RAMESH JAIN

Visual Information Retrieval in Digital Libraries

The emergence of information highways and multimedia computing has resulted in redefining the concept of libraries. It is widely believed that in the next few years, a significant portion of information in libraries will be in the form of multimedia electronic documents. Many approaches are being proposed for storing, retrieving, assimilating, harvesting, and prospecting information from these multimedia documents. Digital libraries are expected to allow users to access information independent of the locations and types of data sources and will provide a unified picture of information. In this paper, we discuss requirements of these emerging information systems and present query methods and data models for these systems. Finally, we briefly present a few examples of approaches that provide a preview of how things will be done in the digital libraries in the near future.

INTRODUCTION

The nature of documents is rapidly changing. A document in a computer is a combination of text, graphics, images, video, and audio. This revolution in the nature of documents, obviously brought on by the technology now available, has resulted in a major change in the role and nature of libraries. Digital or electronic libraries will allow access to information anywhere, anytime, and in the most desired form. Researchers in digital libraries are developing techniques to cope with this major change in the basic nature and functionality of libraries.

Another major change in the nature of libraries will be due to the amount of information available in a library. In fact, the basic notion of a centralized physical library is slowly disappearing. The World Wide Web has resulted in a transparent linking of worldwide information sources. Most of these information sources on the Web are currently multimedia documents. The amount of information on the Web is already beyond easy access without powerful search tools, and it is increasing exponentially. A library can now be considered a means of access to this vast resource on the Web. The libraries of the future will be similar to the World Wide Web than to the traditional physical library. This change results in some very interesting challenges. The most important challenge is to find the right information in this huge body of data.

In order to search digital libraries, many search tools are emerging, and these have become common on the Web. These search engines, which currently work only for text, help users by preparing data directories

that assist in finding all documents relevant in the context of specified key words. Arguably, search engines have played a significant role in the popularity of the Web. Without these tools, we would be wasting significantly more time chasing links on the Web.

Since most documents are now multimedia, search tools for nontextual information will be required. Without search and organization methods for nontextual information, it will be very difficult to use digital libraries.

In this paper, we discuss visual information retrieval methods. We discuss the nature of visual information (graphics, images, and video) and present the techniques being developed to retrieve this information. In a digital library, these tools will work closely with textual searches. In this paper, however, the focus will be only on visual tools.

INFORMATION IN DIGITAL LIBRARIES

A few decades ago, traditional libraries had only books; they were the only mechanism to store and communicate data, information, and knowledge. Technological advances in several fields have made it possible to store and communicate other forms of knowledge as easily as books. A major reason for this is the multimedia revolution. The power of multimedia systems originates in the fact that disparate information can be represented as a bit stream. This is a big advantage because every form of representation, from video to text, can be stored, processed, and communicated using the same device—a computer. By reducing all forms of information to bit streams, we can start focusing on information rather than the sensor used to acquire it and the communication channel used to transport it. We can also use an appropriate presentation method to supply information to a user.

Most information in computers used to be alphanumeric and was already at a higher symbolic level. In multimedia systems, different types of information—images, text, audio, video, and graphics—are used. These media provide information in disparate representations and at different levels, ranging from signal (audio) to symbol (graphical). To combine and compare two information sources, it is essential that both information sources are understood, and this understanding should be at a level where we can compare and contrast information independent of the original representation medium. This is true for us, and if we want computers to seamlessly deal with disparate information sources, then we will have to do this for computers also. In digital libraries, computers should be able to distinguish each form of information. Strictly syntactic knowledge about video, audio, graphic, or any other form of data makes them just a communication channel. The most attractive feature of current multimedia systems is that, even with very little semantic information,

they make different forms of information available in one environment. This facility is an enormous step in the right direction. By bringing all this information into a computing environment, we are developing systems that can deal with this information in a very flexible way.

Images, video, audio, and other information representation have a large volume of data. Technology is progressing rapidly to deal with the required storage and bandwidth problems. These information sources represent low-level information. When considered as a bit stream with the meta information, the explicit semantic information content in these sources is very low. This poses a serious problem in accessing these information sources. Humans are very efficient in abstracting information and then interacting with humans and other devices at a high level. This allows high bandwidth interactions among humans and between human and machines.

Multimedia systems currently have this semantic bottleneck. Techniques must be developed to add semantics to the data acquired from disparate sources in disparate forms. Since documents in digital libraries will be complex, tools to deal with information independent of its overall representation will become essential. A user may just ask a question and the answer may be available in the library in either text, image, graphics, or tabular form. The answer must be provided independent of the representation. In some cases, the answer may be partially available in different forms, and then these partial results must be combined at different information levels to provide the answer.

In this discussion, we consider the semantics of image and video data. No effort is made to address techniques that combine partial information from several sources. We will focus on how to represent information in images and how to organize images and related information to provide answers to a user from a database.

NEW DATABASE OPERATIONS

A digital library should allow the storage, communication, organization, processing, and envisioning of information. It should facilitate interactions by using natural interactions, which include multimedia input and output devices and use of high-level domain knowledge by a user.

Domain knowledge should be so much a part of a system such that a user feels that the system is an intelligent aide. A user should be able to articulate queries using terminology commonly used in his field and should not have to worry about the organization of information in the system.

The system should allow for powerful navigation tools. The user will use vague natural language, and that should be understood by the system to let a user navigate through the system. The nature of queries will be

fuzzy not due to the laziness of the user but due to the nature of information and the size of the database. A general query environment will be like the one shown in figure 1. A user looking for certain information, for example, about a person who he vaguely recalls, specifies important things he remembers about the person. This specification may be that she has big eyes, wide mouth, long hair, and a small forehead. Based on this information, candidate people's pictures are retrieved. The user can then select the closest person that matches the query and modify the query by either specifying features or by using graphical and image-editing tools on the photo. This refines the query image, which is then sent to the system to provide new candidates to satisfy the query. Thus a query is incrementally formulated starting with the original vague idea. This process will terminate when the user is satisfied.

Due to the nature of data, several levels of abstraction in the data, and temporal changes in the data, the types and nature of interactions in such systems will be richer than those in a database or image processing system. We loosely refer to all interactions initiated by a user as queries. The types of queries in such systems can be defined in the following classes:

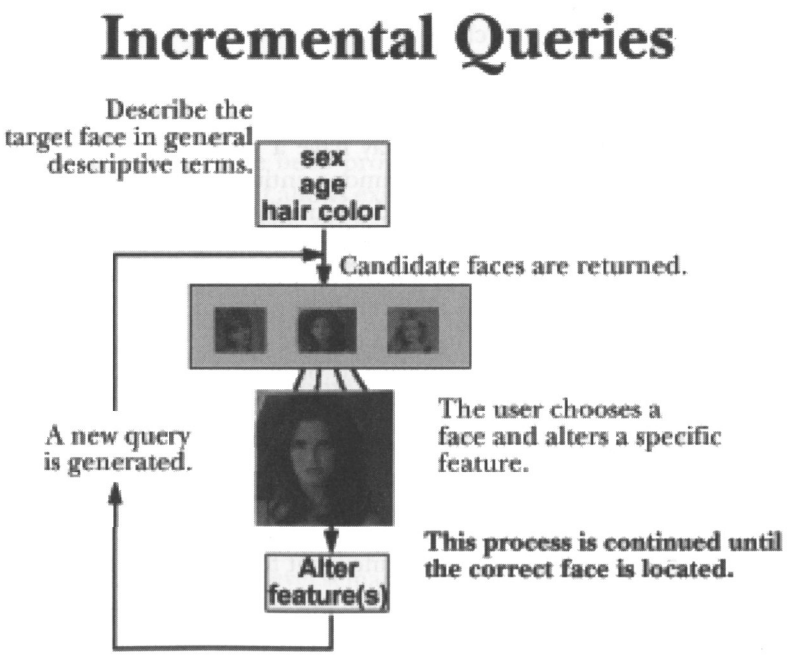

Figure 1. This figure shows that the queries in digital libraries will be incremental in nature. These queries will facilitate navigation and browsing of data.

1. *Search.* Search is one of the most commonly used operations in digital libraries. A user may want to search a library for specific images or documents containing some pictorial information. Tools should be provided to search based on this information. As discussed later, meta features are used to provide some information about images. In many applications, queries can be formulated to search specific images using only meta data. These queries can be answered, in most cases, using conventional database queries. In fact, many early image databases and browsers were designed using this approach.

   A search based on some image or object attributes is more common. To answer these queries, one may have to use visual attributes of images for the search. A major difference in these queries will be the fact that similarity becomes a central operation rather than conventional matching. Techniques to evaluate similarity are an active research topic in many fields of science and technology (Santini & Jain, in press). Many approaches have been proposed to compare several attributes to evaluate the similarity of two objects. In addition to the decision on what attributes to select, a very difficult decision is how to combine those attributes. Methods to combine attributes are domain dependent and subjective. It is clear, however, that in dealing with images and similar data sets, similarity rather than matching will be a key function in searching.

2. *Browse.* When a user approaches a library, the most common operation is browsing documents to locate those that may contain the information of interest. A user may have a vague idea about the attributes of an entity, relationships among entities in an image, or overall impression of an image. Such ideas are formed due to the overall appearance of the image rather than very specific objects and relations among them. In such cases, the user may be interested in browsing the database based on an overall impression or appearance of images rather than searching for a specific entity. The system should allow formulation of fuzzy queries to browse through the database. In browsing mode, there is no specific entity for which a user is looking. The system should provide data sets that are representative of all data in the system. The system should also keep track of what has been shown to the user. Some mechanism to judge the interest level of the user in the data displayed should be developed and this interest level should be logged to determine what to display next.

3. *Temporal Events.* It is estimated that videos will be a major source of information in digital libraries. The number of videos has been rapidly increasing, and video is becoming an integral part of compound documents. In video sequences, one may want to retrieve images based on some events taking place in the sequence. A typical query of this

type may be: Show me all sequences in which player *X* was blocked by player *Y*. These queries will require temporal analysis of video sequences in terms of the events of interest. Some primitive spatio-temporal features must be computed and stored in the database to answer questions concerning events of interest to users.

Abstractions in spatio-temporal space are not yet understood well enough to automatically extract them from video sequences. Though some techniques have been developed to represent relative time ordering of two events, representations for abstraction of events need to be developed to allow users to articulate questions related to temporal events.

4. *Integrated Queries.* Users are interested in getting information independent of the medium. Thus, in a document, the requested information may be either in text, image, graphics, or video form, and the system should provide the information without a user knowing the medium. This facility will require an abstraction of information from every media into one unified representation. We do not know of any efforts being made in this area yet.

DATA MODEL AND VISUAL FEATURES

Information in an image exists at several abstraction levels and should be accessible at these levels. The data model used to store this information must allow the existence of information at these multiple levels. Several data models have been proposed (e.g., see Gupta, 1991). Here we discuss one model that allows explicit representation of abstract levels in images. The VIMSYS data model uses a hierarchical representation of data using various levels of semantic interpretation that may satisfy the needs of digital libraries (Gupta et al., 1991). This data model is shown in Figure 2. At the image representation (IR) level, the actual image data are stored. Image objects (such as lines and regions) are extracted from the image and stored in the image object (IO) layer with no domain interpretation. Each of these objects may be associated with a domain object (DO) in that layer. The semantic interpretation is incorporated in these objects. The domain event (DE) layer can then associate objects of the DO layer with each other, providing the semantic representation of spatial or temporal relationships. This hierarchy provides a mechanism for translating high-level semantic concepts into content-based queries using the corresponding image data. This allows queries based on object similarity to be generated without requiring the user to specify the low-level image structure and attributes of the objects. Another very important aspect of this representation is that the first two levels, IR and IO, are domain-independent levels and the other two, DO and DS, are

domain-dependent levels. We do not know any system yet where this goal of clearly organizing domain-dependent and domain-independent components can be cleanly partitioned and implemented. We believe, however, that this is a worthwhile target. The architecture discussed below is motivated by this desire.

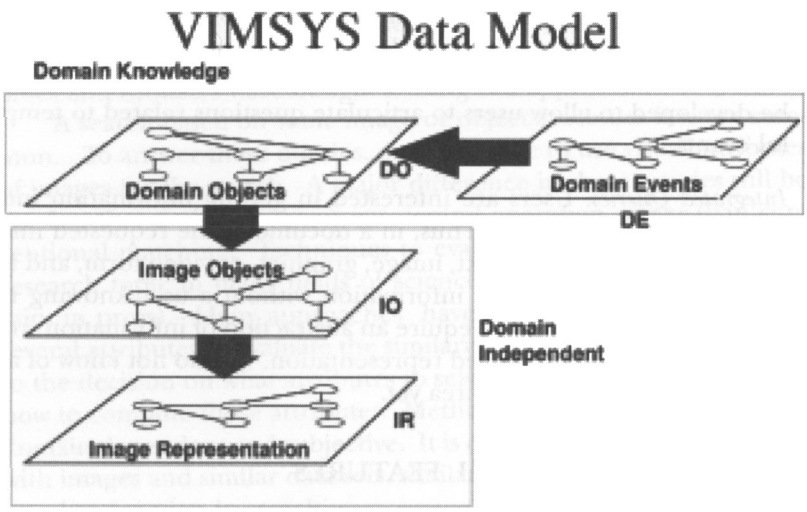

Figure 2. A four-level data model to capture different levels of abstractions in visual information systems is shown here. The image levels are domain independent, the other two levels depend on the domain.

These ideas have been used to develop several systems for the retrieval of images and video information in our group (Gupta et al., 1991; Bach et al., 1992; Swanberg et al., 1993a). Here we discuss each of the system components briefly. In the following discussion, we will discuss this architecture in the context of images and video, but our concepts are applicable to any kind of data.

TYPES OF FEATURES

Features must be extracted from input images and stored in the database. As is well known, different applications may require different features (Jain et al., 1995). Since the features must be stored at the time of data entry, one must carefully decide which features will be used in a system. We consider that all features must be classified in one of the following classes:

1. F_u. This set contains the features which are commonly referred to as meta-features. Some of these features can be automatically acquired

from the associated information on images. These features may include the size of the image, photographer, date taken, resolution, and similar additional information. This group also contains other features that can be called user-specified. Values are assigned to these features by the user at the time of insertion. Many of these features can be read by the system either from the header, file name, or other similar sources. These features cannot be directly extracted from images.

2. F_d. This set contains the features which are derived directly from the image data at the time of insertion of the images in the database. Values are automatically calculated for these features using automatic or semiautomatic functions. These features are called derived features and include those that are commonly required in answering queries. These features are stored in the database.

3. F_c. This set contains the features whose values are not calculated until they are needed. Routines must be provided to calculate these values when they become necessary. These features may be computed from data at the query time. These are called query-only features or computed features.

The first two types of features are actually stored in the database. Metadata can be frequently read from other sources or should be manually entered. Which feature should be in F_d and which should be in F_c is an engineering decision. One must study frequently asked queries and determine the required features. This determines the set to which a particular feature should belong.

The system interface encourages users to formulate queries using metadata and derived features as much as possible. It reluctantly allows use of computed features. To access data, the system can purge the search space significantly using metadata and derived features and then apply computed features to only this reduced set of images. This strategy allows flexibility while maintaining a reasonable response time. The system may be able to predict wait time using number of images from which computed features must be extracted.

INTERFACES

Users of digital libraries will have disparate backgrounds. As a result, the interfaces to these libraries should be such that any novice can use intuitive methods. The operations used in these interactions must require almost no knowledge of the organization of the data and information. Many of these operations cannot be conveniently performed using traditional interfaces. Here we discuss some general issues in

designing interfaces for digital libraries. Interactions with libraries are likely to be multimedia. Due to the nature of the data and several abstraction levels, it is expected that users will require multimodal interface mechanisms.

Our focus on visual queries in digital libraries must allow facilities to formulate the following interactions:

- *General Search*: In general, there will be two modes of navigation in libraries: locating and browsing. In the location mode, a user knows what he or she wants and the goal of the queries will be to get precisely that information. Also in the location mode, many queries may be symbolic because what is required can be articulated using meta data. Some location queries may require visual data. It is expected that search queries will deal mostly with meta data. For these queries, some query language, possibly a variant of SQL, may be used.

- *Query by Pictorial Example (QPE)*: A very powerful expression of a query is to point to a picture and expect that the system will show all pictures similar to the example. This approach is easy to use but very complex to implement. The system must use certain features and some similarity measures to evaluate other pictures that are similar to the example. Effectively, the system must rank all data with respect to the example and then display pictures that are closest to the example. Interestingly, this has been very popular in designing image databases (Niblack et al., 1993).

  In QPE, features and similarity measures must be clearly defined for use in retrieving images. Similarity judgment has been a difficult problem and continues to attract the attention of several researchers (Santini & Jain, in press). The most interesting fact about similarity measures is that they are domain dependent and very subjective. Assuming that we have identified a measure that is acceptable to a user for his or her domain, we face some interesting problems in QPE. All images are compared to the example to evaluate their similarity. This is possible in those cases where the size of the database is such that computations can be done in a reasonable time. When the size of the database grows such that it is not possible to accommodate all data in main memory and such computations become impractical, one must resort to indexing techniques.

  Indexing techniques for spatial data have been developed (Jagadish, 1991; Niblack et al., 1993; Samet, 1984). These techniques are very limited when it comes to addressing the problem of similarity indexing. Techniques like TV-trees are a good step in the right direction but lack several important features (Linet al., 1994).

- *Query Canvas*: Queries may be formulated by starting with an existing picture, scanning a new picture, and modifying these by using the visual and graphical tools available in common picture editing programs, such as Adobe Photoshop. One may cut and paste from several images to articulate a query in the form of an image. It is also possible to start from a clean image and then draw an image using different drawing tools. The basic idea in this approach is to provide a tool to define a picture that may be used in a QPE. This approach allows a user to define a picture that they are looking for using visual tools. This will provide users with a visual query environment.
- *Containment Queries*: In many cases, a user may point to an object or circle an area in an image and request all images that contain similar regions. These queries seem simple and will be if complete segmentation of images is performed and all region properties are stored. Most image database systems store only global characteristics of an image. In these cases, one is looking for all images that are a superset of the region attributes. Once all such images are retrieved, some other filtering techniques could be developed to solve this problem.
- *Semantic Queries*: All the above queries were based on image attributes. In most applications, an image database is likely to be prepared for a specific domain-dependent application, such as human faces, icefloe images, or retinal images. It is important that users can then interact using domain-dependent terms. It is common that people may describe a person using terms like big eyes, wide mouth, small ears, rather than the corresponding image objects.

  Semantic queries require extensive use of domain knowledge. Domain knowledge is necessary both in defining features that will be used by the system and in interpreting user queries. Most image database systems either considered domain knowlede implicitly by defining features or ignored it (Faloutsos et al., 1994). The role of explicit knowledge in image databases is discussed in (Gupta et al., 1991; Swanberg et al., 1993a; Swanberg et al., 1993b).
- *Object Related Queries*: These queries are semantic and ask for the presence of an object. These queries may deal with three-dimensional objects. Since three-dimensional objects are difficult to recognize using automated techniques, these queries may become very complex. Three-dimensional object recognition is a very active research area in machine vision. Queries based on recognizing objects in a query image may be, therefore, very difficult to execute.
- *Spatio-Temporal Queries*: In video sequences, and in many other applications where pictures are obtained over a long period, a user may want to get answers to some spatio-temporal events and concepts. Answers to such questions may require complete analysis of all video

sequences and storing some important features from there. Considering the fact that methods to represent temporal events are not well developed yet, this area requires much research before one can design a system to deal with spatio-temporal queries at the natural language level.

EXAMPLE SYSTEMS

In this section, we present some emerging approaches in visual information retrieval. The information may be retrieved either based on global image properties or on object characteristics. We discuss approaches for both these systems and present example systems.

Image Databases

When one looks at an image, some global impression is formed. This impression is based on some general characteristics of images. Even in those cases where one may be interested in objects in images, global characteristics may help. This is due to the fact that, in the context of the library, many images will contain only one object of interest, and this object will be photographed in relatively controlled conditions. Thus, one may design a powerful system just by considering basic image features. Some very basic image features are color, texture, and shape. Grayscale images are considered here as a special case of color. On first examination, it appears that one should consider attributes of objects in images. From machine vision literature, it is clear that segmentation is a difficult problem (Jain et al., 1995). While dealing with a diverse set of images which are acquired under varying conditions, segmentation may be very difficult. In such cases, one may want to completely ignore domain knowledge and build a database only using image attributes. These attributes may be computed for complete images or for their predefined areas.

Many systems have been designed using image-only attributes. QBIC from IBM uses color, texture, and manually segmented shapes (Faloutsos et al., 1994). QBIC was the first complete system to demonstrate the efficacy of simple attributes in appearance-based retrieval of images from a reasonably sized database. The use of shape in QBIC is problematic, however. Shape is defined for individual segments which must be obtained manually. This also creates an artificial situation in the database because, for each manually obtained segment, one must consider a separate record in the database. Thus if an image has N objects, the database must contain $N+1$ records—one for the image and one each for N segments. Shape measures on complete images are not satisfactory because shape is defined for an image region. Some heuristics have been proposed, but much remains to be done in this area.

Color is considered a global characteristic. Most systems rely on color histograms. Some kind of histogram matching is done to determine similarity of two images. Histogram-based approaches clearly ignore spatial proximity of colors and hence may result in erroneous results. In most cases, however, histogram-based matching is quite effective.

Texture poses a more difficult problem. Most systems use global measures of texture and try to assign some texture attribute to images. These attributes are then used for evaluating similarity of texture in images. Most images contain different types of texture in different parts of the image. The global texture attributes, therefore, could be misleading. These systems use only the first two levels of the VIMSYS data model. Since both of these—IR and IO—levels are domain independent, these image databases are domain independent. Users of these systems must supply the semantics in these systems. The semantics can be provided by using color and texture attributes of objects of interest. One may filter using these attributes and then use domain-dependent features on remaining images to retrieve desired information.

An example of an image database that provides tools to organize and retrieve information using image level information is the PinPoint system

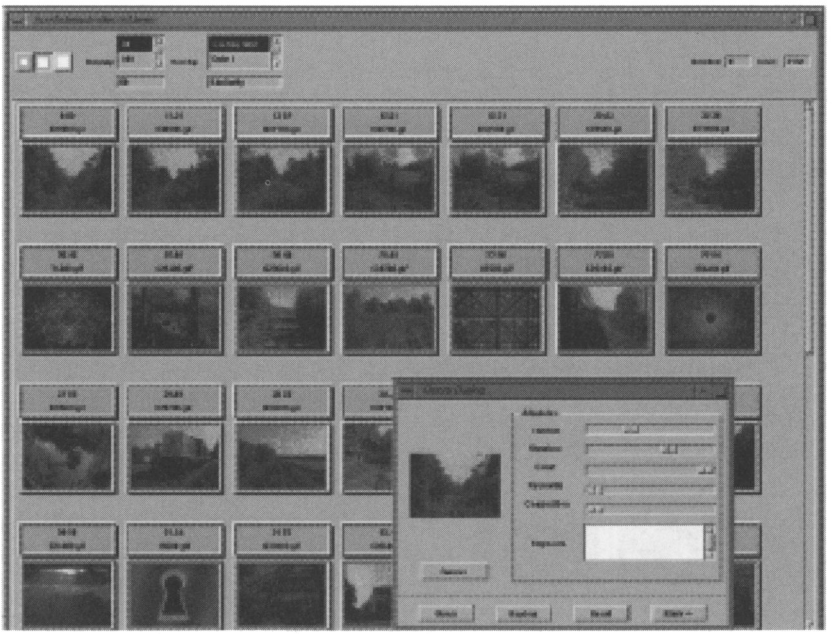

Figure 3. A screen shot of PinPoint showing the query window and all images retrieved using QPE. Notice that a user can adjust weights of features and the feedback to the user is instantaneous.

developed at Virage. This system extracts features to characterize images using color, texture, structure, and composition. These features can be combined using distance functions. This system treats keywords also like features by using a thesaurus to compute distances between keywords in the query and stored images. The weights of the features can be changed to retrieve similar images using different similarity functions. We show a screen shot of this system in Figure 3. This shot shows all images retrieved as similar to the example image, which is the best matching image and hence appears as the first image in similar images. If the images are created using the query canvas, shown in Figure 4, then one can articulate a query by cutting and pasting and by other image manipulation operations. It must be mentioned that this system has no domain-level knowledge.

Interestingly, even without any domain knowledge in this system, users very quickly learn to retrieve images of their choice by using an example image and appropriate weights of the features provided in the system. The system uses color, texture, and structure as features of an image. In color, both global colors, and automatically segmented segments and their locations, defined as composition, are used. For texture, several properties

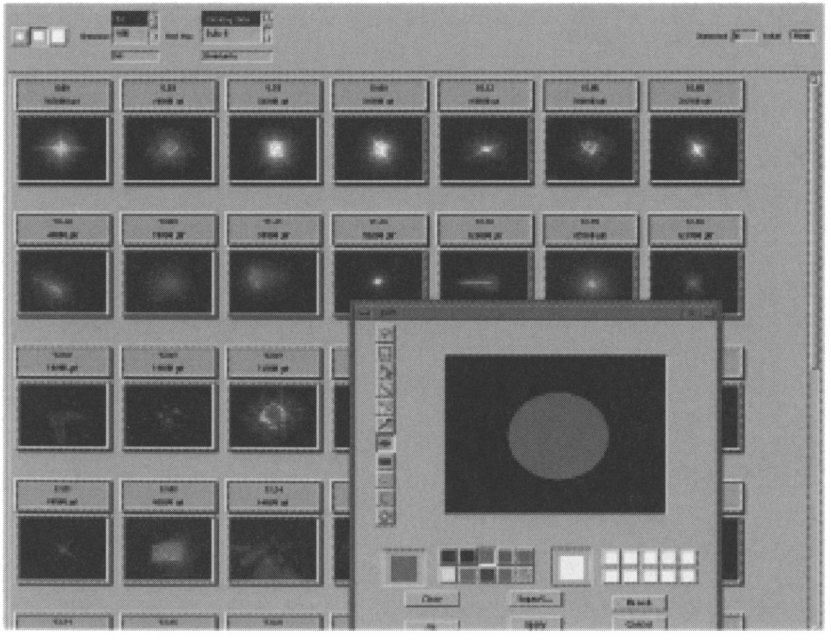

Figure 4. The query canvas allows a user to articulate a query using visual means. One can cut and paste from images and use image manipulation programs to articulate a query.

are computed using standard texture features and are combined to represent an overall measure of the texture. Structure addresses shapes and location of edge segments. It is interesting to see that these purely image-based features, when combined with hand drawn queries on a canvas or an image selected for QPE, perform quite effectively in retrieving semantically relevant objects. This strongly suggests that by defining a pictorial alphabet and suitable rules to use this alphabet, it may be possible to develop powerful domain-dependent systems.

Semantic Knowledge: The Xenomania System

One can use domain knowledge to extract features at insert time and interpret user queries using domain knowledge and statistical characteristics of the information in the database. Many projects in academia and industry address these issues. Here we demonstrate some of these ideas using a face retrieval system called Xenomania implemented at the University of Michigan (Bach et al., 1992).

Xenomania was an interactive system for the retrieval of face images and information. It allows a user to locate a specific person in the database and retrieve the person's image and other information. The user can describe the target face in general terms—e.g., shape of eyes, nose, length of hair—to begin the location process and to retrieve the initial results. After that, the target face may be described using these general terms or by using the actual image contents of the retrieved faces. All aspects of the architecture described above are incorporated into this system (this system is described by Bach et al. [1992]).

We chose the interactive face identification problem because of the lack of well-defined image objects and features and the heavy dependence on both predefined domain knowledge and extensive user participation. Although much work has been done toward modeling of facial features, it is still very difficult to accurately extract and evaluate these features over a variety of faces and situations and even more difficult to assign semantic attributes to these features which are meaningful to users. This application exploits the demands for both extensive predefined domain knowledge and user-incorporated knowledge at every step of processing.

Xenomania relied very heavily on previous research in the field of face recognition. Much work was done regarding the psychological aspects of face recognition which provided a basis for our initial implementation. Many automatic face recognition systems have also been developed. The Xenomania project, however, was not a face recognition system but rather an image database system used for interactive face retrieval. Some face-recognition systems have approached the problem from strictly an image processing point of view with little or no emphasis on descriptive representation of faces. These systems do not incorporate the user

for describing the face or guiding the query refinement once the recognition process has been initiated. The most successful face reocgnition system is based on eigenfaces (Pentland, Moghaddam, & Starner, 1994; Pentland, Picard, & Sclaroff, 1994). This system is also influenced by image recognition approaches. In an eigenface-based system, one can specify an image and the system will retrieve all images that are similar to that. It may be interesting to combine eigenfaces with the descriptive approach used in Xenomania.

Domain Knowledge

As in any image management application, we are faced with the difficulty of determining which attributes are important for each domain object, and how to accurately represent these attributes in the system. However, this is an attractive problem from our point of view, because it gives us the opportunity to investigate different types of object and feature representations. For instance, there are several attributes about an eye that may be important. Individual eye attributes such as area and width will be necessary, as will relative attributes such as the width of the eye compared to the height of the eye. Spatial attributes such as distance between the left eye and the right eye are also important and must be incorporated into the system. Other objects, such as eyebrows, may require entirely different attributes than those for eyes to be maintained in the system. We have based much of our initial implementation on research that has been done to evaluate which facial features and attributes are best suited for face identification and differentiation.

Many image databases are likely to be for specific applications and hence will require strong domain knowledge. The domain objects should be described using the image alphabet or image objects in the VIMSYS model. This task will require close interactions among database designers, image processing experts, and domain experts.

VIDEO DATABASES: TV NEWS ON DEMAND

Video is rapidly becoming the preferred mode of receiving information and video is certainly the most vivid medium for conveying information. Video has gained tremendous popularity since it appeared on the scene. As is well known, television has been one of the most influential inventions of this century. As a result, the last decade has seen rapid growth in camcorder use in all aspects of human activities.

Video is the most impressive medium for communicating and recording events in our life. Its use is limited, however, by its basically sequential nature. To access a particular segment of interest on a tape, one must spend significant time searching for the segment. Video databases have potential to change the way we access and use video.

By storing each individual shot in the database, one can then access any individual frame based on the content of the shot. Each shot can be analyzed to find what is contained in each shot. Frames in each shot can be analyzed to find events in it. By segmenting videos into shots and analyzing those shots, one can extract information that can be put into a database. This database can then be searched to find sequences of interest.

Video databases can be useful in many applications. One application is news on demand. Suppose that each sequence is analyzed and the information in it is stored in a database with pointers to the relevant frames. This database then can be used to view the news of choice to the depth desired by a user and in the sequence desired. We are implementing such a system in our laboratory (Swanberg et al, 1993a; Swanberg et al., 1993b; Hampapur et al., 1994a; Hampapur et al., 1994b). Details of segmentation of the sequence, architecture of the system, role of knowledge in such a system, and all other aspects have been presented in Swanberg et al., 1993a; Swanberg et al., 1993b; and Hampapur et al., 1994. It must be mentioned here that many other systems of this type are being implemented in other places.

The architecture for the video database is composed of four major components: input, database, query environment, and knowledge base. The input module is further divided into two major components: a sequence segmentation subsystem and a feature detection subsystem. The knowledge module has a video object schema definition subsystem to help a user enter knowledge into the system for a specific application. The video object schema definition subsystem provides tools to model the video object schema for an application based on the operators available in the input and query processing systems. Based on the video object schema, the feature detection subsystem analyzes a video frame sequence to extract structure and the semantic information about each object of interest in the video. The extracted objects and related semantic information are then stored in the feature database. According to the video object schema definition, a user query interface is automatically customized. A user can also navigate the video object schema defined from the video object schema definition subsystem as well as its associated video object data through the user query interface.

CONCLUSION AND FUTURE RESEARCH

We have discussed some basic issues in visual information retrieval and presented some example systems. As is clear from these examples, these systems are in the early stages of development, but there is growing research interest in this area. Many powerful approaches are being developed for image and video databases. It is clear that these approaches should work very closely with textual and audio search techniques. We

believe that, as research in these areas progresses, we will see the emergence of powerful multimedia information retrieval techniques. These techniques will allow a user to articulate their queries using the medium of their choice and will retrieve information from distributed multimedia libraries. The next few years are likely to result in significant progress in this area.

ACKNOWLEDGMENTS

The research and ideas presented in this paper evolved during collaborations with several people in the InfoScope project. I am thankful to everyone who actively participated in the project. I want to particularly thank Jeff Bach, Shankar Chatterjee, Amarnath Gupta, Arun Hampapur, Bradley Horowitz, Arun Katkere, Don Kuramura, Saied Moezzi, Edna Nerona, Simone Santini, Chiao-Fe Shu, Deborah Swanberg, David White, and Terry Weymouth for collaboration in different aspects of this work.

REFERENCES

Bach, J.; Paul, S.; & Jain, R. (1992). An interactive image management system for face information retrieval. *IEEE transactions on knowledge and data engineering, 5*(6), 619-628.

Faloutsos, C.; Barber, R.; Flickner, M.; Hafner, J.; Niblack, W.; Petkovic, D.; & Equitz, W. (1994). Efficient and effective querying by image content. *Journal of Intelligent Information Systems, 3*(3/4), 231-262.

Gupta, A.; Weymouth, T.; & Jain, R. (1991). Semantic queries with pictures: The VIMSYS model. In *Proceedings of the 17th International Conference on Very Large Databases* (pp. 69-79). Barcelona, Spain: Morgan Kaufmann Publishers.

Hampapur, A.; Jain, R.; & Weymouth, T. (1994a). Digital video indexing in multimedia systems. In *Proceedings of the Workshop on Indexing and Reuse in Multimedia Systems. Proceedings of the 12th National Conference on Artificial Intelligence.* Seattle, WA: American Association of Artificial Intelligence Press.

Hampapur, A.; Jain, R.; & Weymouth, T. (1994b). Digital video segmentation. In *Proceedings of the ACM Conference on Multimedia.* Association of Computing Machinery.

Jagadish, H. V. (1991). A retrieval technique for similar shapes. In *Proceedings of the ACM SIGMOD International Conference on the Management of Data* (pp. 208-217). Denver, CO: ACM.

Jain, R.; Kasturi, R.; & Schunck, B. (1995). *Machine vision.* New York: McGraw-Hill, Inc.

Lin, K-I.; Jagadish, H. V.; & Faloutsos, C. (1994). The TV-tree: An index structure for high-dimensional data. *VLDB Journal, 3*(4), 517-542.

Niblack, W.; Barber, R.; Equitz, W.; Flickner, M. D.; Glasman, E. H.; Petkovic, D.; Yanker, P.; Faloutsos, C.; & Taubin, G. (1993). The QBIC project: Querying images by content, using color, texture, and shape. In W. Niblack (Ed.), *Storage and Retrieval for Image and Video Databases: Vol. 1908: SPIE Proceedings* (pp.173-187). Bellingham, WA: SPIE.

Pentland, A.; Moghaddam, B.; & Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 84-91). Seattle, WA: IEEE Press.

Pentland, A.; Picard, R. W.; & Sclaroff, S. (1994). Photobook: Tools for content-based manipulation of image databases. In W. Niblack & R. Jain (Eds.), *Storage and Retrieval for Image and Video Databases II: Vol. 2185: SPIE Proceedings* (pp. 34-47). Bellingham, WA: SPIE.

Samet, H. (1984). The quadtree and related hierarchical data structures. *Computing Surveys, 16*(2), 187-260.

Swanberg, D.; Shu, C. F.; & Jain, R. (1993a). Architecture of a multimedia information system for content-based retrieval. In P. V. Rangan (Ed.), *Network and operating system support for digital audio and video: Third international workshop proceedings.* Berlin: Springer-Verlag.

Swanberg, D.; Shu, C. F.; & Jain, R. (1993b). Knowledge-guided parsing in video databases. In *Electronic imaging: Science and technology proceedings* (pp. 13-24). San Jose, CA: IST/SPIE.

Swanberg, D.; Weymouth, T.; & Jain, R. (1992). Domain information model: An extended data model for insertions and query. In *Proceedings of the Multimedia Information Systems* (pp. 39-51). Tempe, AZ: Arizona State University, Intelligent Information Systems Laboratory.

White, D., & Jain, R. (1996). Similarity indexing with the SS-tree. In *Proceedings of the IEEE 12th International Conference on Data Engineering* (pp. 516-523). Los Alamitos, CA: IEEE.