



Intelligent Indexing and Semantic Retrieval of Multimodal Documents

ROHINI K. SRIHARI
ZHONGFEI ZHANG
AIBING RAO

rohini@cedar.buffalo.edu
zhongfei@cedar.buffalo.edu
arao@cedar.buffalo.edu

*Center for Document Analysis and Recognition (CEDAR), UB Commons, 520 Lee Entrance-Suite 202,
State University of New York at Buffalo, Buffalo, NY 14228-2583, USA*

Received April 1, 1999; Revised December 21, 1999; Accepted January 11, 2000

Abstract. Finding useful information from large multimodal document collections such as the WWW without encountering numerous false positives poses a challenge to multimedia information retrieval systems (MMIR). This research addresses the problem of finding pictures. The fact that images do not appear in isolation, but rather with accompanying, *collateral* text is exploited. Taken independently, existing techniques for picture retrieval using (i) text-based and (ii) image-based methods have several limitations. This research presents a general model for multimodal information retrieval that addresses the following issues: (i) users' information need, (ii) expressing information need through composite, multimodal queries, and (iii) determining the most appropriate weighted combination of indexing techniques in order to best satisfy information need. A machine learning approach is proposed for the latter. The focus is on improving *precision* and *recall* in a MMIR system by optimally combining text and image similarity. Experiments are presented which demonstrate the utility of individual indexing systems in improving overall average precision.

Keywords: multimedia information retrieval, content-based retrieval, image indexing, text indexing, multimodal query processing

1. Introduction

With the advent of digital libraries, it is becoming increasingly important to develop capabilities for intelligent indexing and retrieval of *multimodal* documents (Maybury 1997). Multimodal refers to the use of two or more distinct modalities in the information, such as language and pictures. The fact that *multimedia* sources such as speech, ASCII can be used to convey a single modality should be noted. Currently, this is achieved by independently applying text indexing and image indexing techniques to the text and image components respectively. If the goal is indexing and retrieval of specialized, homogeneous, document collections, such an approach may be feasible. However, if the objective is to index multimodal documents without having *a priori* knowledge of their content, then more sophisticated techniques are required. This paper deals with the latter situation.

It will be shown that in order to retrieve images based on their similarity in content, a wide variety of similarity measures need to be employed. This in turn, necessitates several types of indexing techniques, each suited for a particular type of document. Since it is computationally not feasible to deploy all indexing techniques on each multimodal document,

techniques are required to select the most appropriate techniques for a given document. This research exploits the fact that images are typically accompanied by descriptive text, such as photograph captions. It discusses techniques for (i) joint indexing of text and images, (ii) specialized image similarity techniques, and (iii) combining information obtained from text and image indexing in satisfying multimodal queries.

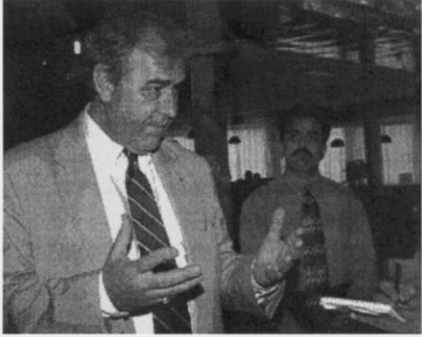
Taken independently, existing techniques for text and image retrieval have several limitations. Text-based methods, while very powerful in matching context (Salton 1989), do not have access to image content. There has been a flurry of interest in using textual captions to retrieve images (Rowe and Guglielmo 1993). Searching captions for keywords and names will not necessarily yield the correct information, as objects mentioned in the caption are not always in the picture. This results in a large number of false positives which need to be eliminated or reduced. In a recent test, a query was posed to a search engine to find pictures of Clinton and Gore resulting in 941 images. After applying filters to eliminate graphics and spurious images (e.g., white space), 547 potential pictures which satisfied the above query remained. A manual inspection revealed that only 76 of the 547 pictures contained pictures of Clinton or Gore! This illustrates the tremendous need to (i) employ image-level verification, and (ii) the need to use text more intelligently.

Typical image-based methods compute general similarity between images based on statistical image properties (Niblack et al. 1993). Examples of such properties are texture and color (Swain and Ballard 1991, Smith 1997). While these methods are robust and efficient, they provide very limited semantic indexing capabilities. There are some techniques which perform object identification; however these techniques are computationally expensive and not sufficiently robust for use in a content-based retrieval system. This is due to a need to balance processing efficiency with indexing capabilities. If object recognition is performed in isolation, this is probably true. More recently, other attempts to extract semantic properties of images based on spatial distribution of color and texture properties have also been attempted (Smith and Chang 1996). Such techniques have drawbacks, primarily due to their weak disambiguation; these are discussed later. Webseer (1998) describes an attempt to utilize both image and text content in a picture search engine. However, text understanding is limited to processing of HTML tags; no attempt to extract descriptions of the picture is made. More important, it does not address the interaction of text and image processing in deriving semantic descriptions of a picture.

This research focuses on improving precision and recall in a multimodal information retrieval system by interactively combining text processing with image processing. The fact that images do not appear in isolation, but rather with accompanying text, which is referred to as *collateral text* is exploited. Figure 1 illustrates such a case. The interaction of text and image content takes place in *both* the indexing and retrieval phases. An application of this research, namely a picture search engine which permits a user to retrieve pictures of people in various contexts is presented. A sample query would be *find pictures of victims of natural disasters*. This query could also be accompanied by an exemplar image. Text indexing is accomplished through standard statistical text indexing techniques and is used to satisfy the general context that the user specifies. Preliminary work in the intelligent use of collateral text in determining *pictorial attributes* is also presented. This information is used to dynamically select the most appropriate image indexing techniques. If the

Caption:

MANAMA, BHN, 4-AUG-1998: Richard Butler, head of the U.N. Special Commission weapons inspection team, talks to a reporter on arrival in Bahrain Tuesday Aug. 4 1998, after talks with Iraqi officials on dismantling Iraq's weapons of mass destruction collapsed. Butler, who will report to the Security Council on Thursday on the impasse, blamed Iraqis for the failure of talks.



Collateral Text:

UNITED NATIONS, Aug.4 (UPI) -- U.N. Secretary-General Kofi Annan says the breakdown in Iraqi weapons talks "may be a major hiccup, but a hiccup that we can overcome, I hope."

.....

when Butler comes back and reports, we will be able to continue our work."

Figure 1. Sample data from the UPI collection.

image is composed primarily of human faces, for example, simple color histogram indexing techniques alone will not suffice. Besides the typical image indexing techniques based on color, texture, shape, etc., specialized techniques such as face detection and recognition are also employed. This information is useful in sorting the resulting set of pictures based on various visual criteria (e.g., the prominence of faces). Experiments have been conducted to effectively combine text content with image content in the *retrieval* stage; the goal is to satisfy the semantics of the query.

The next section discusses the relevance of this work to document image retrieval. Section 3 describes the application in more detail. Section 4 presents a model for multimodal information retrieval. Sections 5 and 6 discuss text and image indexing techniques developed for this application, including background similarity matching of images. Finally, Section 7 describes retrieval experiments using the model described.

1.1. Multimodal document image indexing

Document image understanding (DIU) (Baird et al. 1992) is concerned with the automatic conversion of a scanned document into a set of structured, symbolic entities. Deriving such a semantic representation requires determining the *layout structure* as well as the *logical structure* of a document. At the layout structure, entities represent "blocks" of information; these are classified as text, graphics, or photograph blocks. The logical structure represents

higher-level semantics; for example, a text block and a photograph block could be related through a photograph-caption relationship. Since the ultimate goal is *retrieval* of specified information, it is necessary to employ appropriate *indexing* techniques for the entities. Text blocks may be indexed by first applying optical character recognition (OCR) algorithms, and subsequently using text indexing techniques. While much progress has been made in efficient indexing of certain types of multimodal documents, such as those consisting of text, tables, charts and graphics, indexing of photographic blocks remains an open problem. The word “indexing” is used here to denote the extraction and representation of semantic content.

This research explores the interaction of textual and photographic information in multimodal documents. The focus is on the type of indexing that would be necessary for representing and retrieving photographic information. In particular, techniques are presented for intelligent indexing of photographic blocks based on accompanying text blocks that serve as captions. It will be shown that various image indexing techniques are necessary depending on the type of photograph. For example, photographs of natural scenery may be indexed using techniques such as color histograms, whereas images consisting primarily of people, require advanced indexing techniques such as face detection in conjunction with global similarity techniques. It will be shown that accompanying text plays a key role in determining the appropriate indexing technique. This paper also addresses the subsequent stage of retrieval of images. The need for combining various types of indexing is illustrated, particularly the need for dynamic combination based on the query.

Traditionally, document imaging has referred to the imaging of paper documents. This work also has applications in document image understanding where the document is already in electronic form. An example of this is HTML documents found on the World Wide Web (WWW). The WWW may be viewed as the ultimate, large-scale, dynamically changing, multimedia database. Finding useful information from large-scale multimedia document databases poses a challenge in the area of multimodal information indexing and retrieval. It is a non-trivial task to determine the scope of the text that serves as a “caption” to a photograph in an HTML document. The results of an effort in this area are presented later in the paper. A particularly interesting case is one where the caption text is embedded in the photograph itself. This is the technique used by several major web-based news services such as CNN and MSNBC since it guarantees correct placement of the caption underneath the photograph regardless of the browser or HTML version being used. Techniques for extracting and recognizing embedded text have also been developed as part of the overall effort being described here.

While it is recognized that the problems of determining logical structure, including identifying the scope of text relevant to a photograph are important, they are not the focus of this paper. It is assumed that the text and photographic blocks have been identified, correlated, and are available electronically.

2. Important attributes for picture searches

Before techniques for extracting picture properties from text and images are described, it is useful to examine typical queries used in retrieving pictures. Jorgensen (1996) describes

experimental work in the relative importance of picture attributes to users. Twelve high-level attributes including *literal object, people, human attributes, art historical information, visual elements, color, location, description, abstract, content/story, viewer response and external relationship* were measured. It is interesting to note that *literal object* accounted for up to 31% of the responses. Human form and other human characteristics accounted for approximately 15%. Color, texture, etc. ranked much lower compared to the first two categories. The role of content/story varied widely, from insignificant to highly important. In other words, users dynamically combine image content and context in their queries.

Romer in Romer (1998) describes a wish list for image archive managers, specifically, the types of data descriptions necessary for practical retrieval. The heavy reliance on text-based descriptions is questioned; furthermore, the adaptation of such techniques to multimodal content is required. The need for visual thesauri (Srihari and Burhans 1994, Chang and Lee 1991) is also stressed, since these provide a natural way of cataloging pictures, an important task. An ontology of picture types would be desirable. Finally, Romer describes the need for “a precise definition of image elements and their proximal relationship to one another”. This would permit queries such as *find a man sitting in a carriage in front of Niagara Falls*.

Based on the above analysis, it is clear that object recognition is a highly desirable component of picture description. Although object recognition in general is not possible, for specific classes of objects, and with feedback from text processing, object recognition may be attempted. It is also necessary to extract further semantic attributes of a picture by mapping low-level image features such as color, texture into semantic primitives. Efforts in this area (Smith and Chang 1996) are a start, but suffer from weak disambiguation and hence can be applied in select databases; our work aims to improve this. Improved text-based techniques for predicting image elements and their structural relationships are presented.

3. MIR: A multimodal picture retrieval system

To demonstrate the effectiveness of combining text and image content, a robust, efficient and sophisticated picture search engine has been developed; specifically, this system will selectively retrieve pictures of people and/or similar scenery in various contexts. A sample query could be *find outdoor pictures of Bill Clinton with Hillary talking to reporters on Martha's Vineyard*. This should generate pictures where (i) Bill and Hillary Clinton actually appear in the picture (verified by face detection/recognition), (ii) the picture depicts an outdoor setting, and (iii) the collateral text supports the additional contextual requirements. The word robust means the ability to perform under various data conditions; potential problems could be lack of or limited accompanying text/HTML, complex document layout etc. The system should degrade gracefully under such conditions. Efficiency refers primarily to the time required for retrievals which are performed online. Since image indexing operations are time-consuming, they are performed off-line. Finally, sophistication refers to the specificity of the query/response.

There are three datasets that are being used in this research. The first is a dataset consisting of approximately 5000 images with accompanying text provided to us by United Press International (UPI). These images representing topical news and human interest stories,

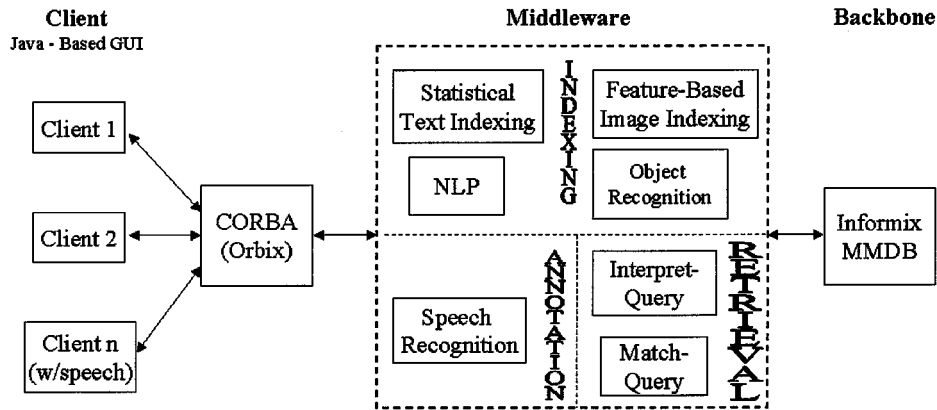


Figure 2. Architecture of multimedia information retrieval system.

were collected over a period of two months. Thus, there is some (but alas, not enough) overlap in the context of pictures, a desirable feature for this research. The majority of these pictures consist of people; there are also a considerable number of pictures depicting various man-made objects (e.g., cars). Thus, this database is quite different from the traditional natural scenery databases used by researchers working in image similarity (Picard et al. 1994, Smith 1997). Figure 1 illustrates a sample image that is accompanied by a caption, as well as more extensive text referred to as *collateral* text. A second dataset consisting of consumer photos (provided by Kodak) is also being used. These photos are accompanied by speech annotations recorded on the built-in microphone of the Kodak DC260 digital camera. The third dataset consists of a set of multimodal documents downloaded from the WWW. These are proving to be more challenging since the issue of identifying text relevant to an image must first be addressed. Furthermore, the images are of poorer quality.

Figure 2 depicts the overall architecture of the MIR system. It consists of three components: (i) the *client* modules used for annotation and retrieval, (ii) the server module (or *middleware*, supporting *annotation*, *indexing* and *retrieval*), and (iii) the *backbone* of the system, in this case, the multimedia database itself. A key feature of MIR is its ability to support *multimodal queries*. A user may input both context words as well as an image exemplar as a query. This example is discussed later on in the experimental results section.

The focus of this paper is on the indexing and especially, the retrieval capabilities of the system. The ability to annotate images using a speech interface is also supported by MIR, but not discussed here further. Such a feature is required in a system whereby consumers can annotate their personal photograph collection, enabling easy retrieval later on. As the architecture illustrates, various types of multimedia indexing are supported; these are discussed in later sections. The retrieval module addresses the task of understanding the semantics of the query, i.e., decomposing the query into its constituent parts. It also determines the best method of presenting the results. Finally, the user interface supports both query refinement and relevance feedback.

4. A model for multimodal information retrieval

Even though there has been much success recently in text-based IR systems, there is still a feeling that the needs of users are not being adequately met. Multimodal IR presents an even greater challenge since it adds more data types/modalities, each having its own retrieval models. The body of literature in multimodal IR is vast, ranging from logic formalisms for expressing the syntax and semantics of multimodal queries (Meghini 1995) to MPEG-4 (<http://www.crs4.it/~luigi/MPEG/MPEG4.html>, 1998) standards for video coding which calls for explicit encoding of semantic scene contents. A popular approach has been to add a layer representing *meta querying* on top of the individual retrieval models. An agent-based architecture for decomposing and processing multimodal queries is discussed in Merialdo and Dubois (1997). In focusing so much on formalisms, especially in the logic-based approaches, researchers sometimes make unrealistic assumptions about the quality of information that can be automatically extracted (e.g., the detection of complex temporal events in video).

Subrahmanian (1998) presents a query language, HM-SQL, for retrieving hybrid, multimedia data. This new language, which is an extension of SQL, includes enhancements that support querying of multimedia entities. For example, a query to *find all image/video objects containing Jane Shady wearing a purple suit and Denis Dopeman* is translated into the HM-SQL query depicted in figure 3.

While this model may suffice for cases where meta-data is manually added to multimedia data, thereby creating a structured multimedia database, it is not applicable to situations where automatic indexing of multimedia takes place. The problems with this model of multimedia retrieval include:

```

SELECT M
FROM smds source1 M
WHERE
  (FindType(M)=Video OR FindType(M)=Image
   AND
   M in FindObjWithFeature(Denis Dopeman)
   AND
   M in FindObjWithFeatureandAttr(Jane
                                   Shady, suit, purple))
UNION
(SELECT M.file
 FROM imagedb idb M
 WHERE
  M IN imagedb:getpic(Dennis Dopeman))

```

Figure 3. HM-SQL query example.

- Assumes that information about objects and relationships is available in a high-level, semantic format. In the above query, it is assumed that information about Jane Shady (derived from either an image or video) is available, including details of what she is wearing. This is not realistic in an automatic indexing situation.
- Assumes that each media source has its own specialized query language to access specialized indexing schemes. More important, it assumes that these indexes can directly retrieve symbolic information, such as finding a picture of Dennis Dopeyman. In automatic indexing situations, such a black box approach is not possible. There is a synergy between various indexing techniques, both for a given media source, as well as across media sources that is necessary. It may be necessary to combine several indexing techniques.
- The reliability of an indexing technique such as *getpic* is not taken into account. This is detailed more in the model to be presented next.

In this research, the focus is on utilizing *automatically* extracted information from multimodal data in improved retrieval. For example, in order to retrieve pictures of Denis Dopeyman, a judicious combination of techniques such as face detection, recognition, specialized similarity matching, as well as text indexing may be required! Furthermore, special operators for combining the various results are required.

A sound multimedia IR model must address the following issues:

- How to decompose the composite multimedia query into the individual query components, that is, determining the semantics of the query;
- How to estimate the reliability of an indexing technique for a given query component;
- How to incorporate multiple combination schemes for low-level indexing techniques thus producing a single ranking;
- How to maximize the expected precision and recall over all possible combination schemes in order to best satisfy the user's information need;
- How to incorporate relevance feedback in the model;
- User interfaces, not addressed in this paper.

If a certain indexing function has a high expected value, it may contribute to a modification of a query component to be used in another indexing function. For example, if image similarity has a high expected utility, the text query string can be modified to be similar to the text components of top-ranked documents based on image similarity. This technique is referred to as *blind relevance feedback*.

4.1. *Semantics of multimodal queries: Information need*

When a user presents a query, either as a text string, or as a composite query involving several media, he is attempting to express his *information need*. Information need is an abstract concept: it represents everything he is looking for, and no more. In the case of multimedia especially, a user typically recognizes when a document matches his information need, but may not be able to describe it in words. Information need is approximated through *concrete* queries. These queries can consist of a single modality, as with text search, or consist of

several modalities, referred to as a *composite query*. Queries can either represent a subset of information needs, or serve as an example of a document matching the information need. In many cases, a composite query is required in order to express an information need. An example of this is *find pictures like this that were taken in the summer of 1998*. Without some external source of context, this query cannot be satisfied.

Obviously, certain modalities are more suited for expressing an information need. For example, an image query may be better suited for finding pictures than a text description, if the artistic qualities of the picture are the central focus. An interesting case is where a composite query is used when one is not required. That is, *redundancy* is present in the composite query. Although a picture query matches the information need best, a redundant text query may also be included. Such queries represent a realization of the weakness of indexing methods associated with a certain modality. If image matching were perfect, then a redundant text query would not be necessary.

Figure 4 illustrates the three levels of multimedia retrieval. At the highest level, there is the abstract information need. The user approximates his information need through queries, single or composite; these are referred to as *media query objects*. Each of these are in turn processed through the *query decomposition* phase into *query components*. Examples of query components that are to be matched include **text strings**, **image features** such as color histograms, **object categories** (e.g., person, Bill Clinton, car), **spatial relationships** (e.g. person on top of car), **picture attributes** such as indoor/outdoor, as well as **meta-data**

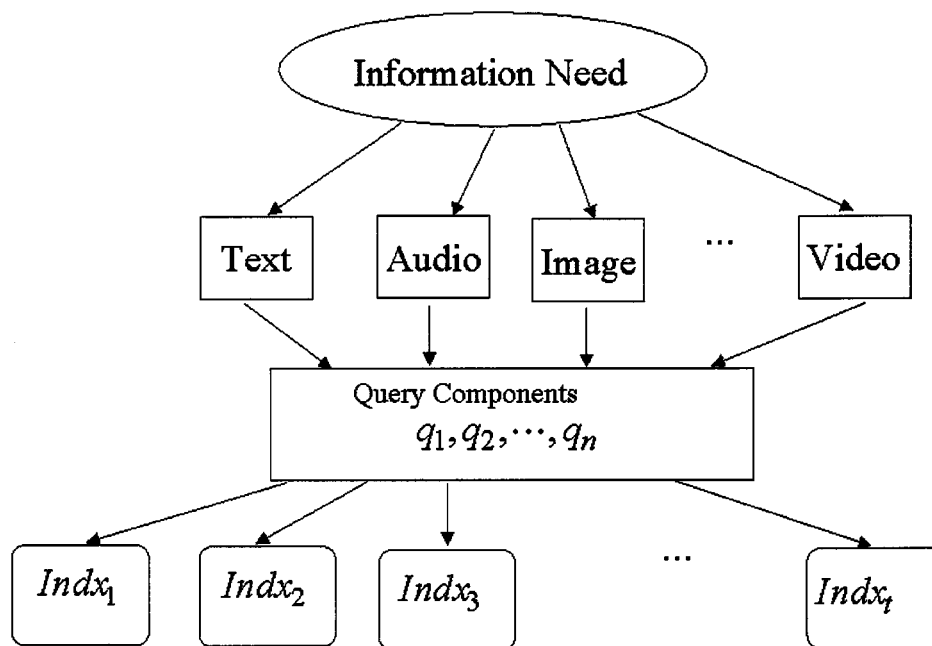


Figure 4. Translating multimedia queries into low-level indexing techniques.

such as date, time, etc. Clearly, the number and types of query components will increase as new media sources such as audio and video are introduced.

In the last stage, query components are matched using a judicious combination of the low-level indexing techniques. Let $\mathcal{D} = \{D_1, D_2, \dots, D_m\}$ be the set of all data items. Each low-level indexing function $indx_i$ ranks the document database based on certain criteria. Given an indexing function $indx_i$ and a query component q_j , let

$$sim_i^j : \mathcal{D} \rightarrow I$$

be the normalized similarity between q_j and the j th component of each document with respect to $indx_i$; I is the closed interval of real numbers from 0.0 to 1.0. Sort \mathcal{D} using $sim_i^j(D_k)$ as the key of document D_k . This results in a ranking of the database, \mathcal{D}_i^j , which is the match for the query based on the j th component under indexing scheme $indx_i$. The ranking is a permutation of \mathcal{D} . Since each of these similarity functions have inherent weaknesses, one interesting issue is how to properly combine these functions to best satisfy the user's information need.

4.2. Maximizing expected utility

This section discusses a model whereby the process of satisfying the user's information need may be viewed as maximizing the expected utility of the various similarity functions. It involves a training phase for determining the suitability of a given sim_i^j .

Let \mathcal{P} be the set of all permutations of the data set \mathcal{D} . Each element of \mathcal{P} may be regarded as a ranking of the database representing a match to some given query. Let $\mathcal{D}_{opt} \in \mathcal{P}$ be an optimal ranking in that it has the best average precision for that query. Here the average precision of the retrieval in response to a query is defined as

$$AP = \frac{1}{R} \sum_{i=1}^R \frac{i}{N_i}$$

where R is the total number of relevant documents and N_i is the number of documents retrieved so far when the i th relevant document is retrieved. Based on the ground truth provided by \mathcal{D}_{opt} , a performance utility with respect to the query can be defined as:

$$AP : \mathcal{P} \rightarrow \mathcal{R}^+$$

which maps each permutation to the average precision of the permutation, here \mathcal{R}^+ is the set of all non-negative reals. Average precision combines precision and recall in one measure; this is the quantity to be maximized in the model being described.

The goal is to approximate \mathcal{D}_{opt} , that is to maximize the average precision of the original query based on the performance of the individual component matches. The goal of the multimedia IR system is to select the weights ω_i^j that maximize the average precision of the permutation $\mathcal{D}_c \in \mathcal{P}$:

$$\mathcal{D}_c = P(\hat{\omega}_i^j)$$

which is generated by sorting \mathcal{D} with key computed by applying $\sum_{i,j} \omega_i^j sim_i^j$ to each data item. This represents the most typical case, *linear combination*. The combination can also be non-linear function such as *filter*.

A training phase is now employed in order to determine the weights ω_i^j . The training data consists of a set of queries Q_1, Q_2, \dots, Q_n as well as the corresponding relevance judgment. For each Q_k , decompose it into the corresponding set of query components q_j . For a weight ω_i^j , (i) apply $\sum_{i,j} \omega_i^j sim_i^j$ to each data item (ii) sort \mathcal{D} using the above values as a sorting key, resulting in the permutation $P_k(\omega_i^j)$, and (iii) then calculate $\mathbf{AP}(P_k(\omega_i^j))$. Finally, the average

$$\frac{1}{n} \sum_{k=1}^n \mathbf{AP}(P_k(\omega_i^j))$$

is the overall performance by employing weights ω_i^j for the set of test queries. The goal is to maximize this performance which corresponds to determining the optimal set of weights $\hat{\omega}_i^j$.

In the retrieval stage, a query is decomposed in a similar manner. The component queries q_j are used to compute sim_i^j ; then sort \mathcal{D} using $\sum_{i,j} \hat{\omega}_i^j sim_i^j$ values as sort keys. Finally, the training phase described above can be conducted for several models of combining indexing techniques, including linear combinations and filtering. As an example, consider the query *find pictures of Bill Clinton during the last election*. In this case, the context is applied first, and then filtered based on the results of face detection. The command to perform this is:

```
Filter((SIM(Text, q_{text})),
      SIM(attr=faces))
```

4.3. Query decomposition

Here, the focus is on the interpretation of the query, as handled by the procedure *Int_Query* which attempts to understand the users request and decompose it accordingly.

User input includes one or more of the following: (i) *text_query*, a text string; (ii) *image_query*, an image; (iii) *topic_query*, one or more concepts selected from a pre-defined set of topics, such as *sports, politics, entertainment, etc.*; and (iv) *user_preferences*, a set of choices made by the user indicating the relative importance of context, image content, certain image features, and preferred display choices. The specific objective of the *Int_Query* procedure is to determine the arguments to each of the sim_i^j components mentioned above.

Determining arguments to the statistical text and image similarity functions are straightforward. The text string comprising the query is processed resulting in content terms to be used in vector-space matching algorithm. In the case of a query image, the image features are available already, or are computed if necessary. Determining arguments to the object and relationship similarities are more involved. Some NLP analysis of the *Text_String* is required to determine which people, objects, events and spatial relationships are implied by the query. It is at this stage that text can help guide image indexing.

The next two sections discuss text indexing and image indexing. They describe how the *sim* function operates on these two modalities.

5. Text indexing

There are three main objectives to research in text processing for this effort: (i) statistical text indexing to capture general context, (ii) using text information to *guide* the image indexing process, i.e., to select the most appropriate indexing techniques, and (iii) advanced NLP techniques to extract picture attributes of a picture based on collateral text.

5.1. Statistical text indexing

The goal here is to capture the general context represented by collateral text. Though not useful in deriving exact picture descriptions, statistical text indexing plays a key role in a robust multimodal IR system. The SMART (Salton 1989) vector-space text indexing system is employed in this work; this has been interfaced with MIR. The problem being faced here differs from traditional document similarity matching since the text being indexed, viz, collateral and caption text is frequently very sparse. For this reason, experimentation with the use of NLP pre-processing in conjunction with statistical indexing has been conducted. NLP pre-processing refers to methods such as Named Entity (NE) tagging (Bikel et al. 1997) which classify groups of words as person name, location, etc. For example in the phrase, *Tiger Woods at the River Oaks Club, River Oaks Club* would be classified as a location. Applying NE tagging to captions and collateral text reduces errors typically associated with words having multiple uses. A query to “find pictures of woods containing oaks” should not give a high rank to the above caption. NE tagging of queries and captions leads to improved precision and recall. NE tagging is also a useful pre-processing step to parsing, described below. Finally, the output of NE tagging assists in determining appropriate image indexing techniques.

5.2. Extracting picture attributes through NLP techniques

The goal of image indexing is to automatically produce a semantic representation of the picture contents. Based on the current state of the art in image processing and computer vision techniques, only modest progress has been made towards this goal. In this work, the fact that images are accompanied by text is exploited. By applying sophisticated natural language processing (NLP) techniques to such text, it is possible to derive vital information about a picture’s content. Some organizations such as Kodak are manually annotating picture and video clip databases to permit flexible retrieval. Annotation consists of adding logical assertions regarding important entities and relationships in a picture. These are then used in an expert system for retrieval. Aslandogan et al. (1997) describes a system for image retrieval based on matching manually entered entities and attributes of pictures. The goal is to *automatically* derive the following information which photo archivists have deemed to be important in picture retrieval:

- Determining which **objects** and **people** are present in the scene; the **location** and **time** are also of importance as is the **focus** of the picture. Srihari (1995) discusses syntactic and semantic rules for this task.
- Preserving **event** (or activity) as well as **spatial** relationships which are mentioned in the text.
- Determining further **attributes** of the picture such as indoor vs. outdoor, mood, etc.

The above information also is used to guide an intelligent image indexing system. For example, if people are predicted to be in the picture, face detection will be called for. Furthermore, image similarity retrieval involving such images may specify either foreground or background matching. On the other hand, a picture determined to consist primarily of scenery will be indexed using conventional color histogram techniques as these will typically suffice.

Information extraction (IE) techniques (Sundheim 1995), particularly shallow techniques can be used effectively for this purpose. Unlike text understanding systems, IE is concerned only with extracting relevant data which has been specified a priori using fixed templates; such is the situation here.

In the semantic indexing system developed here, syntactical groups (noun groups, verb groups, etc) are first identified as *objects*; then short phrases representing semantics are grouped over these objects. The semantics are extracted based on the type codes assigned to the objects, and represented as *relation vectors* between two objects. The matching proceeds as a two-level process. In the first level match (*node-level* match) we establish a one-one mapping between objects in documents and objects in queries by comparing their similarities based on their semantic distance in WordNet. The second level match (*arc-level* match) proceeds by comparing the relation vectors between the corresponding objects. A similarity score is finally computed based on a conditional probability formula relating the two levels. The overall system diagram is given in figure 5, where SEM denotes the semantics files, OBJ denotes the object list files, and VEC represents the vector files.

The following are the key phases in semantic indexing of documents.

- *Identify Objects*: To identify objects within a document, we first used a rule-based Part of Speech Tagger (Brill 1992) to tag each term in the tokenized document. The identification is done using regular expressions involving the POS tags, and consists of 6 possible patterns ranging from “don’t care” patterns such as modal words to noun and verb group patterns. Each object consists of a head and a list of modifiers. A type code is assigned to each word in an object according to 25 unique noun beginners (Miller 1998) and 15 unique verb beginners (Fellbaum 1998), the type of an object is determined by the type code of its head.
- *Short Phrase Extraction*: Based on regular expressions involving the list of objects identified for each document, short phrases are extracted. There are 5 phrase patterns ranging from complex nominals to conjunctive groups. Each extracted phrase is represented by 5 constituents: *subject*, *action*, *object*, *subject situations* and *situations*, of which each of the first three is an object, and the last two are two lists of FWs and objects.

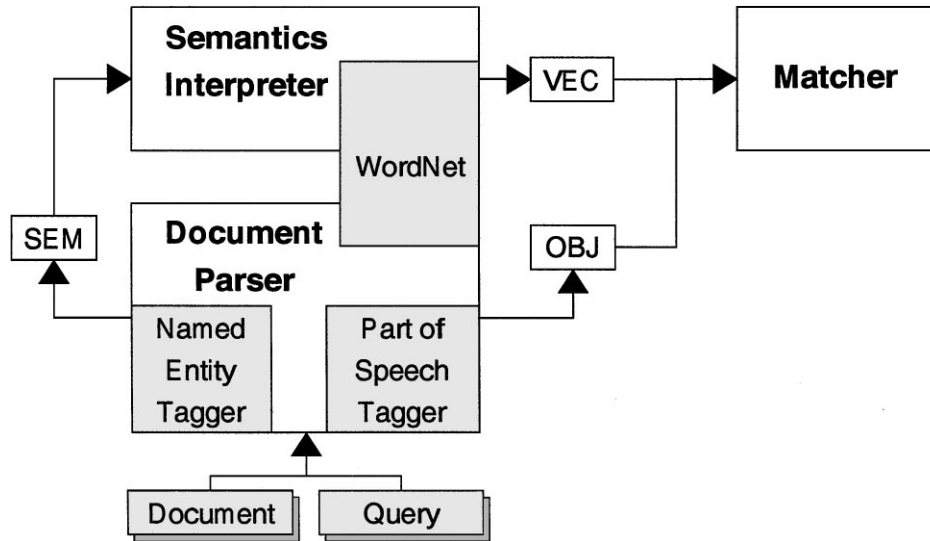


Figure 5. Architecture of the semantic text indexing system.

The similarity between a query and an indexed document is based on both the object-level (or node-level) match as well as the event match. At this point, only node-level match has been implemented. To obtain the node-level similarity score of a document, we compute the semantic distances of noun and verb objects between documents and in queries using WordNet. Since an object consists of a head and a list of modifiers, the similarity between two objects is taken as a linear combination of the similarity of the heads and that of the modifiers. Thus the problem is reduced to computing the similarity between two words.

The similarity of two words is determined by a slightly modified formula from Aslandogan et al. (1997):

$$Sim(w_1, w_2) = \frac{idf_{w_1} \times idf_{w_2}}{Dist(w_1, w_2) + 1}$$

where $Dist(w_1, w_2)$ is the distance between word w_1 and w_2 , and idf_{w_i} is the *inverse document frequency* of word w_i in the data collection. The distance between two words is determined by a weighted edge count in WordNet following hypernym/hyponym links. The weighting scheme adopted reflects the likelihood of a particular sense of a word, together with the specificity of the words along the semantic paths based on the notion of *basic-level lexicalized concepts* (Rosch et al. 1976).

For words of different syntactic categories (e.g., noun vs. verb), a conversion to noun is attempted for the verb by trying to find if it has a noun entry in WordNet. The similarity is then computed between the two nouns, but it is penalized by a predefined factor.

The final node-level similarity score for a document is thus defined as

$$Sim = Sim_R \times C_Q \times c \times C_D$$

where Sim_R is the accumulated words similarity (*raw* similarity), C_Q and C_D are the percentage of terms being matched in the query and the document (*coverage*), respectively, and c is simply a constant for weighting C_D .

6. Image indexing

Imagery is probably the most frequently encountered modality, next to text, in multimedia information retrieval. Most of the existing techniques in the literature of *content based retrieval* use low-level or intermediate-level image features such as color, texture, shape, and/or motion for indexing and retrieval. These methods have the advantage of efficient retrieval of information. On the other hand, they also suffer from the disadvantage of ineffective retrieval. This is especially true when the images in the database convey more structured semantics, such as spatial relations between objects. In such cases, the low-level and intermediate-level features may not always correlate to the semantics conveyed by the image.

In this section, various image indexing techniques are described which are used in the experiments described later on. These include: (i) face detection, (ii) color histogram matching, and (iii) background similarity matching. The latter two techniques attempt to provide more semantic matching.

6.1. Face detection

The applications of face detection and/or face recognition include (i) filtering, i.e. determining whether or not a particular image contains human being, (ii) identifying individuals i.e., handling queries for certain well-known people using face recognition, and (iii) improving the accuracy of similarity matching. Color histogram techniques do not work well for images containing faces; the experiments presented in this paper illustrate this. However, after applying face detection to the original images, the face areas are automatically “cropped” out, and the rest of the image may be used for histogram based similarity matching.

Face detection and/or recognition has received focused attention in the literature of computer vision and pattern recognition for many years. A good survey of this topic may be found in Chellappa et al. (1995). Typically, face detection and recognition are treated separately in the literature, and the solutions proposed are normally independent from one another. In this research, a *streamlined solution* to both face detection and face recognition is pursued whereby both detection and recognition are conducted in the same color feature space; the output of the detection stage is directly fed to the recognition stage. Furthermore, face recognition is a self-learning system, meaning that the face library used in face recognition is obtained through automatic labeling systems (PICTION) that employ face detection and caption understanding (Srihari 1995). This permits automatic construction of the face library, as opposed to interactive, manual data collection.

During construction of the face library, a detected face is automatically saved into the face library if an identification can be made through text clues (Srihari 1995). In query mode, the detected face needs to be searched in the library to determine the identity of this individual.

Judging similarity between two face images to decide whether they represent the same individual can be difficult since images may reflect variations in pose, orientation, size, facial expression and background. This requires face identification in the general image domain, which is the current research effort underway.

6.2. Color histogram matching

In this section, the varying performance of color histogram matching with respect to the type of images is discussed. This motivates the need for different types of color similarity matching, depending on the type of query image. Since it may not be possible to select the most appropriate color matching technique based on the query image alone, accompanying text queries can provide clues.

An experiment using images from the Eastman Kodak database was performed. Two classes of images are selected: *Purely Scenery Images* and *Images with Human Faces*. In the following, *PSI* and *HFI* are used to denote these two classes of images, respectively. There are 58 images for *PSI* and 66 images for *HFI*. In both classes, images include the following four categories: (i) *Water Scene*: Images about river, boats, swimming, etc.; (ii) *Street Scene*: Images about street, buildings, houses, etc.; (iii) *Indoor Scene*: Images about furniture, babies, etc.; (iv) *Mountain and Sky*: Images of mountains under the sky, with or without humans. With the above images, three databases are formed: *PSI*(58 images), *HFI*(66 images) and *TOTAL*(Combining above two classes, 124 images). The images in these different data sets are of varying degrees of complexity. The goal of this experiment is to demonstrate the varying effectiveness of color similarity matching.

The histogram-based retrieval system first generates the color histogram vectors for each image. For each database, with each image as a query image, an index is generated. Finally, for each image, the *precision* and *recall* based on the above similarity groups are calculated for the top 10 matches as well as the *average precision and recall*. The average *precision-recall* graphs are shown in figure 6.

Figure 6 shows that image retrieval performs better when the database consists of purely scenery as opposed to pictures with objects in the foreground. This is the motivation to propose a theory of image similarity retrieval based on foreground and background partition; this is discussed in the next section.

6.3. Combining face detection with conventional similarity-matching techniques

In this work, an attempt is made to combine *object detection*, but not necessarily *object recognition* with conventional image similarity techniques for semantic indexing. The current focus is on retrieving images consisting of people or scenery. This requires the capabilities of face detection and/or recognition in the general image domain.

In the following text, a representation of image features is proposed by combining semantic features (i.e. objects) with primitive statistical features (i.e. objects) with primitive statistical features (e.g. histogram vectors).

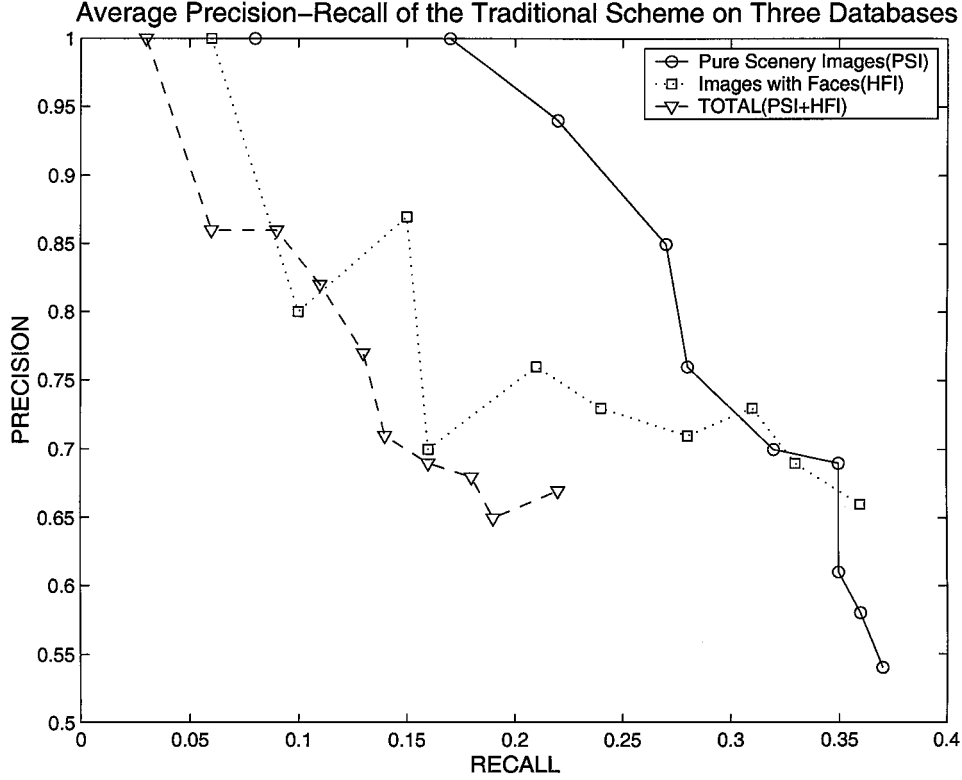


Figure 6. Average precision-recall for different databases. It shows that the algorithm performs best when the test data set is purely scenery. Number of retrieved images for each query is 10.

6.3.1. Background image computations. First several definitions are required. Let

$$FS = \{(F_i, S_i)\}_{i=1}^N \quad (1)$$

be a set of pairs consisting of statistical feature functions such as color histogram, texture features, shape information, etc., and the corresponding similarity measures of images. Let $OD = \{od_j\}_{j=1}^M$ be a set of object detectors such as face detector, building detector, etc.

For a given image I , run object detectors on I ; this results in a sequence of objects represented by their positions and dimensions. The image I is then partitioned into a vector of subimage records with the form

$$(R_1, R_2, \dots, R_K, B) \quad (2)$$

where B is the *background* subimage which is the remaining portion of I after all the object subimages have been cropped out and

$$R_j = (O_j^1, O_j^2, \dots, O_j^{n_j}) \quad (3)$$

is a set of records of the objects detected by the j th detector od_j and n_j is the number of objects. Note that some of them may be *null* which means n_j is zero. The background subimage B is then passed as input to the set of feature functions F_i ($i = 1, \dots, N$) to obtain a statistical feature vector

$$B = (f_1, f_2, \dots, f_N) \quad (4)$$

which consists of histogram vectors, texture vectors, etc.

6.3.2. Background similarity. Given two images p, q and their feature vectors

$$(R_1^p, R_2^p, \dots, R_K^p, B^p) \quad (5)$$

$$(R_1^q, R_2^q, \dots, R_K^q, B^q) \quad (6)$$

together with the object lists

$$R_j^p = (P_j^1, P_j^2, \dots, P_j^{n_j^p}) \quad (7)$$

$$R_j^q = (Q_j^1, Q_j^2, \dots, Q_j^{n_j^q}) \quad (8)$$

the simplest measure is to use the traditional measure on the feature vectors of the background subimage by ignoring the object subimage components, this is called *background similarity* and is defined as

$$S_B(p, q) = \mathcal{S}(S_B^1, S_B^2, \dots, S_B^N) \quad (9)$$

where S_B^i is the similarity measure with respect to the i th feature F_i of the background subimage, i.e.

$$S_B^i = S_i(F_i(B^p), F_i(B^q)) \quad (10)$$

with $\langle F_i, S_i \rangle$ as in Eq. (1), and \mathcal{S} is a cumulative function which combines similarity measure on each feature into a final measure. Conventional similarity matching techniques for image retrieval based on the background subimages may now be employed.

To explore other similarity measures with semantic objects present, it is necessary to first consider the measure between a single object component of P and Q , i.e., the measure

between R_j^p and R_j^q for $j = 1, 2, \dots, K$ respectively. Possible choices are: (i) *Object Count Difference*: difference between the numbers of objects of same model, and (ii) *Object Difference*: difference between the spatial relationships of the objects.

6.3.3. Results of background similarity matching. Assume the image databases and the notations used in the previous section. For each image in the database *HFI*, the face detection system (Zhang 1998, Srihari and Zhang 1998) previously described is used to detect faces. For those images where automatic face detection failed, manual detection was performed. In this work, the focus is on background similarity, rather than on face detection alone. Errors of the face detector are typically false negatives (misses). It is recognized that such errors will lead to decreased recall.

Figure 7 illustrates the difference in performance in pure image similarity versus background similarity on the database *HFI*.

Figure 8 illustrates the results of the same experiment on the database *TOTAL*. This database consists of both scenery and people pictures.

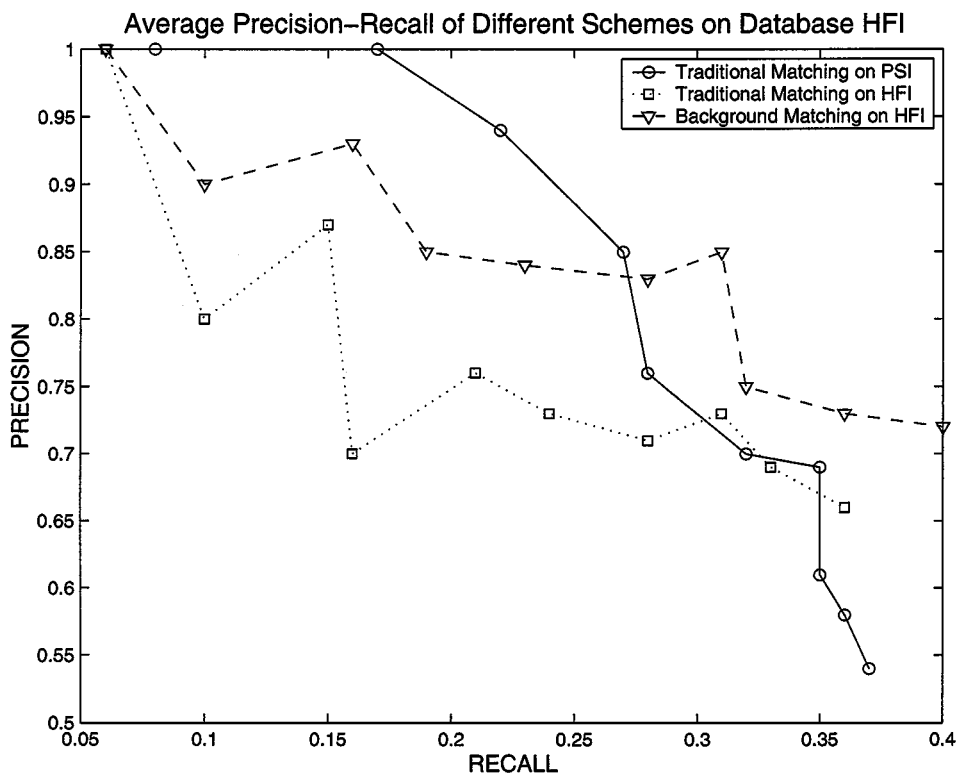


Figure 7. Average precision-recall for different matching schemes on database *HFI*. The background matching greatly improves the performance on database *HFI*, it approximates and sometimes outperforms the performance on purely scenery case. Note: (1). Number of retrieved images: 10; (2). The original system on scenery data set PSI are also shown for reference.

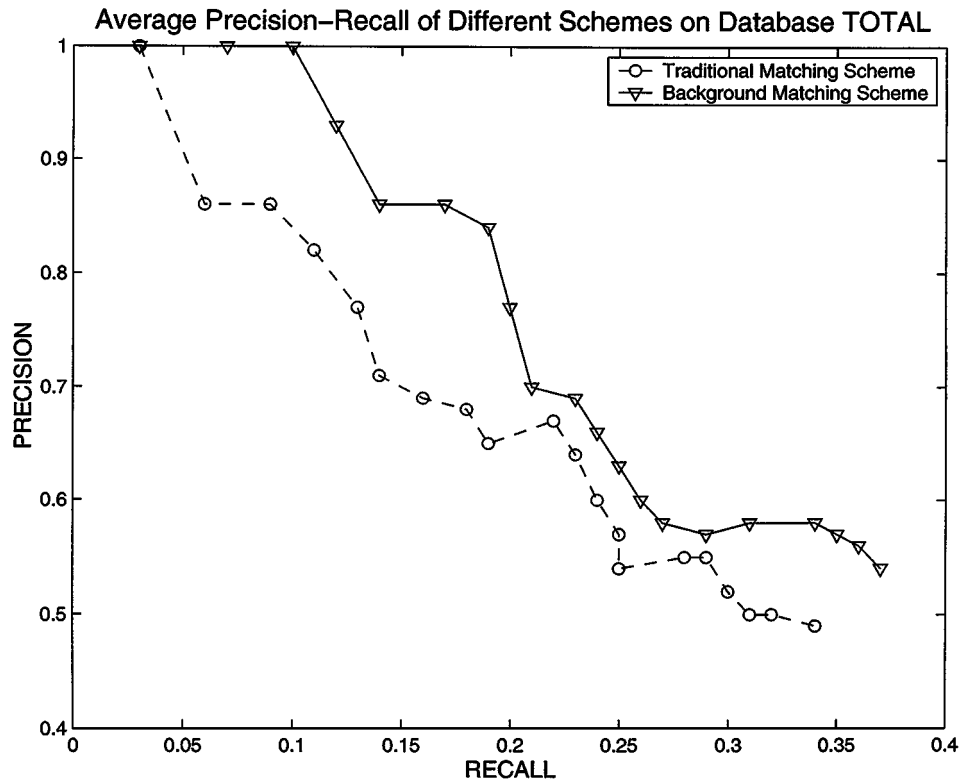


Figure 8. Average precision-recall for different matching schemes on database *TOTAL*. The background matching scheme also greatly improves the performance on database *TOTAL*. Number of retrieved images: 20.

From figures 7 and 8, it can be seen that the new background matching technique significantly improves the performance on both databases.

In summary, image processing capability currently consists of (i) a face detection module based on color feature classification to determine whether or not an image contains human faces, and (ii) a histogram based similarity matching module to determine whether or not two images “look” similar. By combining the two techniques, the system is able to retrieve images with similar backgrounds and with human beings as “foregrounds”. Extension to the capability of comparing the “foregrounds” as well is being conducted, which requires face identification in general image domain. Also Extension of this technique to other types of object is underway

7. Retrieval experiments

The central hypothesis that was being tested in these experiments was as follows: *expressing a query in terms of several, possibly redundant, multimedia query components results in better performance than expressing the query using a single, most expressive media type.*

For example, consider the goal of finding pictures which are visually similar to a given query picture. Due to the less than perfect performance of image similarity algorithms, it proved to be better to combine the original picture query with a redundant, less expressive text query. Similarly, when looking for pictures of a certain individual, say Bill Clinton, using multiple queries, including text, image similarity matching, and object category (face), proved to be better than using text matching alone.

The experiments also revealed (i) the relative effectiveness of individual media query components for certain types of queries, and (ii) the effectiveness of various combination strategies (i.e., linear combination versus filtering). For these experiments, the automatic training algorithm was not used; the weights were arrived at experimentally.

7.1. Dataset, queries

Compiling a set of multimedia queries along with a dataset on which to run them proved to be a challenging task. It was observed that given a moderate size database consisting of a few hundred arbitrarily chosen photographs, it was not possible to conduct effective tests using multimedia queries. There were very few pictures that could be classified as being in the same category based on visual similarity. Furthermore, there was very little overlap in the topic areas represented by the photographs and accompanying captions. Thus, it was not possible to test the effectiveness of multimedia queries.

To overcome these difficulties, a dataset was constructed that would support a chosen query set. In particular, 46 topic areas were first identified, ranging from names of people, to general events such as golf. Several Internet sources were identified that contained rich repositories of information, both pictures and text pertaining to these topics. A minimum of 50 documents per topic were downloaded from these sources. For each topic, one of two types of queries was formulated. These included:

- *image as primary query*: in these queries, the information need was best expressed through a query image. A redundant, less expressive text description was written for this image. The dataset for this topic was manually examined, and relevant images were marked.
- *text description as primary query*: in these queries, the information need was best expressed through a detailed text description, that is, a caption. Several images that were judged to be relevant to the query were marked. A secondary query, in terms of an image which contained some properties of the caption, but not all, was determined.

A total of 46 queries were used in this experiment, 15 where the image was the primary query mode, 15 where the text was the primary query mode, 10 involving simple searches for people, and finally, 6 queries involving background similarity.

It is important to note that relevance ranking was judged with respect to the query modality used. Within each of the categories above, various types of queries were included that called for finding objects, people, relationships between objects, as well as general picture qualities. The last category was reflected in the relevance judgments assigned to the image-centered queries. A special type of query called for finding pictures of people against specific background, such as Niagara Falls. Figure 9 illustrates some sample multimodal queries used in the experiments.

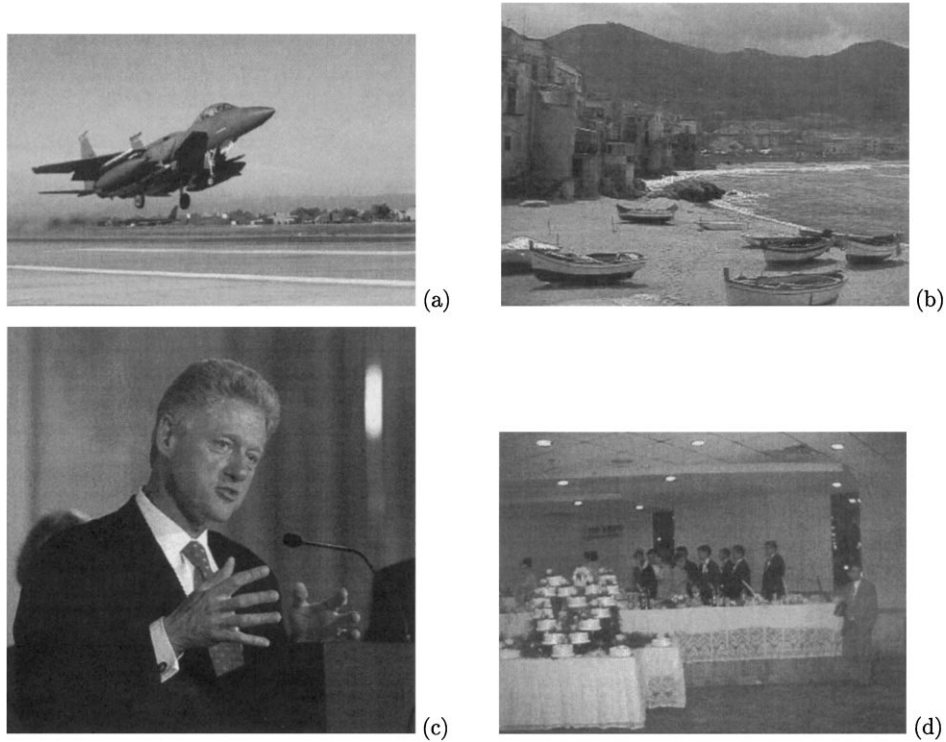


Figure 9. Sample queries used in experiments: (a) Image as primary query and secondary text string query *A fighter jet taking off from an air base*; (b) text string *beach and sea scene* as primary query with image as secondary query; (c) primary query is text string *President Clinton* with image as secondary query; (d) primary query is text string *Wedding reception scene, with banquet table and cake in background*.

7.2. Results

Table 1 presents the results of combining text and image similarity computation on both image-centered queries as well as queries where text was the primary mode. The results were calculated using the same set of queries on the topic databases (TDB) and on the combined databases (CDB). In all the tables in the following, the values shown are average values of average precisions obtained over multiple queries.

Figure 10 shows the average precision-recall of the text similarity, the image similarity, the optimal linear combination (LC) and the unbiased LC (Image Weight = 0.5) of these two similarity matching techniques of the image-centered queries on the topic database (TDB) and the combined database (CDB). Figure 11 shows the average precision-recall of the text-centered queries. It can be seen that the optimal combination, namely, the one that yields the best average precision corresponds to using image weights of approximately .3 for the individual databases, and significantly higher for the combined databases. As would be expected, the overall performance degrades on the combined database.

Table 1. Results of combining text and image similarities.

	Image weight										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TDB											
Image-based	0.631	0.712	0.713	0.722	0.725	0.737	0.725	0.711	0.671	0.590	0.516
Text-based	0.758	0.810	0.807	0.814	0.814	0.799	0.776	0.751	0.681	0.588	0.473
CDB											
Image-based	0.195	0.228	0.234	0.241	0.245	0.254	0.289	0.307	0.267	0.189	0.160
Text-based	0.199	0.224	0.224	0.228	0.235	0.237	0.240	0.233	0.240	0.190	0.146

In this table, image weight of 0 corresponds to using text similarity only; image weight of 1.0 corresponds to using image similarity only. The highlighted optimal combination is somewhere in between. The first two rows represent results using topic databases (TDB) for each query. The last two rows show results on the combined database (CDB), i.e., combining all 46 topic databases.

Average Precision-Recall of Image-Based Queries

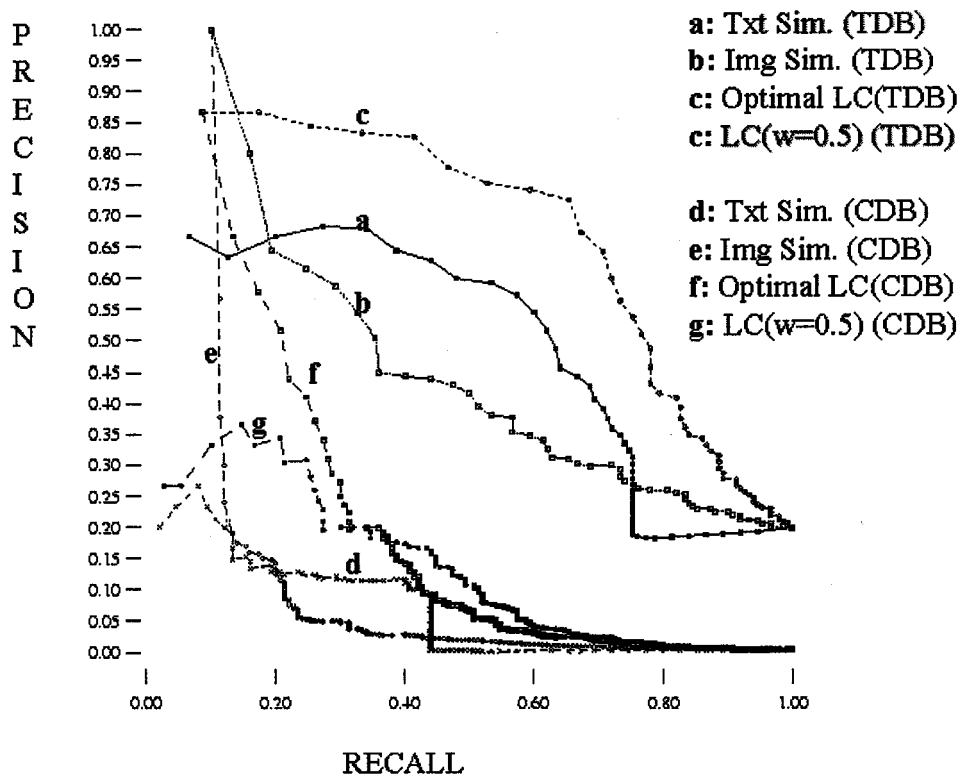


Figure 10. The average is taken over 15 image-centered queries.

Average Precision-Recall of Text-Based Queries

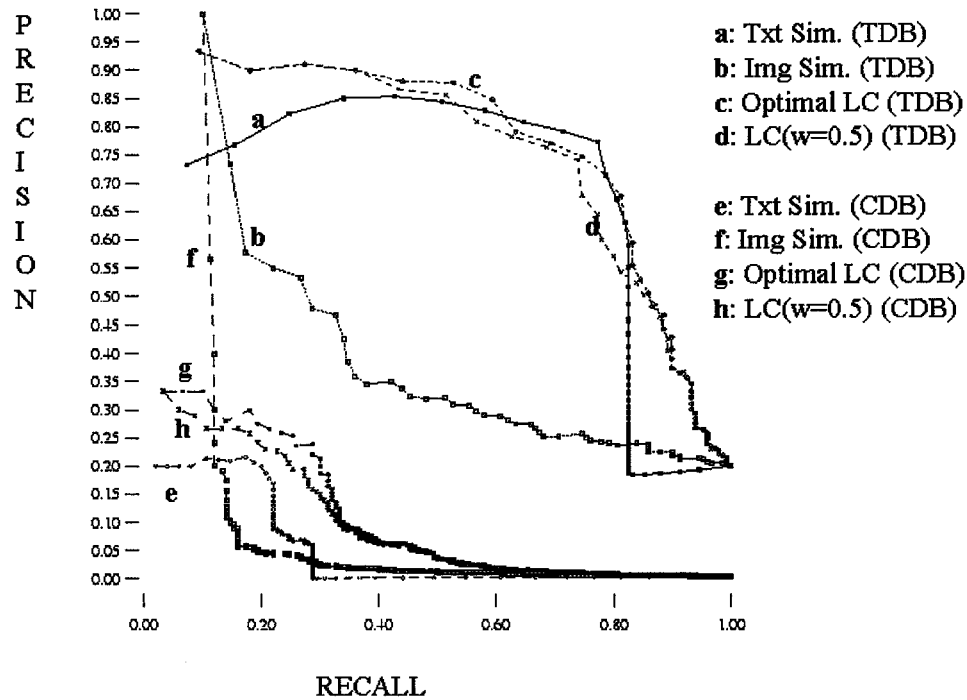


Figure 11. The average is taken over 15 text-centered queries.

Table 2 presents the results for a specific type of query, namely queries calling for pictures of people. A total of 10 queries were used in this experiment, 5 where the image was the primary query, and 5 where the text was the primary query. Figure 12 presents the average precision-recall of the queries in this case. As the data illustrates, text similarity dominated the overall performance, especially in the specialized databases. Thus, the presence of the name in the text was a very strong indicator that the person actually appeared in the picture. This was due to the bias in the way the datasets for these queries was constructed. In the

Table 2. Results of queries calling for pictures of named people, e.g., Bill Clinton.

	Image weight										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TDB	0.941	0.902	0.892	0.880	0.868	0.847	0.803	0.739	0.643	0.493	0.371
CDB	0.338	0.350	0.353	0.356	0.361	0.351	0.349	0.343	0.305	0.156	0.131

In this table, a total of 10 queries, half each of image as primary query mode and text as primary query mode were used. The table does not distinguish between these two.

Average Precision-Recall of Queries Calling for Named People

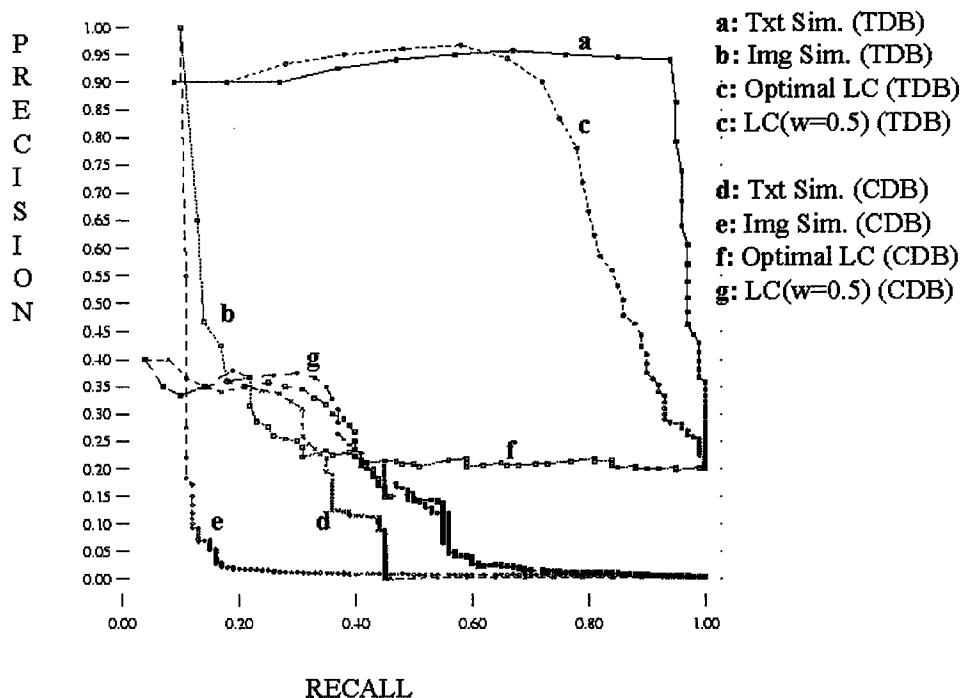


Figure 12. The average is taken over 10 queries calling for people pictures.

combined database, the text was less reliable. The mention of a name did not imply his or her presence in an accompanying picture.

Table 3 presents the average precision for queries involving background similarity matches. Figure 13 presents the average precision-recall of this case. A set of 6 queries, each calling for pictures of a group of people posing against various backgrounds was used. As in the second experiment, half of these used a text description as a primary query, the other half used image as primary query. Furthermore, two different types of image similarity

Table 3. Results of queries calling for matching pictures taken against the same background.

	Image weight										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
General Img Sim	0.277	0.568	0.564	0.570	0.584	0.596	0.612	0.617	0.650	0.676	0.645
Backgrd Img Sim	0.277	0.595	0.589	0.597	0.605	0.619	0.634	0.647	0.682	0.695	0.683

In this table, a total of 6 queries were tried, with half using image as primary query mode, the other half using text as primary query mode. Two types of image similarity, the first using general color similarity, and the second, specialized background matching were tested.

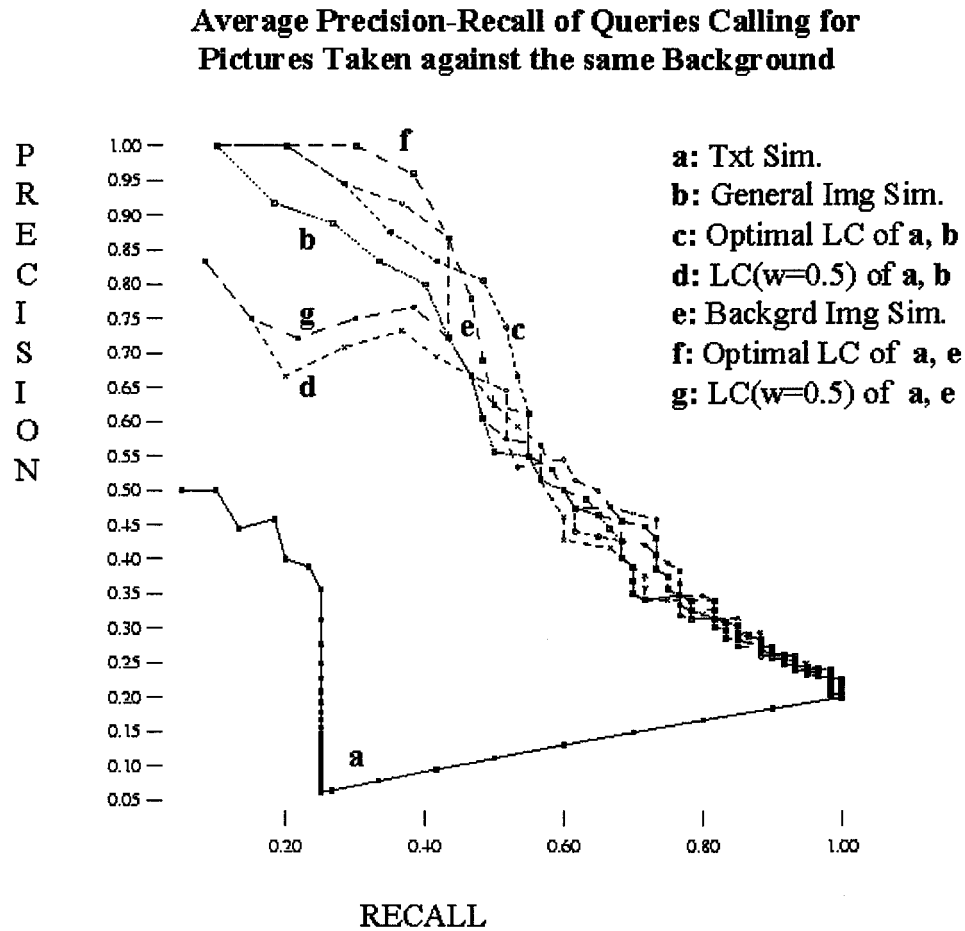
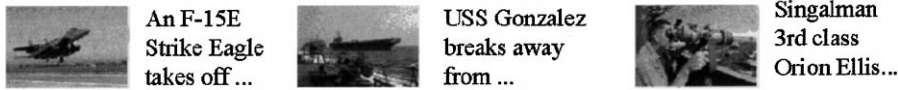


Figure 13. The average is taken over 6 queries requesting pictures with the same background.

computation were tested here. The first involved general color histogram matching; the second was a specialized background similarity matching technique introduced earlier in the paper. As the results indicate, image similarity dominated the performance in these types of queries. Furthermore, background image similarity matching proved to be a modest improvement over conventional image similarity matching.

Figure 14 illustrates results for the query shown in figure 9(a). Figure 15 illustrates results for the query shown in figure 9(b). Figure 16 illustrates results for the query shown in figure 9(c). Due to the restrictions on reproducing some of these images, it was not possible to show the most striking results for each case.

Top 3 matches by image similarity



Top 3 matches by text similarity



Top 3 matches by linear combination with image weight 0.6

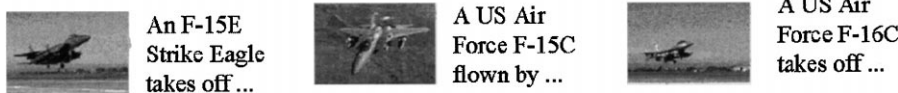
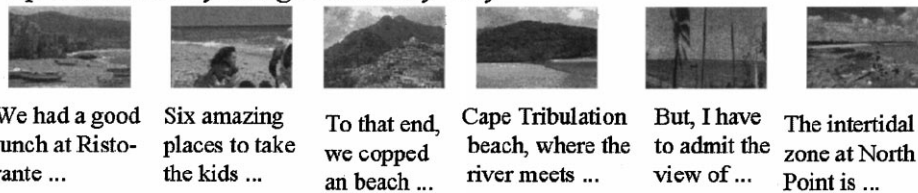
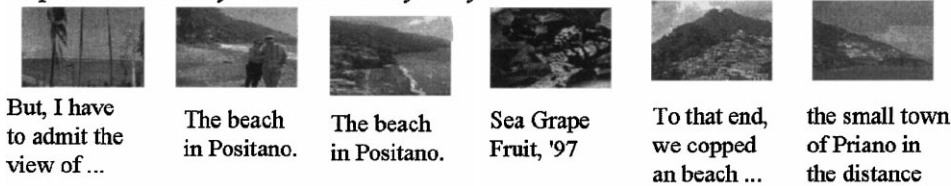


Figure 14. Results for image as primary query shown in figure 9(a).

Top 6 matches by image similarity only



Top 6 matches by text similarity only



Top 6 matches by linear combination with image weight 0.7

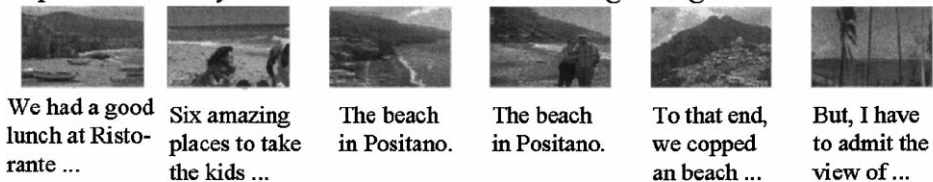


Figure 15. Results for text as primary query shown in figure 9(b).

Top 4 matches by image similarity only



President Clinton announces, January 14, that he ...



President Bill Clinton shakes hands with ...



French President Jacques Chirac escorts ...



Rosie O'Donnell joins host Alan Thicke of ...

Top 4 matches by text similarity only



French President Jacques Chirac escorts ...



French President Jacques Chirac greets ...



French President Jacques Chirac escorts ...



President Clinton and Minority Leader Richard ...

Top 4 matches by linear combination with image weight 0.7



President Clinton announces, January 14, that he ...



French President Jacques Chirac escorts ...



President Bill Clinton shakes hands with ...



President Bill Clinton delivers his State of ...

Figure 16. Results for find person query shown in figure 9(c).

7.3. Discussion

Some general observations can be made based upon the results. First, regardless of the manner in which a user prefers to express his information need (text or image), the retrieval system performs best when a combination of both is used. Thus, the user may be better off by providing a second, redundant query in a different modality.

The performance of text similarity is very good in the specialized topic databases. Adding more words to the text description will result in finer-grained matches. However, in a large database, such as the web, the presence of several words will lead to many false positives since each word could match a potentially different set of documents. There are numerous documents that contain the word Clinton; however many of the pictures present in these documents may not contain Clinton. An example of this is *protesters march before President Clinton's visit on Tuesday*. A remedy to this calls for more intelligent text indexing. Rather than treating the text query as a "set of keywords", a more semantic interpretation such as that described in an earlier section is called for. Image similarity matching, although computationally expensive, proves to be a stable technique.

Text can be used for another purpose as well. By analyzing a text query, it is sometimes possible to selectively apply different image similarity techniques. For example, a query

calling for pictures of people would benefit by employing face detection. The multimedia IR model presented earlier that includes a learning phase should account for this.

Finally, experiments such as the above would be more convincing if they were applied to a very large, heterogeneous document database such as the web. At this point, due to computational limitations, it is not possible to index such a large collection using the various suite of techniques described here. Thus, smaller, representative collections are chosen. The smaller collections immediately suffer from a bias due to the way in which they are chosen. What is necessary is a method of randomly sampling the web, such that confusion in terms of naturally occurring phenomena is present. Examples of these include: (i) multiple pictures of different people against the same background, (ii) pictures allowing us to test queries for highly structured pictures in terms of objects and relationships structure, and (iii) pictures reflecting various picture attributes such as indoor versus outdoor. It is also necessary to devise semi-automatic methods of assigning relevance judgments to this dataset for each query.

8. Summary

This paper has presented a model for multimodal information retrieval. The model addresses important issues such as satisfying users' information needs based on optimally using low-level indexing techniques. A method for training the system to learn optimal combinations has been discussed. Techniques have been presented for intelligently exploiting text in order to derive relevant attributes of the accompanying picture. The need for specialized image similarity computations based on the presence of objects and people in an image has been justified. This includes the use of object detection, namely face detection. It is felt that such techniques are required in order to perform semantic retrieval. Techniques for measuring image similarity based on background matching only have been presented. Experimental results show that an intelligent combination of such information leads to improved precision of queries, at the expense of recall. The latter may not be a serious problem when the database is extremely large, and *any* or *some* matches are required, not necessarily all matches.

Ongoing work includes more extensive testing of the system, as well as incorporating NLP results in query matching. Longer term research includes the search for an intermediate image representation that reflects the presence of objects in a picture, without having to perform full-fledged object recognition. While the work here has shown that face detection can be successfully employed, generalization to other classes of objects is necessary. It is sufficient to detect the presence of objects, without having to classify them. This will enable more sophisticated image similarity retrieval.

The lower-level query language for retrieving information about images based on their semantic content requires further refinement. Furthermore, a system for translating a high-level query such as *find outdoor pictures of John Doe* into the corresponding lower-level queries is required. This mapping should be transparent to the user.

The work presented here does not include relevance feedback techniques as of yet. Since the feedback is multimodal data, it is not clear on what basis the user has determined a particular document to be relevant. Currently, image retrieval systems with relevance feedback require the user to employ awkward graphical interfaces such as sliding bars to

determine the relative importance of a feature. Ideally, the system should be able to learn this automatically from the feedback.

It is necessary to develop more appropriate evaluation methodology. Since searching for pictures is a browsing process, rather than a static event, the need for dynamic evaluation models must be addressed. Measures such as precision and recall must be extended to reflect the utility of a given session, rather than a single query.

Finally, the design of a multi-agent system (Huhns and Shing 1998) that is capable of accessing information from various sites is being explored. The highlight is a self-learning user agent. Information which is unique to the user, and that may assist in indexing and retrieval will be housed in this agent. This includes for example, sample images of the user's face, along with faces of other people that are likely to be of interest. Such information helps in automatic face recognition at a later point. Audio files representing the user's speech samples may also be of use. Much work remains on defining a methodology whereby the information in the user agent could get automatically updated, based on past interaction with the system. Such a system would enable a higher degree of intelligence in both the indexing and retrieval stages.

References

- Aslandogan YA, Thier C, Yu CT, Zou J and Risse N (1997) Using Semantic contents and WordNet in Image Retrieval. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 286–295.
- Baird HS, Bunke H and Yamamoto K (1992) Ed. Structured Document Image Analysis. Springer-Verlag.
- Bikel DM, Miller S, Schwartz R and Weischedel R (1997) Nymble: A high-performance learning name-finder. In: Proceedings of the Fifth Conference on Applied Natural Language Processing. Morgan Kaufmann, pp. 194–201.
- Brill E (1992) A simple rule-based part of speech tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing, ACL.
- Chang CC and Lee SY (1991) Retrieval of similar pictures on pictorial databases. *Pattern Recognition*, 24(7):675–680.
- Chellappa R, Wilson CL and Sirohey S (1995) Human and machine recognition of faces: A survey. *Proc. of the IEEE*, 83(5).
- Fellbaum C (1998) A semantic network of English verbs. In: Fellbaum C, Ed., *WordNet: an Electronic Lexical Database*. MIT Press, Ch. 3.
- Huhns M and Munindar S (1998) Ed. *Readings in Agents*. Morgan Kaufmann.
- Jorgensen C (1996) An investigation of pictorial image attributes in descriptive tasks. In: Rogowitz BE and Allenbach JP, Eds., *Proceedings of SPIE Vol. 2657: Human Vision and Electronic Imaging*. SPIE Press, pp. 241–251.
- Maybury MT (1997) Ed. *Intelligent Multimedia Information Retrieval*, AAAI Press/MIT Press.
- Meghini C (1995) An image retrieval model based on classical logic. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 300–309.
- Merialdo B and Dubois F (1997) An agent-based architecture for content-based multimedia browsing. In: Maybury MT, Ed., *Intelligent Multimedia Information Retrieval*. AAAI Press/MIT Press.
- Miller GA (1998) Nouns in WordNet. In: Fellbaum C, Ed., *WordNet: an Electronic Lexical Database*. MIT Press, Ch. 1.
- Niblack W et al. (1993) The QBIC Project: Querying Images by Content Using Color, Texture, and Shape. In: *Storage and Retrieval for Image and Video Databases*. SPIE.
- Picard RW, Pentland A and Sclaroff S (1994) *Photobook: Content-based manipulation of image databases*. M.I.T Media Laboratory Perceptual Computing Technical Report, 255.

- Romer DM (1998) Research Agenda for Cultural Heritage on Information Networks: Image and Multimedia Retrieval. <http://www.ahip.getty.edu/agenda>.
- Rosch E, Mervis CB, Gray W, Johnson D and Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psychology*, 8:382–349.
- Rowe M and Guglielmo E (1993) Exploiting captions in retrieval of multimedia data. *Information Processing and Management*, 29(4):453–461.
- Salton G (1989) Automatic Text Processing. Addison-Wesley.
- Smith JR (1997) Integrated spatial and feature image systems: retrieval, analysis and compression. PhD Thesis, Columbia Univ.
- Smith JR and Chang SF (1996) Visualeek: A fully automated content-based image query System. In: Proc. of ACM Multimedia 96.
- Srihari RK (1995) Use of collateral text in understanding photos. *Artificial Intelligence Review* 8 (special issue on Integration of NLP and Vision): 409–430.
- Srihari RK and Burhans DT (1994) Visual semantics: Extracting visual information from text accompanying Pictures. In: Proceedings of AAAI-94, Seattle, WA, pp. 793–798.
- Srihari RK and Zhang Z (1998) Finding pictures in context. In: Proc. of IAPR International Workshop on Multimedia Information Analysis & Retrieval. Springer-Verlag Press, pp. 109–123.
- Subrahmanian VS (1998) Principles of Multimedia Database Systems. Morgan Kaufmann.
- Sundheim B (1995) Ed. Proceedings of the Sixth Message Understanding Conference (MUC-6). Morgan Kaufmann.
- Swain MJ and Ballard DH (1991) Color indexing. *International Journal of Computer Vision*, 7(1):11–32.
- Webseer (1998) <http://webseer.cs.uchicago.edu>.
- What is MPEG-4 (1998) <http://www.crs4.it/~luigi/MPEG/mpeg4.html>.
- Zhang Z (1998) Invited paper. Recognizing human faces in complex context. In: Proc. of the International Conference on Imaging Science, Systems, and Technology. CSREA Press, pp. 218–225.