Exploiting Multimodal Context in Image Retrieval

ROHINI K. SRIHARI AND ZHONGFEI ZHANG

ABSTRACT

THIS RESEARCH EXPLORES THE INTERACTION of textual and photographic information in multimodal documents. The World Wide Web (WWW) may be viewed as the ultimate, large-scale, dynamically changing, multimedia database. Finding useful information from the WWW without encountering numerous false positives (the current case) poses a challenge to multimedia information retrieval systems (MMIR). 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 collateral text-based methods and image-based methods have several limitations. Text-based methods, while very powerful in matching context, do not have access to image content. Image-based methods compute general similarity between images and provide limited semantics. This research focuses on improving precision and recall in an MMIR system by interactively combining text processing with image processing (IP) in both the indexing and retrieval phases. A picture search engine is demonstrated as an application.

INTRODUCTION

This research explores the interaction of textual and photographic information in multimodal documents. The World Wide Web (WWW) may be viewed as the ultimate, large-scale, dynamically changing, multi-

Rohini K. Srihari, Department of Computer Science, Center for Document Analysis and Recognition (CEDAR), UB Commons, 520 Lee Entrance—Suite 202, State University of New York, Buffalo, NY 14228-2567
Zhongfei Zhang, Computer Science Department, Watson School of Engineering and Applied Science, State University of New York at Binghamton, Vestal, NY 13902
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 496-520

media database. Finding useful information from the WWW poses a challenge in the area of multimodal information indexing and retrieval. The word "indexing" is used here to denote the extraction and representation of semantic content. 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 that permits a user to retrieve pictures of people in various contexts—is presented.

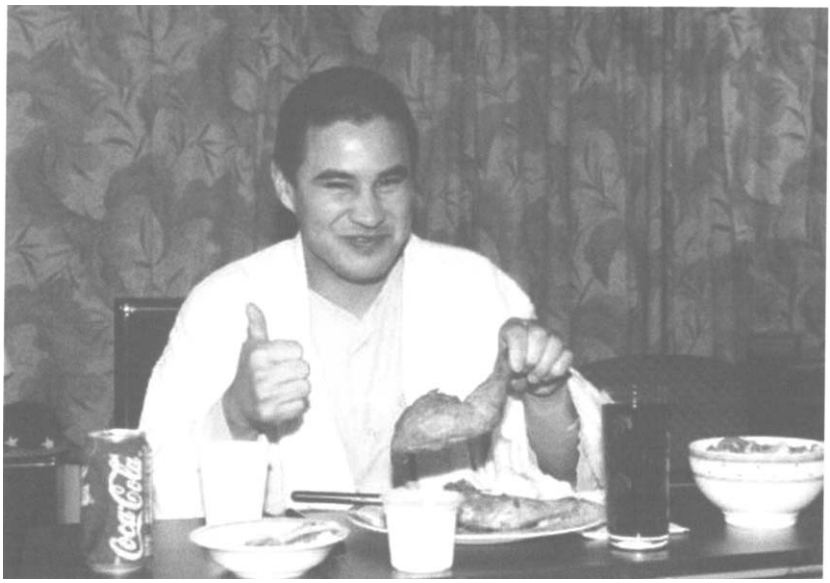

Figure 1. Staff Sgt. Andrew Ramirez digs into a plate of chicken, his first hot meal since release from captivity, at the Landstuhl Regional Medical Center, Landstuhl, Germany (U. S. Air Force photo by Senior Airman Elizabeth Weinberg). Released photo by: SRA Brian M. Boisvert, 786th Communications Squadron Record ID No. (VIRIN): 990502-F07285W-001). The picture was obtained from the U. S. Department of DefenseLink Web page located at http://defenselink.mil/multimedia/.

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 & 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 that 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 and resulted in 941 images. After applying our own filters to eliminate graphics and spurious images (e.g., white space), 547 potential pictures that satisfied the 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 employ image-level verification and to use text more intelligently.

Typical image-based methods compute general similarity between images based on statistical image properties (Flickner et al., 1995). Examples of such properties are texture and color (Swain & Ballard, 1991). While these methods are robust and efficient, they provide very limited semantic indexing capabilities. There are some techniques that 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 & Chang, 1996). Such techniques have drawbacks, primarily due to their weak disambiguation. These are discussed later. Webseer (http://webseer.cs.uchicago.edu) 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.

In this article, a system for finding pictures in context is described. A sample query would be *Find pictures of victims of natural disasters.* Specifically, experiments have been conducted to effectively combine text content with image content in the retrieval stage. Text indexing is accomplished through standard statistical text indexing techniques and is used to satisfy the general context that the user specifies. Image processing consists of face detection and recognition. This is used to present the resulting set of pictures based on various visual criteria (e.g., the prominence of faces). Experiments have been conducted on two different scenarios for this task; results from both are presented. Preliminary work in the intelligent use of collateral text in determining pictorial attributes is also presented. Such techniques can be used independently or combined with image processing techniques to provide visual verification. Thus this represents the integration of text and image processing techniques in the indexing stage.

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—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 thirty-one of the responses. Human form and other human characteristics accounted for approximately fifteen responses. Color, texture, and so on 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 (1993) 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 & Burhans, 1994; Chang & 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 (1995) 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 and texture into semantic primitives. Efforts in this area (see Smith & 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.

Webpic: 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, Webpic will selectively retrieve pictures of people 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 (1) Bill and Hillary Clinton actually appear in the picture (verified by face detection/recognition), and (2) 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, and so on. 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 offline. Finally, sophistication refers to the specificity of the query/response. In order to provide adequate responses to specific queries, it is necessary to perform more complex indexing of these data.

Figure 2 depicts the overall structure of the system. It consists of three phases. Phase 1 is the data acquisition phase—multimodal documents from WWW news sites (e.g., MSNBC, CNN, USA Today) are downloaded. In order to control the quality of data that are initially downloaded, a Web crawler in Java has been implemented to do more extensive filtering of both text and images.

The inputs to the system are a set of name keys (names of people) and an initial set of URLs to initiate the search. Some preprocessing tools are employed during this phase. One such tool is an image-based *photograph versus graphic* filter. This filter is designed and implemented based on histogram analysis. Presumably, a photograph histogram has a much wider spectrum than that of a graphic image.

A *collateral text extractor*, whose task is to determine the scope of text relevant to a given picture, is also employed. Caption text appears in a wide variety of styles. News sites such as CNN and MSNBC use *explicit* captions for pictures. These are indicated through the use of special fonts and careful placement using HTML commands as illustrated in Figure 1. In other Web pages, captions are not set off explicitly but, rather, are *implicit* by virtue of their proximity to the picture.

Explicit captions are detected based on the presence of strong HTML clues as well as the usage of key phrases such as "left, foreground, rear" and so on. These can be used to predict picture contents. General collateral text is detected based on the presence of words from the "ALT" tag, caption words, spatial proximity to picture, and so on. Such text, while not a powerful predictor of the contents of a picture, establishes the context of a picture. An image-based caption extractor that extracts ASCII text that has been embedded in images (a common practice among news oriented sites) has been developed in our laboratory and is available for use.

Phase 2 is the content analysis or indexing phase (performed offline). Phase 2 illustrates that both natural language processing (NLP) and image processing result in factual assertions to the database. This represents a more semantic analysis of the data than general text and image indexing based on statistical features. This is discussed in later sections.

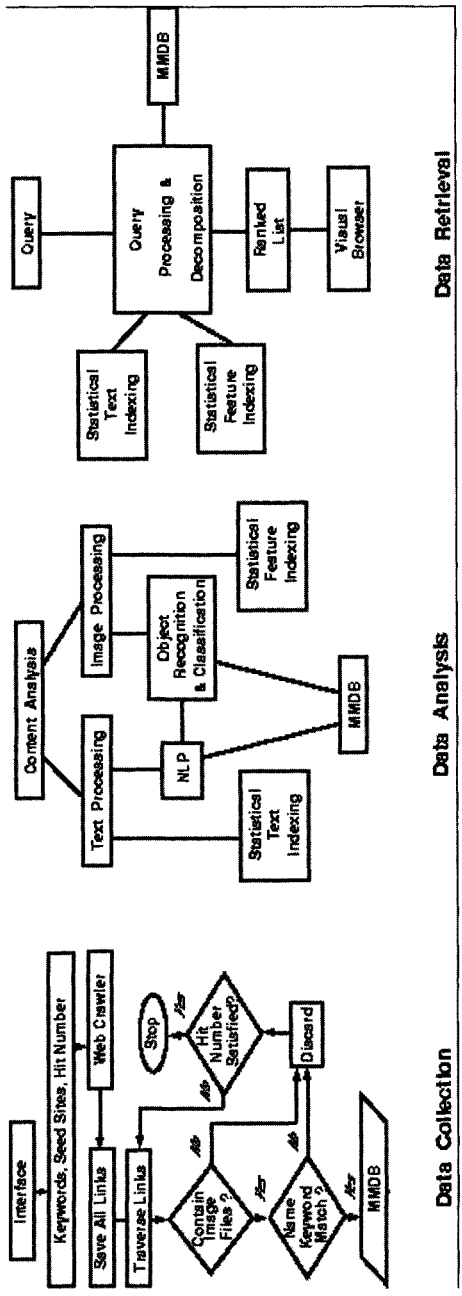

Figure 2. Overall Control Structure of the Proposed System.

Phase 3, retrieval, demonstrates the need to decompose the query into its constituent parts. A Web-based graphical user interface (GUI) has been developed for this. As Figure 3 illustrates, the system permits users to view the results of a match based on different visual criteria. This is especially useful in cases where the user knows the general context of the picture but would like to interactively browse and select pictures containing his or her desired visual attributes. The interface also illustrates that further query refinement using techniques such as image similarity are possible. Finally, although the example illustrates a primary context query, it is possible for the original query to be based on pure image matching techniques. The basic database infrastucture for a multimodal database has been built using Illustra (Illustra is a relational database management system from Informix Inc.).

This is used for data storage as well as representing factual (exact) information. Illustra's ability to define new data types and associated indexing and matching functions is useful for this project.

Metadata

For each picture and its accompanying text, the following metadata are extracted and stored. The metadata model described here is currently applicable only to text and image sources. However, it can be easily extended to accommodate audio and video sources as well:

- Text_Indx: text index, using statistical vector-space indexing techniques. This is useful in judging similarity of two contexts.
- Img_Indx1,Img_Indx2,...Img_Indxk: indexes for various image features based on statistical techniques. This includes color, texture, shape, as well as other properties useful in judging the similarity of two images.
- PDT: this is a template containing information about people, objects, events, locations, and dates mentioned in the text accompanying a picture. Such information is extracted through NLP techniques and will be discussed in the text processing section. Similarity of these templates involves a sophisticated *unification* algorithm.
- Objects: this is a template containing information about objects detected in the image (image coordinates) and their spatial relationships. It also includes information pertaining to general scene classification (e.g., indoor/outdoor, man-made/natural, and so on).

Text Indexing

Text Processing

The goal of natural language processing research in this project is to examine the use of language patterns in collateral text to indicate scene

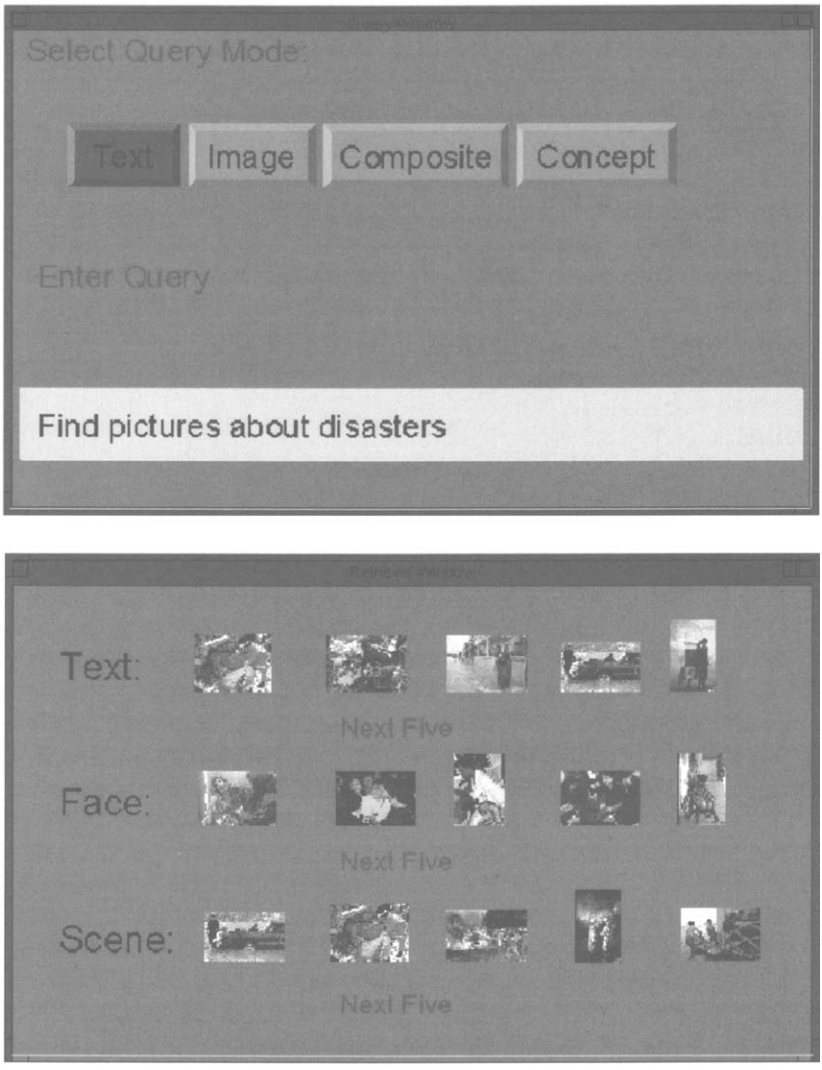

Figure 3. The Multimodal GUI Used in Retrieval.

contents in an accompanying picture. In this section, NLP techniques to achieve this goal are described. The objective is to extract properties of the accompanying picture as well as cataloging the context in which the picture appeared. Specifically, the interest is in deriving the following information that 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.
- Preserving event (or activity) as well as spatial relationships that are mentioned in the text. Spatial information, when present, can be used for automatically identifying people in pictures.

  Consider the caption *President Clinton and his family visited Niagara Falls yesterday. The First Lady and Chelsea went for a ride on the Maid of the Mist.* This should not match the query *find pictures of Clinton on the Maid of the Mist.* However, the caption *Clinton rode the Maid of the Mist Sunday* should be returned. Current IR systems that rely on statistical processing would return both captions. NLP techniques are required for correct processing in this case.
- Determining further attributes of the picture such as indoor versus outdoor, mood, and so on.
- Representing and classifying the general context indicated by the text—e.g., political, entertainment, and so on.

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) describe a system for image retrieval based on matching manually entered entities and attributes of pictures, whereas our objective is to *automatically extract* as much information as possible from natural language captions.

Specifically, the goal is to complete *picture description templates* (PDT) which represent image characteristics. Templates of this type are used by photo repository systems, such as the Kodak Picture Exchange (Romer, 1993). The templates carry information about people, objects, relationships, location, as well as other image properties. These properties include: (1) indoor versus outdoor setting, (2) active versus passive scene—i.e., an action shot versus a posed photo, (3) individual versus crowd scene, (4) daytime versus night-time, and (5) mood.

As an example, consider Figure 4 which shows the output template from processing the caption *A woman adds to the floral tribute to Princess Diana outside the gates of Kensington Palace* of Figure 1. Information extraction (IE) techniques (Sundheim, 1995), particularly shallow techniques, can be used effectively for this purpose. Unlike text understanding sys-

People: person (female, PER1)
Objects: flowers
Activity: pay_tribute (PER1, Princess Diana)
Location: Kensington Palace, "outdoor"
Event Date: Monday, Sept. 2, 1997
Focus: PER1

Figure 4. Picture Description Template (PDT).

tems, IE is concerned only with extracting relevant data that have been specified a priori using fixed templates. Such is the situation here.

Specific techniques for deriving the above information are now presented. The techniques fall into three general categories: statistical text indexing, light parsing, and extracting picture attributes.

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 information retrieval system. There has been considerable research in the area of document indexing and retrieval, particularly the vector space indexing techniques (Salton, 1989). The problem being faced here differs from traditional document matching since the text being indexed—viz, collateral text—is frequently very sparse. Minor adjustments are made to existing techniques in order to overcome the sparseness problem. This includes: (1) the use of word triggers (computed from a large corpus) to expand each content word into a set of semantically similar words, and (2) the use of natural language pre-processing in conjunction with statistical indexing. Word triggers refer to the frequent co-occurrence of certain word pairs in a given window size of text (e.g., fifty words). Natural language pre-processing refers to methods, such as Named Entity Tagging (described below), which classify groups of words as person name, location, and so on. While the use of NLP in document indexing and retrieval has met with limited success, the brevity of collateral text calls for more advanced processing.

Light Parsing: Extracting Patterns of Interest

The previous section described general content indexing; these techniques are based on statistics of word, word-pair frequencies, and so on. In this subtask, the focus is on more in-depth syntactic processing of the relevant text; this is treated as an information extraction task. Such systems consist of several hierarchical layers, each of which attempts to extract more specific information from unformatted text.

In the case of photographs, *template entities* are the objects and people appearing in the photograph, *template relationships* include spatial relationships between objects/people, as well as event/activity information. The first layer consists of *named entity tagging*; this is an extremely useful pre-processing technique and has been the subject of considerable research.

Named entity (NE) tagging refers to the process of grouping words and classifying these groups as person name, organization name, place, date, and so on. 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 collateral text reduces errors typically associated with words having multiple uses. For example, a query to "Find pictures of oaks along a river" should not retrieve the above caption since *River Oaks Club* is tagged as a location. Bikel et al. (1997) describe a statistical method for NE tagging; given a manually truthed corpus of captions and collateral text, it is straightforward to develop an NE tagger. At this point, a rule-based system for NE tagging has been implemented which is giving better than 90 percent accuracy performance.

The next layers of the hierarchical grammar are used for recognizing domain-independent syntactic structures such as noun and verb groupings (assuming that named entity tagging has already taken place); this leads to identification of template entities and basic relationships (i.e., SVO structure). The processing in these layers is confined to the bounds of single sentences. The final layer is where intersentential information is correlated, thus leading to merging of templates. It is here that the final decision on entries in the picture description template are made. For example, one sentence in a caption may refer to Princess Diana seen at her country estate, while a later sentence may refer to the fact that the estate is located outside the village of Althorp, England. In such a situation, template merging would result in the information that, in the specified picture, the location is Althorp, England. This is a form of co-reference that is being exploited. The template also includes general characteristics of the picture which may be detected from either the caption or collateral text. This is discussed in the next section.

The demands for efficient and robust natural language processing systems have caused researchers to investigate alternate formalisms for language modeling. Current information extraction requirements call for the processing of up to 80 MB of text per hour. Researchers have increasingly turned to finite-state processing techniques (Roche & Schabes, 1997). Roche (1997) says that "for the problem of parsing natural language sentences, finite-state models are both efficient and very accurate even in complex linguistic situations" (p. 241). A finite state transducer (FST) is a special case of a finite state automaton (FSA) in which each arc is labeled by a pair of symbols (input and output)

rather than a single symbol. A rule compiler (Kartutnen & Beesley, 1992) takes *regular relations* as input and constructs the corresponding FST. Operations supported by FST that are useful in grammar construction are union, intersection and, particularly, composition. Domain-specific pattern rules (to extract special attributes for a select domain) can be written as a new FST; this new FST can easily be composed with the base system. Hobbs et al. (1997) employs a cascaded set of FSTs to implement a hierarchical grammar for IE. The picture description grammar is currently being implemented as a cascaded FST.

Extracting Picture Attributes

Once the parsing process has been completed, it is possible to attach further attributes to the picture. This includes attributes such as indoor versus outdoor, mood, and so on. By employing the roles that entities take on in the picture description templates, as well as referring to ontologies and gazetteers, it is possible, in some cases, to extract further attributes. For example, if a caption refers to *Clinton on the White House lawn*, it is characterized as an outdoor picture. This is essentially a *unification* process between location types. Chakravarthy (1994) discusses the use of WordNet in performing such characterization.

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* or *image indexing and retrieval* use low-level or intermediate-level image features such as color, texture, shape, and/or motion for indexing and retrieval. Although these methods may be efficient in retrieval, the retrieval precision may not be good enough, as typically it may not be true that image features always reflect their semantic contents.

In this article, the focus is mainly on image retrieval of people or scenes in a general context. This requires capabilities of face detection and/or recognition in the general image domain. By a general image domain, it is meant that the appearances of the objects in question (e.g., faces) in different images may vary in size, pose, orientation, expression, background, as well as contrast. Since color images are very popular in use and very easy to obtain, these have been chosen for experimentation.

The potential applications of the capability of face detection and/or face recognition include: (1) filtering—i.e., determining whether or not a particular image contains a human being, (2) identifying individuals—i.e., handling queries for certain well-known people using face recognition, and (3) improving the accuracy of similarity matching. For images involving human faces, it is very difficult to check similarity based on

histograms of the entire images. Color histogram techniques do not work well for images containing faces. However, after applying face detection to the original images, the face areas may be automatically "cropped" out, and the rest of the image may still 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 years. A good survey on 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 of each other. In this task, a *streamlined solution* to both face detection and face recognition is pursued. By a streamlined solution, it is meant that both detection and recognition are conducted in the same color feature space, and the output of the detection stage is directly fed into the input of the recognition stage. Another major difference between the present research and work described earlier in the literature is that the proposed system is a self-learning system, meaning that the face library used in face recognition is obtained through face detection and text understanding using the earlier research system PICTION (Srihari, 1995b). This allows the stage of face data collection for construction of the face library as an automatic part of data mining, as opposed to interactive manual data collection usually conducted for face recognition. Note that in many situations it is impossible to do manual data collection for certain individuals, such as Bill Clinton. For those people, their face samples can only be obtained through the WWW, newspapers, and so on. Thus, automatic data collection is not only efficient but is also necessary.

Face detection is approached as pattern classification in a color feature space. The detection process is accomplished in two major steps: feature classification and candidate generation. In the feature classification stage, each pixel is classified as face or nonface based on a standard Bayesian rule (Fukunaga, 1990). The classification is conducted based on pre-tuned regions for the human face in a color feature space. The color features used in this approach are hue and chrominance. The pre-tuning of the classification region in the color feature space is conducted by sampling over 100 faces of different races from different Web sites. In the candidate generation stage, first a morphological operation is applied to remove the noise, and then a connected component search is used to collect all the "clusters" that indicate the existence of human faces. Since the pre-tuned color feature region may also classify other parts of the human body as candidates, let alone certain other objects that may happen to be within the region in the color feature space, heuristic checking is used to verify the shape of the returned bounding box to see if it conforms to the "golden ratio" law.[1] Figure 5 shows the whole process of face detection and recognition for a Web

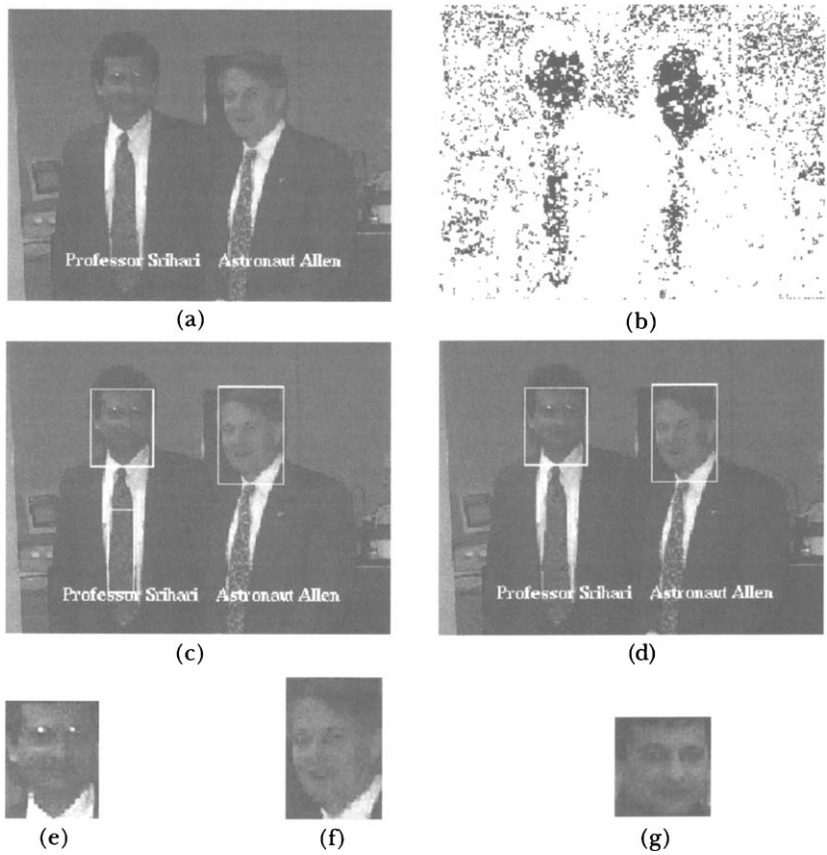

Figure 5. An Example of the Result of Automatic Face Detection and Face Recognition. (a) An original image from internet web. (b) The binary image after classification. (c) Result after morphological operations and connected component search. (d) Final detection result after applying heuristic checking to reject false positives. (e) The first face candidate returned for face recognition. (f) The second face candidate returned for face recognition. (g) Another face image of the same individual as in (e).

image. Note that each detected face is automatically saved into the face library if it has a strong textual indication of who this person is (self-learning to build up the face library), or the face image is searched in the face library to find who the person is, if the query asks to retrieve images of this individual (query stage).

In the face recognition stage, there are two modes of operation. In the mode of face library construction, it is assumed that each face image has its collateral textual information to indicate identities of the people in the image. Face detection is first applied to detect all the faces. Based on

the collateral information (Srihari, 1995b), the identities for each face may be found and thus saved into the library automatically. In the mode of query, on the other hand, the detected face needs to be searched in the library to find out the identity of this individual.

Figure 5(e) and (g) are two face images of the same individual. This is a problem of finding semantic similarity between two face images in the general image domain to identify whether or not two face images contain the same individuals. This is one of the current research directions underway. Promising experimental results based on preliminary tests show that it is possible to include the capability of querying individuals in image retrieval by conducting semantic similarity matching.

To summarize, image processing capability currently consists of: (1) a face detection module based on color feature classification to determine whether or not an image contains human faces, and (2) a histogram-based similarity matching module to determine whether or not two images "look" similar.

Multimodal Query Processing

Even though there has been much success recently in text-based information retrieval 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 standards for video coding which call 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 unreal assumptions about the quality of information that can be automatically extracted (e.g., the detection of complex temporal events in video).

The present research focuses not on the formalism used to represent the queries, rather, the focus is on the effect of utilizing *automatically* extracted information from multimodal data in improved retrieval. Processing queries requires the use of: (1) information generated from statistical text indexing, (2) information generated from natural language processing of text, and (3) information generated from image indexing—in this case, face detection and recognition—as well as color, shape, and texture indexing.

Thus, matching a query to a captioned image in the database could involve four types of similarity computation:

1. *(Text_Indx$_q$, Text_Indx$_{CapImg}$):* text-based similarity, statistical approach;
2. *SIM(Img_Indx(j)$_q$,Img_Indx$_{CapImg}$):* j=1,...,k: image similarity, for each image feature statistical approach;
3. *SIM(PDT$_q$,PDT$_{CapImg}$):* text-based concept similarity, symbolic approach; and
4. *SIM(Objects$_q$,Objects$_{CapImg}$):* image-based content similarity, symbolic approach.

Syntax and Semantics of Multimodal Queries

Similarity matching techniques for each information source are discussed in the next section. Here the discussion centers on the interpretation of the query, as handled by the procedure *Interpret_Query* which attempts to understand the user's request and decompose it accordingly.

User input includes one or more of the following: (1) *text_query*, a text string; (2) *image_query*, an image; (3) *topic_query*, one or more concepts selected from a pre-defined set of topics, such as *sports, politics, entertainment,* and so on; and (4) *user_preferences*, a set of choices made by the user indicating preferred display choices and so on. These are used by the *Interpret_Query* module in determining *ranking schemes.*

The specific objective of the *Interpret_Query* procedure is: (1) to determine the arguments to each of the *SIM(x,y)* components mentioned above, and (2) to determine the set of ranking schemes that will be used in presenting the information to the user. Determining arguments to the text and image similarity functions are straightforward. The text string comprising the query is processed, resulting in content terms to be used in a vector-space matching algorithm. In the case of a query image, the image features are available already, or are computed if necessary. Determining the arguments to the picture description template similarity and object similarity are more involved. Some natural language processing analysis of the *Text_String* is required to determine which people, objects, events, and spatial relationships are implied by the query.

Another important issue is to decide on how information should be combined. For example, for an unambiguous query such as *Find pictures of Bill Clinton,* the face detection and recognition results will be automatically applied to produce a single ranking of images satisfying the query. However, for a more subjective query, such as *Find pictures of victims of natural disasters,* the general context is first applied. The results are then sorted based on various visual criteria, thus allowing the user to browse and make a selection.

Each ranking scheme *(RSk)* defines a ranking (CapImg(k,1),CapImg (k,2),...,CapImg(k,nk)) of the images in the database. Currently, a simple technique to generate ranking schemes is employed. For each information source that is involved in a query, several sort criteria are applied in varying order. These sort criteria reflect the relative importance of each

information source. For example, for queries involving finding people in various contexts, two sorted lists will be presented to the user. The first weights the context more and the second weights the face detection results more—i.e., presence of face, relative size of face.

Matching Queries to Data

Text-based similarity is based on statistical indexing techniques; while not as precise as natural language processing techniques, it is very robust. Image-based similarity techniques using color, shape, texture, and so on have been discussed extensively in the content-based image retrieval literature. Image-based content similarity includes any visual information that has been verified by using object recognition techniques (e.g., number of faces, gender) or semantic classification (e.g., indoor versus outdoor).

When matching based on the similarity of picture description templates, it is necessary to employ unification techniques. For example, a search for *Dalmation* should match a picture whose PDT contains *dog*. That is, *Unify(Dalmation,dog)* should return a non-zero value. An approach similar to that of Aslandogan et al. (1997) to perform inexact matching is being adopted. The use of ontologies is required for several purposes in this phase. First, they are required to map entities into their *basic categories* (Rosch et al., 1976); research has shown that people most often query by basic categories (e.g., *dog* rather than *Doberman*). If the caption refers to the location as an *auditorium*, for example, it is necessary to map this into *building* for the purpose of retrieval. Similar mapping needs to take place on query terms. Srihari (1995a) and Aslandogan et al. (1997) discuss the use of WordNet in matching picture entities with queries. WordNet provides critical information in determining hierarchical relationships between entity classes and event classes.

Query Refinement and Relevance Feedback

Since users are not always sure of what they are looking for, an adaptive system is required. After specifying an initial query, the results are sorted into various classes based on the ranking schemes suggested by *Interpret_Query*. Users may choose to refine the query by either modifying the text query, concept query, or select images that best match their needs. The latter are used in a relevance feedback process, where users can interactively select pictures that satisfy their needs. Although the technique is well-understood in the text domain (Chang, 1998; Robertson, 1986; Rocchio, 1971; Ide, 1971; Croft & Harper, 1979; Fuhr & Buckley, 1991), it is still in the experimental stage in the image domain (Smith, 1997). Popular techniques include Rocchio's (1971) relevance feedback formula for the vector model and its variations (Ide, 1971), and the Croft-Harper formula (1979) for the probabilistic retrieval model and its modifications (Fuhr & Buckley, 1991; Robertson, 1986). Query refinement consists of

adjusting the weights assigned to each feature; this is the technique adopted in the text domain. Of course, the difficult aspect is determining which features are important. The multiple ranking scheme described in the previous section is of use here since each ranking corresponds to the importance of certain features (or metadata). By selecting images in certain ranking schemes, the system is able to learn which features are useful. This process can continue iteratively until the user finds the required picture. The user interface supports the visual browsing that is an integral part of image retrieval.

Retrieval Experiments

There were two experiments conducted in picture retrieval from multimodal documents. Each reflected a different strategy of combining information obtained by text indexing and image indexing. Both of these are now described.

Single Ranking Method

In this experiment, the queries are first processed using text indexing methods. This produces a ranking *Px1, ... , Pxn* as indicated in Figure 6.

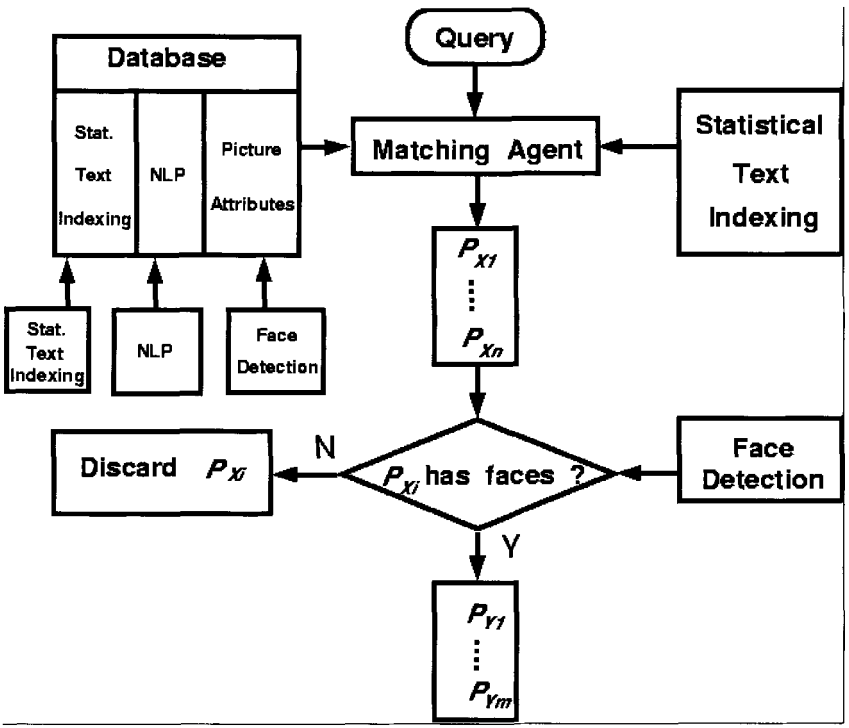

Figure 6. Single-Ranking Strategy for Combining Text and Image Information in Retrieval.

Those pictures *pxi* which do not contain faces are subsequently eliminated; this information is obtained by running the automatic face detection module.

Figure 7 presents the results of an experiment conducted on 198 images that were downloaded from various news Web sites. The original data set consisted of 941 images. Of these, 117 were empty (white space) and 277 were discarded as being graphics. From these, a subset of 198 images was chosen for this experiment.

	Text Only	Text + Manual Insp	Text + Face Det
At 5 docs	1.0	1.0	1.0(3)
At 10 docs	1.0	0.70	1.0(3)
At 15 docs	0.80	0.75	1.0(2)
At 30 docs	0.77	0.67	NA

Figure 7. Results of Single Ranking Strategy of Combining Text and Image Content. The last column indicates the result of text indexing combined with face detection. The number in parentheses indicates the number of images in the given quantile that were discarded due to failure to detect faces.

There were ten queries, each involving the search for pictures of named individual(s); some specified contexts also, such as *find pictures of Hillary Clinton at the Democratic Convention.* Due to the demands of truthing, the results for one query are reported; more comprehensive evaluation is currently underway. Figure 7 indicates precision rates using various criteria for content verification: (1) using text indexing (SMART) alone, (2) using text indexing and manual visual inspection, and (3) using text indexing and automatic face identification. As the table indicates, using text alone can be misleading—when inspected, many of the pictures do not contain the specified face. By applying face detection to the result of text indexing, photographs that do not have a high likelihood of containing faces are discarded. The last column indicates that this strategy is effective in increasing precision rates. The number in parentheses indicates the number of images in the qiven quantile that were discarded due to failure to detect faces.

Sample output is shown in Figures 8 and 9. Figure 8 illustrates the output based on text indexing alone. The last picture illustrates that text alone can be misleading. Figure 9 illustrates the re-ranked output based on results from face detection. This has the desired result that the top images are all relevant. However, a careful examination reveals that, due to the face detector's occasional failure to detect faces in images, relevant images are inadvertently being discarded. Thus this technique increases precision but lowers recall. However, if the source of images is the WWW, this may not be of concern. The face detector is continually being improved to make it more robust to varied lighting conditions.

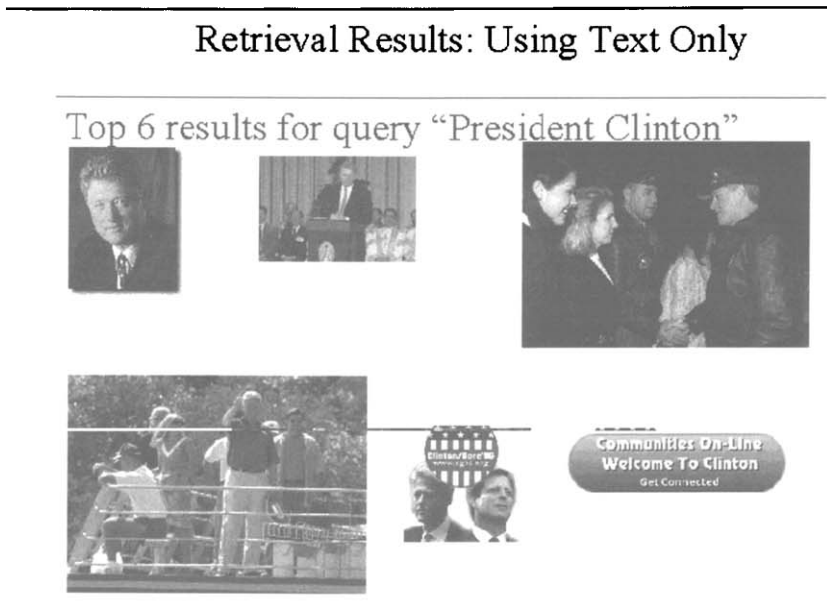

Figure 8. Top Six Images Based on Text Indexing Only.

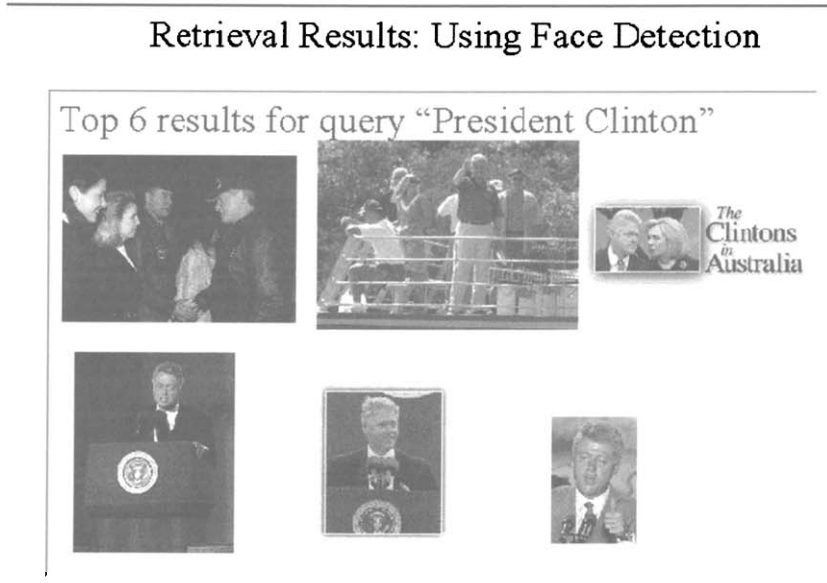

Figure 9. Top Six Images Based on Combining Text Indexing and Face Detection.

Multiple Ranking Method

In this experiment, a multiple ranking method for presenting candidate images to the user is employed. This strategy is depicted in Figure 10. The context is first verified using statistical text indexing. These candidate images are then sorted based on various visual properties. The first property is the presence of faces, the second represents the absence of faces (reflecting an emphasis on general scene context rather than individuals). This reflects the assumption that users do not know a priori exactly what kind of pictorial attributes they are looking for—i.e., that they would like to browse. Figure 11 depicts the top ranked images for the query *victims of disasters.*

Many of these refer to the recent air crash in Indonesia, partially blamed on heavy smoke from forest fires. Some images depict victims, some depict politicians discussing the situation. Based on an imposed threshold, only the top ten images returned by text retrieval were considered. As the results show, this produces a different ranking of images, where the lower row clearly emphasizes people. Had a lower threshold for text retrieval been used, the difference would have been more dramatic.

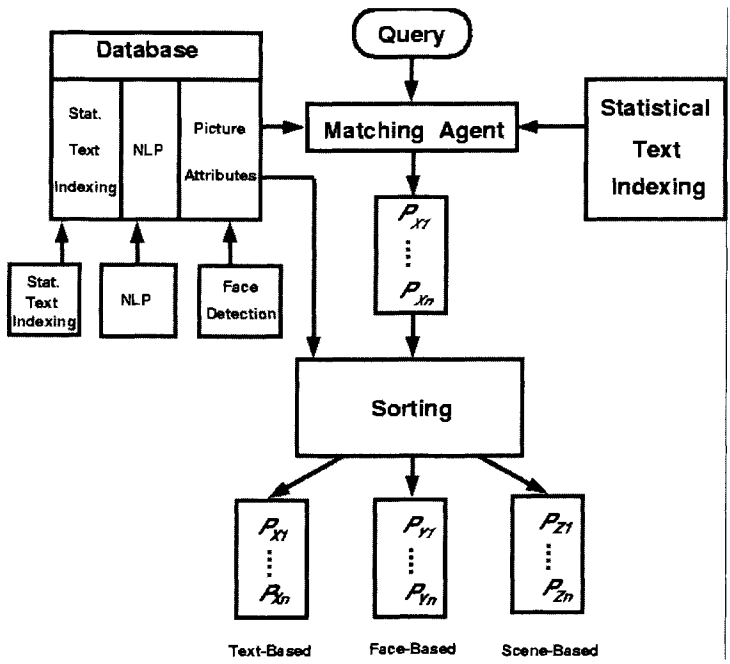

Figure 10. Multiple Ranking Strategy for Combining Text and Image Information in Retrieval.

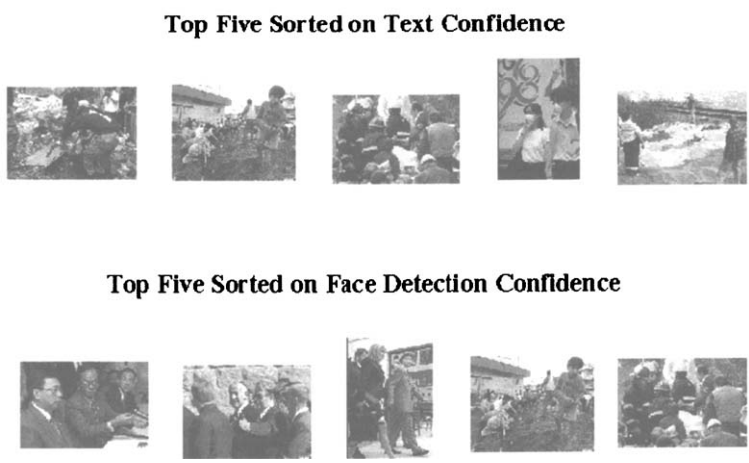

Figure 11. Top Five Images Based on Combining Text Indexing and Face Detection.

Evaluating precision and recall for such a technique is challenging. The precision rate for a given sorting criterion is based on both the text relevance and the presence of the required pictorial attributes (e.g., presence of faces). The text retrieval precision for the top ten images is 90 percent. However, when "presence of faces" is used as a sorting criterion, the precision in the top ten images drops to 40 percent. This is primarily due to the presence of very small faces in the image which are found by the face detector. Since the manual annotators were instructed to disregard faces below a certain size, these are judged to be erroneous (e.g., the last picture in the second row of Figure 11). Thus, assigning relevance judgments based on pictorial attributes must be reinvestigated.

Future Directions

Future directions include improvements on several fronts. First, it is necessary to incorporate information derived from natural language processing as well as statistical image indexing into the retrieval model. Second, the experiments conducted so far have involved only a single query modality, namely text. The next step is to permit multimodal queries, whereby the user can specify an information request using a combination of text (representing contextual constraints) and images (representing exemplars). A relevance feedback mechanism whereby the system can "learn" from user feedback is called for.

Finally, there is a need for more comprehensive testing and evaluation of the techniques developed thus far. The development of evaluation frameworks suitable for multimedia information retrieval systems is still an emerging research area. It is the focus of the MIRA (1999) project,

a consortium of IR researchers in Europe. They make a strong case for *dynamic evaluation* techniques for such applications as opposed to the static evaluation techniques used in text retrieval systems. Rather than evaluating the initial results of a single query, researchers are proposing that the evaluation should be associated with an entire session consisting of continuously refined queries. For example, a monotonically increasing performance curve indicates a good session. They also suggest that new interaction-oriented tasks (apart from search and retrieval) must be supported and evaluated. An example of the latter would be the ability to clarify and formulate information needs.

In this research effort, the following measures of performance are of interest: (1)effectiveness of the ranking scheme generated based on the user's query input and preferences, (2) performance of each individual ranking scheme, and (3) performance of the face detection and recognition modules.

Conclusion

This article has presented a system for searching multimodal documents for pictures in context. Several techniques for extracting metadata from both images and text have been introduced. Two different techniques for combining information from text processing and image processing in the retrieval stage have been presented. This work represents efforts toward satisfying users' needs to browse efficiently for pictures. It is also one of the first efforts to automatically derive semantic attributes of a picture, and to subsequently use this in content-based retrieval. Retrieval experiments discussed in this article have utilized only two of the four indexing schemes that have been developed. These show the promise of integrating several modalities in both the indexing and retrieval stages.

Notes

[1] It is believed that for a typical human face, the ratio of the width to the height of the face is always around the magic value of 2/(1+05), which is called the *golden ratio* (Farkas & Munro, 1987).

References

Aslandogan, Y. A.; Their, C.; Yu, C.T.; Zou, J.; & Rishe, N. (1997). Using semantic contents and WordNet in image retrieval. In *SIGIR '97* (Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia, PA) (pp. 286-295). New York: Association for Computing Machinery.

Chang, C. C., & Lee, S. Y. (1991). Retrieval of similar pictures on pictorial databases. *PatternRecognition, 24*(7), 675-680.

Chang, W. C. (1998). *A framework for global integration of distributed visual information systems.* Unpublished doctoral dissertation, State University of New York, Buffalo.

Charkravarthy, A. S. (1994). Representing information need with semantic relations. In *COLING-94* (The 15th International Conference on Computational Linguistics, August 5-9, 1994, Kyoto, Japan) (pp. 737-741). Morristown, NJ: ACL.

Chellappa, R.; Wilson, C.; & Sirohey, S. (1995). Human and machine recognition of faces: A survey. *Proceedings of the IEEE, 83*(5), 705-741.

Croft, W., & Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 35*(4), 285-295.

Farkas, L. G., & Munro, I. R. (1987). *Anthropometric facial proportions in medicine.* Springfield, IL: Charles C. Thomas.

Fuhr, N., & Buckley, C. (1991). A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems, 9*(3), 223-248.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2d ed.). Boston: Academic Press.

Hobbs, J. R.; Appelt, D.; Bear, J.; Israel, D.; Kameyama, M.; Stickel, M.; Tyson, M. (1997). Fastus: A cascaded finite-state transducer for extracting information from natural language text. In E. Roche & Y. Schabes (Eds.), *Finite-state language processing* (pp. 383-406). Cambridge, MA: MIT.

Ide, E. (1971). New experiments in relevance feedback. In G. Salton (Ed.), *The SMART system: Experiments in automatic document processing* (pp. 337-354). Englewood Cliffs, NJ:Prentice Hall.

Jorgensen, C. (1996). An investigation of pictorial image attributes in descriptive tasks. In B. E. Rogowitz & J. P. Allenbach (Eds.), *Human vision and electronic imaging* (Proceedings of the Society for Optical Engineering (vol. 2657, pp. 241-251). Bellingham, WA: SPIE.

Kartutnen, L., & Beesley, K. R. (1992). *Two-level rule compiler.* Unpublished Xerox PARC Tech. Report No. TR ISTL-92-2.

Meghini, C. (1995). An image retrieval model based on classical logic. In *SIGIR '95* (Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 9-13, 1995, Seattle, WA) (pp. 300-309). New York: Association for Computing Machinery Press.

Merialdo, B., & Dubois, F. (1997). An agent-based architecture for content-based multimedia browsing. In M. T. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 281-294). Cambridge, MA: AAAI Press.

MIRA. (1999). *Evaluation frameworks for interactive multimedia information retrieval applications.* Retrieved July 7, 1999 from the World Wide Web: http://www.dcs.gla.ac.uk/mira.

Robertson, S. (1986). On relevance weight estimation and query expansion. *Journal of Documentation, 42*(3), 182-188.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system: Experiments in automatic document processing* (pp. 313-323). Englewood Cliffs, NJ: Prentice-Hall.

Roche, E. (1997). Parsing with finite-state transducers. In E. Roche & Y. Schabes (Eds.), *Finite-state language processing* (pp. 241-280). Cambridge, MA: MIT.

Romer, D. M. (1993). *A keyword is worth 1,000 images* (Kodak Internal Tech. Rep.). Rochester, NY: Eastman Kodak.

Romer, D. M. (1995). *Research agenda for cultural heritage on information networks.* Retrieved July 7, 1999 from the World Wide Web: http://www.ahip.getty.edu/agenda.

Rosch, E.; Mervis, C. B.; Gray, W. D.; Johnson, D. M.; Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Rowe, N., & Guglielmo, E. (1993). Exploiting captions in retrieval of multimedia data. *Information Processing and Management, 29*(4), 453-461.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Reading, MA: Addison-Wesley.

Smith, J. R. (1997). *Integrated spatial and feature image systems: Retrieval, analysis, and compression.* Unpublished doctoral dissertation, Columbia University, New York.

Smith, J. R., & Chang, S.-F. (1996). VisualSEEK: A fully automated content-based image query system. In *Proceedings of ACM Multimedia '96* (November 18-22, 1996, Boston, MA) (pp. 87-98). New York: Association for Computing Machinery Press.

Srihari, R. K. (1995a). Automatic indexing and content-based retrieval of captioned images. *Computer, 28*(9), 49-56.

Srihari, R. K. (1995b). Use of captions and other collateral text in understanding photographs. *Artificial Intelligence Review, 8*(5-6), 409-430.

Srihari, R. K., & Burhans, D. T. (1994). Visual semantics: Extracting visual information from text accompanying pictures. In *Proceedings of the Twelfth National Conference on Artificial Intelligence* (pp. 793-798). Menlo Park, CA: AAAI Press.

Sundheim, B. (Ed.). (1995). *MUC-6* (Proceedings of the 6th Message Understanding Conference, November 6-8, 1995, Columbia, MD). San Francisco: Morgan Kaufmann.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision, 7*(1), 11-32.

Additional Reference

Bikel, D. M.; Miller, S.; Schwartz, R.; & Weischedel, R. (1997). Nymble: A high-performance learning name-finder. In *Proceedings of the 5th Conference on Applied Natural Language Processing* (March 31-April 3 1997, Washington DC) (pp. 194-201). Boston: MIT Press.