Syeda-Mahmood [61] also describes a system for content-based selection of technical drawing images without conversion. The approach is based on a selection and recognition paradigm where key regions are first selected and a model-based recognition engine confirms the existence of a part using pose estimation. Curvature and text features are extracted from regions in the image, without regard to individual components. During selection, a query image is used and features extracted are matched against the database. The approach benefits from the ability to process drawings without detailed knowledge of or models for the drawing's content.

Koga et al. [37] use a structural approach to index graphs. They attempt to identify types of graphs by their axes and combine classification with label information for content-based retrieval. In a more general image retrieval environment, Gudivada and Raghavan [26] incorporate spatial constraints in their query mechanism and demonstrate the approach on iconic images.

In much earlier work on maps, Amlani and Kasturi [2] described a query language for indexing images of paper-based maps. These problems differ from other document retrieval problems because of the attention that must be paid not only to component symbols, but also to their spatial relations on the page. Samet and Soffer [49, 50, 51] also consider the map retrieval problem by indexing the legend of a map to provide insight into the map content. They allow the user to visually create spatial queries using the legend symbols and retrieve map locations which satisfy the query

constraints. Although the database is populated semi-manually, the use of spatial indexing of recognized features is noteworthy.

Despite the efforts made in this area, it is clear that even if perfect conversion of graphic entities to CAD-based representation were achievable, we would still require complex interpretationany to achieve any signficant level of retrieval capability.

5.2 Logo Recognition

Recently, the problem of logo recognition has received a great deal of attention, primarily as an effective way to index or cluster documents such as letters or memos based on the identity of the originating institution. The fact that logos tend to be unique to a given institution, and that they remain relatively stable over long periods of time, makes them one of the few graphic components that can be indexed in this way.

In [21], Doermann et al. provide a unique multi-level approach to indexing logo databases. The approach uses text and contour features to prune the database using recognized characters and shapes. Similarity invariants are computed for unrecognized contours to obtain a more refined match. Since the logos are extracted from document images, the local invariant features are used to overcome the inherent problems with both over- and under-segmentation and to deal robustly with small changes in structure. The method allows the system to deal robustly with all affine transformations of the logo, as well as partial occlusion and mis-segmentation. Indexing can also be used to identify similar logos in a logo database.

Jaisimha [36] observes that complex logos, especially on faxes, can be highly degraded and that traditional recognition schemes are inferior. Their first stage normalizes the logo and in the second stage uses a wavelet projection signature to represent the logo. A distance measure between the signatures is used to identify and rank similar logos. The approach appears robust over a wide range of degradations.

Recently, Suda et al. [60] took a pattern matching/signal processing approach to

logo and word recognition by using low-pass filtering and matching, while Cesarini et al. [7] used a back-propagation neural network for logo recognition. A number of related problems including oriental seal identification and road sign identification have also been described in the literature and may be useful for indexing.

6 Related Topics

Several topics directly related to document image indexing are worth mentioning to complete this survey. One is the problem of finding or matching instances of a document image with known content in a large document database. This problem has applications both for maintaining database integrity by eliminating duplicates and for retrieval itself. The second is the use of features across different document components as demonstrated by the use of caption information for indexing images. Today's documents are clearly multimodal, so making use of all cues is essential to approach human judgments of relevance. The final topic is the that of document image compression. It has been shown by several researchers that document-specific properties can be exploited for compression, and the resulting compressed representation can be operated on directly by document analysis processes.

6.1 Document Image Matching

In [32], Hull first addresses the problem of content-based matching and describes a method for matching documents which have the same character content but which may have been reformatted or distorted prior to re-imaging. The approach represents each document by a set of robust local features which can be used to hash into a database of descriptors. The features extracted from both the query example and database must be invariant to geometric distortions. By extracting multiple descriptors for a given document, the index can be made more robust to errors in feature extraction. The measure of similarity used is simply the number of features the query

document and the database instance have in common. Experiments were performed using the character count for each word in short sequences of words, providing a set of simple yet robust descriptors for small databases. With as few as ten descriptors, 100% accuracy was obtained for a clean query string and a small clean database.

In other work [31, 30], Hull used pass codes generated from CCITT group 3 or 4 compressed fax images to attempt to match document images. Using a one inch square region of the image, the pass codes are extracted from the skew-corrected compressed image and matched against each previous document using a modified Hasudorff distance. On a database of 800 images, the system was over 95% effective in identifying duplicates.

Similar work [19] describes a method for detecting duplicate documents in very large image databases using only features extracted from the image. The method is also based on a robust "signature" extracted from each document image which is used to index into a table of previously processed documents. To obtain a signature, the page image is divided into small strips and successive strips are processed to identify candidate line segments. Lines near the top of the page which may contain running heads are avoided and a representative line which contains a sufficient number of symbols is extracted. A 3-bit shape alphabet is used to generate a set of keys for indexing via shape coding. Overall, the system demonstrates recall rates of greater than 90%, even for noisy data. The system is able to deal with differences between scanned documents such as resolution, skew and image quality. The approach has a number of advantages over OCR or other recognition-based methods, including speed and robustness to imaging distortions. A unique advantage of this method over related methods is that it uses an indexing approach (as opposed to matching), and can easily be scaled to millions of signatures.

6.2 Indexing Image Captions

Another topic is indexing image captions and attempting to establish a relationship between the caption's content and the picture it describes. It has been observed that the caption can be a valuable tool, both as a method to retrieve relevant images, and for image interpretation.

Srihari makes use of the association between an image and its caption to attempt to do recognition and content-based retrieval [58, 59]. In the PICTION system, she attempts to parse the caption and extract relative location and name information. Face candidates are identified automatically in the image, and the caption information is used to label them. Using part-of-speech tagging, natural language processing and domain knowledge, the system is able to make use of gender differences in the language as well as spatial relationships ("pictured from left to right") to establish a correspondence and effectively recognize the images.

In the MARIE project [46], Rowe uses a semantic network to address a problem similar to Srihari's, but in a domain where the captions are from military photos, can be very noisy and may require high-level domain-dependent knowledge to interpret. Since the system extends the work beyond the domain of faces, more complex shape indexing is required.

Future image retrieval systems operating in a multi-media environment will certainly require such capabilities.

6.3 Document Image Compression

Finally, we highlight a current research area which has significant impact on the ability to deal robustly with document images. It is clear that for any large image database, storage costs must be considered, and image compression is one way of reducing storage requirements. In the document domain, the measure of a good compression algorithm may have a number of parameters beyond the traditional space reduction.

For example, we find that digital libraries which contain document images benefit not only from compression, which reduces the amount of storage necessary, but also from the ability to process and search the underlying documents easily and efficiently.

Recent advances in document image compression have capitalized on symbol-level redundancies to achieve high levels of compression and provide basic access to compressed representation to facilitate, for example, image retrieval. The general technique, first proposed by Ascher and Nagy [3], exploits redundancies by representing each pattern class by a single prototype and representing the image as a set of pointers to these prototypes.

The concept of a dynamic library was formalized by Witten et al., [69], who take small regions that are believed to correspond to textual symbols, cluster them and a create a prototype for each cluster. The image regions are then represented uniquely by a combination of this prototype, the prototype's location in the image, and an encoding of the error image.

Kia and Doermann extended this work to provide a representation which facilitates the efficient representation of both the page and the residual. The representation supports basic component-level document image processing tasks such as skew detection and correction, keyword search, sub-image retrieval, and language identification as well as lossy representation and progressive transmission. Howard [29] demonstrates the advantages of efficient organization. Enhancements in clustering, such as the work done by Bloomberg and Vincent [6] and Zhang and Danskin [70], can greatly improve performance both in lossless and lossy representations.

The implications of being able to treat document images more like the manipulatable, searchable entities we have come to expect in the age of electronic documents are encouraging.

7 Discussion and Conclusions

Although the concept of a raw document image database is attractive, comprehensive solutions which do not require complete and accurate conversion to a machine-readable form continue to be elusive for practical systems. In fact, the continuum which exists between conversion- (or recognition-) based approaches and what can be considered image-based approaches makes defining each class somewhat difficult. In general, the extent to which queries appeal directly to image properties, rather than symbolic representations, is a reasonable way to classify techniques. A further distinction can be made between indexing based on text content and indexing based on document structure. In this paper we have attempted to provide some background about past research on both. We have summarized the primary retrieval and abstracting efforts in Table 7.

Ultimately both indexing and retrieval will make use of the powerful features offered both in the visual representation of the page, and in the underlying content of the text, graphics and images. Such systems will need to address complex tradeoffs between algorithm speed, image quality and retrieval recall and precision.

	Task		Medium			Feature					
Reference	Ret	Abs	OCR	Doc Img	Gr Img	C/W Shp	Str	Texture	Exmp	Other	Comments
Lopresti[40]	x		x								Enhanced IR
Ohta[45]	x		x								
Fujisawa[25]	x		x								
Takasu[63]	x		x								
Cornell[39]	x		x	x							
Chen[10, 11]	x		x				x				
DeCurtins[18]	x			x			x				
Spitz[57, 56, 55]	x	x		x		x					
Chen[9, 8]		x		x		x	x			x	Image Abstracting
Doermann[21]		x		x			x	x	x		Functional Desc
Herrmann[28]	x			x			x				
Jasimha[35]	x			x			x			x	Logos and Signatures
Doermann[22, 52]	x			x			x	x	x	x	Zone attribute
Lorenz[41]	x				x		x				
Syeda-Mahmood[61]	x				x		x			x	
Samet[49, 50, 51]	x				x		x			x	Legend-driven

Table 1: Summary of primary image retrieval and abstracting efforts showing initial datatype, and the features used (character/word shape, page, line or text structure, texture, document example, or other).

References

[1] *Symposiums on Document Analysis and Information Retrieval, University of Nevada, Las Vegas*, 1992 and 1993.

[2] M. Amlani and R. Kasturi. A query processor for information extraction from images of paper-based maps. In *RIAO 88*, pages 991–1000, 1988.

[3] R.N. Ascher and G. Nagy. A means for achieving a high degree of compaction on scan-digitized printed text. *IEEE Transactions on Computers*, 23(11):1174–1179, 1974.

[4] S.C. Bagley and G.E. Kopec. Applications of text image editing. In *Proceedings of the SPIE - Image Handling and Reproduction Systems Integration*, 1991.

[5] S.C. Bagley and G.E. Kopec. Image based text processing. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 551–558, 1991.

[6] D.S. Bloomberg and L.M. Vincent. Blur hit-miss transform and its use in document image pattern detection. In *Proceedings of the SPIE - Document Recognition II*, pages 278–292, 1995.

[7] F. Cesarini, E. Francesconi, M. Gori, S. Marinai, J. Sheng, and G. Soda. A neural based architecture for spot-noisy logo recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 175–179, 1997.

[8] F. Chen and D. Bloomberg. Extraction of indicative summary sentences from imaged documents. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 227–232, 1997.

[9] F. R. Chen and D. S. Bloomberg. Extraction of thematically relevant text from images. In *Symposium on Document Analysis and Information Retrieval*, pages 163–178, 1996.

[10] F. R. Chen, L. D. Wilcox, and D. S. Bloomberg. Detecting and locating partially specified keywords in scanned images using hidden Markov models. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 133–138, 1993.

[11] F.R. Chen, D.S. Bloomberg, and L.D. Wilcox. Spotting phrases in lines of imaged text. In *Proceedings of the SPIE - Document Recognition II*, pages 256–269, 1995.

[12] F.R. Chen, L.D Wilcox, and D.S. Bloomberg. Word spotting in scanned images using hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–4, 1993.

[13] F.R. Chen, L.D. Wilcox, and D.S. Bloomberg. A comparison of discrete and continuous hidden Markov models for phrase spotting in text images. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 398–402, 1995.

[14] F.R. Chen and M.M. Withgott. The use of emphasis to automatically summarize a spoken discourse. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 229–232, 1992.

[15] D. Chetverikov, J. Liang, J. Komuves, and R.M. Haralick. Zone classification using texture features. In *Proceedings of the International Conference on Pattern Recognition*, pages 676–680, 1996.

[16] W.B. Croft, S.M. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. In *Symposium on Document Analysis and Information Retrieval*, pages 115–126, 1994.

[17] J. Cullen, J. Hull, and P. Hart. Document image database retrieval and browsing using texture analysis. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 718–721, 1997.

[18] J.L. DeCurtins and E.C. Chen. Keyword spotting via word shape recognition. In *Proceedings of the SPIE - Document Recognition II*, pages 270–277, 1995.

[19] D. Doermann, H. Li, and O. Kia. The detection of duplicates in document image databases. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 314–318, 1997.

[20] D. Doermann, E. Rivlin, and A. Rosenfeld. The function of documents. *IJCV*, 1998. To appear.

[21] D. Doermann, E. Rivlin, and I. Weiss. Applying algebraic and differential invariants for logo recognition. *Machine Vision and Applications*, 9(2):73–86, 1996.

[22] D. Doermann, J. Sauvola, H. Kauniskangas, C. Shin, M. Pietikainen, and A.Rosenfeld. The development of a general framework for intelligent document image retrieval. In *Document Analysis Systems*, pages 605–632, 1996.

[23] D. Doermann and S. Yao. Generating synthetic data for text analysis systems. In *Symposium on Document Analysis and Information Retrieval*, pages 449–467, 1995.

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

[25] H. Fujisawa and K. Marukawa. Full text search and document recognition of Japanese text. In *Symposium on Document Analysis and Information Retrieval*, pages 55–80, 1995.

[26] V.N. Gudivada and V.V. Raghavan. Spatial similarity based retrieval in image databases. In *Symposium on Document Analysis and Information Retrieval*, pages 255–270, 1993.

[27] D. K. Harman, editor. *The Fourth Text Retrieval Conference (TREC-4)*, Gaithersburg, MD, 1996. NIST.

[28] P. Herrmann and G. Schlagetar. Retrieval of document images using layout knowledge. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 537–540, 1993.

[29] P. Howard. Lossless and lossy compression of text images by soft pattern matching. In *Proceedings of the IEEE Data Compression Conference*, pages 210–219, 1996.

[30] J. Hull and J. Cullen. Document image similarity and equivalence detection. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 308–312, 1997.

[31] J. Hull, J. Cullen, and M. Peairs. Document image matching and retrieval techniques. In *Proceedings of the 1997 Symposium on Document Image Understanding Technology*, pages 31–38, 1997.

[32] J.J. Hull. Document image matching and retrieval with multiple distortion-invariant descriptors. In *International Workshop on Document Analysis Systems*, pages 383–400, 1994.

[33] A. K. Jain and Y. Zhong. Page segmentation using texture analysis. *Pattern Recognition*, 29(5):743–770, 1996.

[34] A.K. Jain, S.K. Bhattacharjee, and Y. Chen. On texture in document images. In *Proceedings of Computer Vision and Pattern Recognition*, pages 677–680, 1992.

[35] M.Y. Jaisimha, A. Bruce, and T. Nguyen. Docbrowse: A system for textual and graphical querying on degraded document image data. In *Document Analysis Systems*, pages 581–604, 1996.

[36] M.Y. Jaisimha, E. A. Riskin, and R. Ladner. Model-based restoration of document images for OCR. In *Proceedings of the SPIE - Document Recognition III*, pages 297–308, 1996.

[37] M. Koga, T. Murakami, Y. Shima, and H. Fujisawa. Structure analysis method of graph image for document image retrieval. In *Proceedings of the SPIE - Character Recognition Technologies*, pages 291–295, 1993.

[38] J. Kupiec, J. Pedersen, and F.R. Chen. A trainable document summarizer. In *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, 1995.

[39] C. Lagoze, E. Shaw, J. R. Davis, and D. B. Krafft. Dienst: Implementation reference manual. Technical report, Cornell University, 1995.

[40] D. Lopresti and J. Zhou. Retrieval strategies for noisy text. In *Symposium on Document Analysis and Information Retrieval*, pages 255–270, 1996.

[41] O. Lorenz and G. Monagan. A retrieval system for graphical documents. In *Symposium on Document Analysis and Information Retrieval*, pages 291–300, 1995.

[42] H.P. Luhn. The automatic creation of literature abstracts. *I.B.M. Journal of Research and Development*, pages 159–165, 1956.

[43] R. Manmatha. Multimedia indexing and retrieval research at the Center for Intelligent Information Retrieval. In *Proceedings of the 1997 Symposium on Document Image Understanding Technology*, pages 16–30, 1997.

[44] A. Myka and U. Guntzer. Measuring the effectsx of OCR errors on similarity linking. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 968–973, 1997.

[45] M. Ohta, A. Takasu, and J. Adachi. Retrieval methods for English text with misrecognized OCR characters. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 950–956, 1997.

[46] N.C. Rowe. Retrieving captioned pictures using statistical correlations and a theory of caption-picture co-reference. In *Symposium on Document Analysis and Information Retrieval*, pages 525–534, 1995.

[47] G. Salton. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

[48] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[49] H. Samet and A. Soffer. A legend-driven geographic symbol recognition system. In *Proceedings of the International Conference on Pattern Recognition*, pages 350–355, 1994.

[50] H. Samet and A. Soffer. A map acquisition, storage, indexing, and retrieval system. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 992–996, 1995.

[51] H. Samet and A. Soffer. MARCO: MAp Retrieval by COntent. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):783–798, 1996.

[52] J. Sauvola, D. Doermann, H. Kauniskangas, C. Shin, M. Koivusaari, and M. Pietikainen. Graphical tools and techniques for querying document databases. In *Proceedings of the First Brazilian Symposium on Document Image Analysis*, pages 213–224, 1997.

[53] G. L. De Silva and J.J. Hull. Proper noun detection in document images. *Pattern Recognition*, 27(2):311–320, 1994.

[54] A. Soffer. Image categorization using texture features. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 233–237, 1997.

[55] A. L. Spitz. Logotype detection in compressed images using alignment signatures. In *Symposium on Document Analysis and Information Retrieval*, pages 303–310, 1996.

[56] A.L. Spitz. An OCR based on character shape codes and lexical information. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 723–728, 1995.

[57] A.L. Spitz. Using character shape codes for word spotting in document images. In *Shape, Structure and Pattern Recognition*, pages 382–389. World Scientific, Singapore, 1995.

[58] R.K. Srihari. Automatic indexing and content-based retrieval of captioned photographs. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1165–1168, 1995.

[59] R.K. Srihari. Automatic indexing and content-based retrieval of captioned photographs. *IEEE Computer*, 28(9):49–56, 1995.

[60] P. Suda, C. Bridoux, B. Kammerer, and G. Maderlechner. Logo and word matching using a general approach to signal registration. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 61–65, 1997.

[61] T.F. Syeda-Mahmood. Indexing of technical manual docuement databases. In *SPIE - Storage and Retrieval for Image and Video Databases III*, pages 430–441, 1995.

[62] K. Taghva, J. Borsack, and A. Condit. Expert system for automatically correcting OCR output. In *Proceedings of the SPIE - Document Recognition*, pages 270–278, 1994.

[63] A. Takasu. An approximate string match for garbled text with various accuracy. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 957–961, 1997.

[64] A. Takasu, S. Satoh, and E. Katsura. A document understanding method for database construction of an electronic library. In *Proceedings of the International Conference on Pattern Recognition*, pages 463–466, 1994.

[65] H. Tanaka and A. Kogawara. High speed string edit methods using hierarchical files and hashing technique. In *Proceedings of the International Conference on Pattern Recognition*, pages 334–336, 1988.

[66] J.M. Trenkle and R.C. Vogt. Word recognition for information retrieval in the image domain. In *Symposium on Document Analysis and Information Retrieval*, pages 105–122, 1993.

[67] K. Tsuda, S. Senda, M. Minoh, and K. Ikeda. Clustering OCR-ed texts for browsing document image database. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 171–174, 1995.

[68] D. Wang and S.N. Srihari. Classification of newspaper image blocks using texture analysis. *Computer Vision, Graphics and Image Processing*, 47:327–352, 1989.

[69] I. Witten, T. Bell, H. Emberson, S. Inglis, and A. Moffat. Textual image compression: Two-stage lossy/lossless encoding of textual images. *Proceedings of the IEEE*, 82:878–888, 1994.

[70] Q. Zhang and J. Danskin. Entropy-based pattern matching for document image compression. In *Proceedings of the International Conference on Image Processing*, pages 221–224, 1996.

List of Tables

1 Summary of primary image retrieval and abstracting efforts showing initial datatype, and the features used (character/word shape, page, line or text structure, texture, document example, or other). 27