Techniques and Systems for Image and Video Retrieval*

Y. Alp Aslandogan, Clement T. Yu

Department of EECS, University of Illinois at Chicago

{yaslando,yu}@eecs.uic.edu

Abstract

Storage and retrieval of multimedia has become a requirement for many contemporary information systems. These systems need to provide browsing, querying, navigation and sometimes composition capabilities involving various forms of media. In this survey we review techniques and systems for image and video retrieval. We first look at visual and non-visual features for image retrieval and techniques for using them. Temporal aspects of video retrieval are discussed next. We review several research and commercial systems including WWW-based systems and conclude with future directions.

1 Introduction

The increasing availability of multimedia information combined with the decreasing storage and processing costs have changed the requirements on information systems drastically. Today, general purpose database systems are incorporating support for multimedia storage and retrieval, as well as features which used to be found in specialized imaging systems or multimedia databases.

Increased use of multimedia has important implications for overall information system design regarding storage, processing, retrieval and transmission. In this paper, we provide an overview of techniques for convenient access to images and video; in addition, several state-of-the-art systems are sketched.

In the following section we describe what visual and non-visual features are used for image retrieval and the querying primitives that make use of these features. In Section 3 we focus on techniques specific to video retrieval. In light of these discussions, we look at some of the currently available image and

*Research supported in part by NSF grant IRI-9509253.

video retrieval systems in Section 4. Systems for image retrieval on the World Wide Web are dealt with separately in subsection 4.3. We discuss future directions and conclude in Section 5.

2 Image Retrieval

Image retrieval is concerned with retrieving images relevant to users' queries from a large collection. Relevance is determined by the nature of the application. In a fabric image database, relevant images may be those that match a sample in terms of texture and color. In a news photography database, date, time and the occasion in which the photograph was taken may be just as important as the actual visual content. Many relational database systems support fields for binary large objects (BLOBs) and facilitate access by user-defined attributes such as date, time, media type, image resolution, and source. On the other hand, content based systems analyze the visual content of images and index extracted features. We are also seeing a rapid emergence of object oriented and extensible relational database systems which offer standard database features and support user defined procedures.

2.1 Visual Content Based Image Retrieval

Visual content can be modeled as a hierarchy of abstractions. At the first level are the raw pixels with color or brightness information. Further processing yields features such as edges, corners, lines, curves and color regions. A higher abstraction layer may combine and interpret these features as objects and their attributes. At the highest level are the human level concepts involving one or more objects and relationships among them. An example concept might be "a person giving a speech." Although automatic detection and recognition methods are available for certain objects and their attributes, their effectiveness is highly dependent on image complexity. Most objects, attribute values and high-level concepts cannot be extracted accurately by automatic methods. In such cases semi-automatic methods or user-supplied keywords and annotations are employed. In the following we describe the various levels of visual features and the techniques for handling them.

2.1.1 Visual Features: Color, Texture and Shape

Color plays a significant role in image retrieval. Different color representation schemes include red-green-blue (RGB), chromaticity and luminance system of CIE (International Commission on Illumination),

hue-saturation-intensity (HSI), and others. The RGB scheme is most commonly used in display devices. Hence digital images typically employ this format. HSI scheme more accurately reflects the human perception of color.

All perceivable colors can be reproduced by a proper combination of red, green and blue components. A 24-bit per pixel RGB color image may have 2^{24} or approximately 16.7 million distinct colors. In order to reduce the number of colors for efficiency in image processing, colors are quantized with a suitable algorithm.

Texture is a visual pattern where there are a large number of visible elements densely and evenly arranged. A texture element is a uniform intensity region of simple shape which is repeated. Texture can be analyzed at the pixel window level or texture element level. The former approach is called *statistical* analysis and the latter *structural* analysis. Generally, structural analysis is used when texture elements can be clearly identified, whereas statistical analysis is used for fine (micro) textures [TT90].

Statistical measures characterize variation of intensity in a texture window. Example measures include contrast (high contrast zebra skin versus low contrast elephant skin), *coarseness* (dense pebbles vs coarse stones), *directionality* (directed fabric versus undirected lawn). Fourier spectra are also used to characterize textures. By obtaining the Fourier transform of a texture window, a signature is generated. Windows with same or similar signatures can be combined to form texture regions.

Structural texture analysis extracts texture elements in the image, determines their shapes and estimates their placement rules. Placement rules describe how the texture elements are placed relative to each other on the image and include measures such as the number of immediate neighbors (connectivity), the number of elements in unit space (density), and whether they are layed out homogeneously (regularity). By analyzing the deformations in the shapes of the texture elements and their placement rules, more information can be obtained about the scene or the objects in it. For instance, a density increase and size decrease along a certain direction might indicate an increase in distance in that direction.

Shape-based image retrieval is one of the hardest problems in general image retrieval. This is mainly due to the difficulty of segmenting objects of interest in the images. Consequently, shape retrieval is typically limited to well distinguished objects in the image [FBF+94, PPS94].

In order to detect and determine the border of an object, an image may need to be preprocessed. The preprocessing algorithm or filter depends on the application. Different object types such as skin lesions, brain tumors, persons, flowers, and airplanes may require different algorithms. If the object of

interest is known to be darker than the background, then a simple intensity thresholding scheme may isolate the object. For more complex scenes, noise removal and transformations invariant to scale and rotation may be needed. Once the object is detected and located, its boundary can be found by edge detection and boundary-following algorithms [SHB93]. The detection and shape characterization of the objects becomes more difficult for complex scenes where there are many objects with occlusions and shading.

Once the object border is determined the object shape can be characterized by measures such as area, eccentricity (e.g. the ratio of major and minor axes), circularity (closeness to a circle of equal area), shape signature (a sequence of boundary-to-center distance numbers), shape moments[1], curvature (a measure of how fast the border contour is turning), fractal dimension (degree of self similarity) and others. All of these are represented by numerical values and can be used as keys in a multidimensional index structure to facilitate retrieval.

2.1.2 Indexing and Retrieval

For indexing visual features a common approach is to obtain numeric values for n features and then represent the image or object as a point in the n-dimensional space. Multidimensional access methods such as K-D-B-trees, quad-trees [PF95, Sam89], R-trees [Gut84] or their variants (R*-trees, hB-trees, X-trees, TV-trees, SS-trees, SR-trees, etc.), are then used to index and retrieve relevant images. Three problems need to be solved for this scheme to work properly: First, most multidimensional methods work on the assumption that different dimensions are independent, and hence the Euclidean distance is applicable. Second, unless specifically encoded, feature layout information is lost. In other words, the locations of these features can no longer be recovered from the index. The third problem is the number of dimensions. The index structures become very inefficient as the number of dimensions grow. In order to solve these problems, several approaches have been developed. We first look at the color indexing problem. Texture and shape retrieval share some of these problems and similar solutions are applicable.

A. Color Indexing and Retrieval

A color histogram records the number of pixels in an image for each color. Two color histograms can be compared by summing the absolute differences or the squared differences of the number of pixels in

[1] For a description of various shape moments see for instance [SHB93].

each color. Such a scheme is simple and tolerant to small changes in a scene. However, it suffers from all three of the problems mentioned above.

Cross Correlation

The similarity of color i and color j may not be 0 even though $i \neq j$. Many colors that are very different in terms of their RGB values may be perceived as similar by humans. Niblack et al [FBF+94] use a matrix in which each entry a_{ij} gives the similarity between color i and color j. But because of the complexity of the computation, the images are preprocessed with a filter which underestimates the actual histogram distance. The filter involves a transform with orthonormal basis functions. After the transform, the dimensions are independent. Hence well established multi-dimensional spatial access methods can now be applied. Since this transform serves as a lower bound for the actual distance, there are no false dismissals. The drawback is that there are false positives, which can be eliminated by going through a verification phase.

Layout

In order to use the location information, significant color regions are extracted and their locations are recorded [HCP95, SC96b]. The regions can be represented by minimum bounding rectangles and stored in a structure like R-trees. A search on a specific color at a specific location can be performed in two steps: In the first step, two searches are performed based on color and on location. The intersection of the results of these searches yields images which satisfy both conditions.

In a slightly simplified version of this scheme [PZM96], pixels belonging to significant color regions and those that do not form two histograms, which are compared separately.

Dimensionality

Color histograms may include hundreds of colors. In order to efficiently compute histogram distances, the dimensionality should be reduced. Transform methods such as K-L and Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) or various wavelet transforms can be used to reduce the number of significant dimensions. Another idea is to find significant colors by color region extraction and compare images based on presence of significant colors only [SC96b]. Spatial partitioning strategies such as the "Pyramid-Technique" [BBK98] map n-dimensional spaces into a 1-dimensional space and use a B+ tree to manage the transformed data. The resulting access structure shows significantly better performance for large number of dimensions

compared to methods such as R-tree variants.

B. Texture and Shape Retrieval

Texture and shape differ from color in that they are defined not for individual pixels but for windows or regions. Texture segmentation involves determining the regions of image with homogeneous texture. Once the regions are determined, their bounding boxes may be used in an access structure like an R-tree. Dimensionality and cross correlation problems also apply to texture and can be solved by similar methods as in color. Fractal codes capture the self similarity of texture regions [HCA95] and are used for texture-based indexing and retrieval.

Shapes can be characterized by methods mentioned in section 2.1.1 and represented by n-dimensional numeric vectors which become points in n-dimensional space. Another approach is to approximate the shapes by well defined, simpler shapes. For instance, triangulation or a rectangular block cover can be used to represent an irregular shape. In this latter scheme, the shape is approximated by a collection of triangles or rectangles, whose dimensions and locations are recorded. The advantage of this approach is that storage requirements are lower, comparison is simpler and the original shape can be recovered with small error. The two methods can be combined to have the advantages of both.

Sketch-based retrieval can be considered a special case of shape retrieval. Here, the user may describe a single object or the whole image by the layout of objects in it. Sketch retrieval may be facilitated by an *edge map* which consists of detected and thinned edges in an image [KKOH92]. Thinning provides a binary (black and white) edge image which greatly reduces the amount of information to be stored and compared.

2.1.3 Object Detection and Recognition

Object detection involves verifying the presence of an object in an image and possibly locating it precisely for recognition. In both *feature based* and *template based* recognition, standardization of global image features and registration (alignment) of reference points are important. The images may need to be transformed to another space for handling changes in illumination, size and orientation. Both global and local features play important roles in object recognition. In *local feature-based* object recognition, one or more local features are extracted and the objects of interest are modeled in terms of these features. For instance, a human face can be modeled by the sizes of the eyes, the distance between the eye and the nose, etc. Recognition then can be transformed into a graph matching problem. In holistic or *global*

feature-based object recognition, characteristics of the object as a whole or a template of the desired object is compared against target images [JZL96]. For instance, to recognize a person, an unknown facial image (or its transform) is matched (as a whole) against (transforms of) images of known people. Psychophysical and neurophysiological studies suggest that both local and global features play a role in human face recognition [CWS95].

Transform methods such as Fourier, Wavelet or K-L also provide characteristics that can be used to detect objects of interest [CSW95, PPS94]. Many objects have unique spectra when transformed with the above methods which serves as a signature for the object. These signatures can be used to detect the presence of the object.

2.1.4 Spatial Relationships

Efficient methods for indexing and retrieving images based on the spatial relationships (such as $left\ of$, *inside*, and *above*) among objects in the image were developed [Ege93, CSY87, YM98]. Deduction of spatial relationships such as *A left of B, B left of C* $\Rightarrow$ *A left of C* are employed to retrieve images which have spatial relationships not explicitly stated in the user query. Chu et. al [HCT96] detect objects such as bones in X-rays, brain tumors and breast outlines in medical images and employ a knowledge-based image data model. The model represents selected features and spatial relationships among them in the form of a type abstraction hierarchy. The SEMCOG system [LCHH97] developed at NEC implements spatial relationship inference mechanism. Topological relationships within the context of minimum bounding rectangles are investigated in [PSTE95].

2.2 Non-visual Features

Commercial imaging systems typically use relational database technology with enhancements for image data types. In these systems, image-specific fields such as the source, the date and the time the image was taken, the media type, the resolution, the input device, the compression method, etc. as well as free text annotations are the primary non-visual features for indexing.

Captions and annotations are free text descriptions of a scene. They are natural for the users and standard text retrieval methods can be applied. However, they also present major challenges for the retrieval system. Two users may describe the same scene in very different manners. They may use different words, emphasize different aspects of the image and describe at different detail. One way to

match different descriptions of the same scene is to expand the query and the database image descriptions with an electronic thesaurus [SQ96, ATY+97]. However, the inherent ambiguity in natural language and typically short descriptions may make word sense disambiguation a difficult task [Voo93].

In order to deal with the challenges of description based retrieval, various methods have been developed such as restricting the sentence types, using inference rules, relevance feedback [HCK90, JG95, ATY+95] and structured descriptions.

Structured descriptions may be natural language sentences with restrictions, symbolic or iconic descriptions involving objects, attributes and relationships [LCHH97, ATY+97, YM98].

2.3 Querying

Possible queries involving one or more features are listed below. Combination queries may involve any number of these query primitives as long as the retrieval system supports such queries.

Simple Visual Feature Query: The user specifies certain values possibly with percentages for a feature. Example: "Retrieve images which contain 25% red, 50% blue, 25% yellow".

Feature Combination Query: The user combines different features and specifies their values and weights. Example: "Retrieve images with green color and tree texture where color has weight 75% and texture has weight 25%".

Localized feature query: The user specifies feature values and locations by placing regions on a canvas. Example: "Retrieve images with sky blue at the upper half and green at the bottom half".

Query By Example: The system generates a random set of images. The user selects one of these and retrieves similar images. Similarity can be determined based on user selected features. For instance "retrieve images which contain textures similar to this example". A slightly different version of this type of query is where the user cuts a region from an example image and pastes it onto the query canvas.

Object versus Image: The user can describe the features of an object in an image as opposed to describing a complete image. Example: "Retrieve images containing a red car near the center."

User Defined Attribute Query: The user specifies the values of the user defined attributes. Example: "Retrieve images where location is Washington DC and the date is July 4 and the resolution is at least 300dpi (dots per inch)".

Object Relationship Query: The user specifies objects, their attributes and the relationships among them. Example: "Retrieve images where an old man is holding a child in his arms".

Concept Queries: Some systems allow the user to define simple concepts based on the features extracted by the system [OS95, AHKR96]. For instance, the user may define the concept of a beach as "small yellow circle at top, large blue region in the middle and sand color in the lower half".

3 Techniques for Video Retrieval

Video retrieval involves content analysis and feature extraction, content modeling, indexing and querying. Video naturally has a hierarchy of units with individual frames at the base level and higher level *segments* such as *shots*, *scenes* and *episodes*. An important task in analyzing video content is to detect segment boundaries.

3.1 Video Segmentation

A *shot* is a sequentially recorded set of frames representing a continuous action in time and space by a single camera. A sequence of shots focusing on the same point or location of interest is a *scene*. A series of related scenes form an *episode* [WDG95]. An abrupt shot change is called a *cut*. Camera operations such as zooming, tilting, panning make it difficult to detect shot changes. Techniques for shot change detection include the following:

- **Direct Pixel or Histogram Comparison**: Pixels of consequent frames can be compared pairwise. If a significant percentage of pixels differ, a shot change is detected. This is a costly operation and is sensitive to minor camera operations like zooming. A more robust method is histogram comparison . A shot change is detected if the histograms of two consequent frames differ significantly [WKSS96]. However, this method can not handle gradual changes.

- **Compressed domain features**: Compressed video provides additional clues such as DCT transform coefficients and motion vectors which can be used for cut detection [KDL96, CSM96]. In MPEG video compression standard [Gal91] the image is compressed in units of 16x16 pixel macroblocks. The motion vectors of subsequent frames are coherent unless there is a shot change. By comparing the DCT coefficients and the motion vectors of these blocks with preceding and succeeding blocks, shot changes can be detected.

- **Text Recognition and Closed Captions for Video Indexing**: A newly emerging field is using textual information whenever available for video indexing. Optical character recognition (OCR) of the text appearing on the video or closed captions available for certain broadcast TV programs are used for segmentation and retrieval [Lie96, Moh96]. In the case of OCR on video, scene models involving likely keywords can be used for model-based segmentation. For instance, an anchor shot of a particular news agency would involve specific texts in the background and a human subject in the foreground. In the case of closed captions, text retrieval techniques can be combined with visual analysis to obtain a better semantic segmentation of the video. Video (TV) capture cards that can monitor and store closed captions, and alert the user for selected keywords are readily available (http://www.diamondmm.com/products/current/dtv-2000.cfm).

In *model based segmentation* [ZTSG95] models of different scenes are built using a-priori knowledge about the scenes. First, the video is divided into different shots by using one or more of the above techniques. Then the shots are classified based on the models. An example classification might involve anchor person shots, weather forecast shots and commercials.

3.2 Object Detection and Tracking

In video, there are two sources of information that can be used to detect and track objects: Visual features (such as color and texture) and motion information. A typical strategy is to initially segment regions based on color and texture information. After the initial segmentation, regions with similar motion vectors can be merged subject to certain constraints such as adjacency [ZKS93, JMA79].

Human skin color and DCT transform coefficients in MPEG as well as broad shape information can also be used for detecting human faces in compressed video [CSW95].

Systems for detecting particular movements such as entering, exiting a scene and placing/removing objects using motion vectors are being developed [Cou96]. It is possible to recognize certain facial expressions and gestures using models of face or hand movements.

3.3 Content Modeling, Indexing and Retrieval

The temporal nature and comparatively huge size of video data requires special browsing and querying functions. A common approach for quick browsing is to detect shot changes and associate a small icon of a key frame for each shot [NT91]. Retrieval using icons, text and image (frame) features is possible.

The hierarchical and compositional model of video [WDG95] consists of a hierarchy of segments such as shots, scenes and episodes. This model facilitates querying and composition at different levels and thus enables a very rich set of temporal and spatial operations. Example temporal operations include "follows", "contains" and "transition". Example spatial operations are "parallel to", and "below". The Hierarchical Temporal Language (HTL) [SYV97, YM98] also uses a hierarchical model of video consisting of units such as frames, shots, and sub-plots. The language uses the classical temporal operators to specify properties of video sequences as well as certain modal operators to specify properties of sequences at different levels in the hierarchy. The semantics of the language is designed for similarity-based retrieval.

Object based querying involves detection and tracking of moving objects and queries based on an example of the object provided/selected by the user [CA96].

An important criteria for the performance of a video retrieval system, or a Multimedia On Demand (MOD) system in general is the *quality of service*. Example quality of service parameters are delay jitter and the skew (synchronization difference) between the mono-media streams that make up the multimedia data. Since most users of such systems access the system via some sort of a network, continuity and the synchronization of the media streams have to be ensured under stringent communication subsystem limitations. Various buffering and disk scheduling techniques have been proposed and implemented for ensuring quality of service in such systems [Ran93, Mar97, Aur98]. A sample multimedia-on-demand system is illustrated in figure 1.

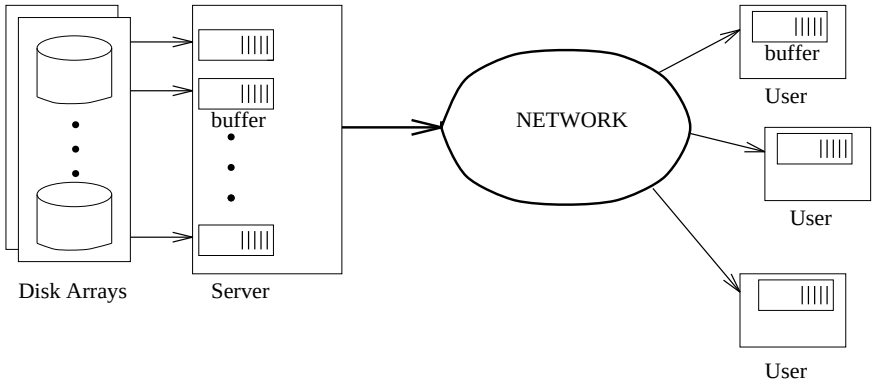

Figure 1: A sample multimedia-on-demand system with buffering and disk array replication.

4 Systems

Several research and commercial systems provide automatic indexing and querying based on visual features such as color and texture. These include Photobook, VisualSEEk, Cypress, QBIC and Virage. Certain unique features of these systems will be discussed in the following subsections.

4.1 Research Systems

The Photobook system [PPS94] enables users to plug in their own content analysis procedures and selecting among different content models based on user feedback via a learning agent. Sample applications include a face recognition system, image retrieval by texture similarity, brain map, and semi-automatic annotation based on user-given labels and visual similarity. Cypress [OS95] lets users define concepts using visual features like color. For instance, a user may coin the term "beach" for a certain combination of yellow color (sun), beige (sand), and blue (sea). VisualSEEk [SC96b] allows localized feature queries and histogram refinement for feedback using a web-based tool.

Systems such as CVEPS [CSM96], and JACOB [CA96] support automatic video segment decomposition, and video indexing based on key frames or objects. The users can employ image analysis and retrieval components to index and retrieve key frames or video objects based on their visual features or spatial layout. The CVEPS system also provide these functions as well as video editing in compressed domain. JACOB uses artificial neural networks for automatic shot detection.

Among systems using captions or annotations for image retrieval, the caption based image retrieval system of Dublin City University uses WordNet, an electronic dictionary/theasurus, for query expansion [SQ96]. Rohini and Srihari [RS95] describe a system that uses a semantic model for interpreting captions in order to guide person recognition . The SCORE system [ATY+97, ATY+95] uses an extended Entity-Relationship model to represent image contents and WordNet to expand queries as well as database descriptions. The SEMCOG system [LCHH97] performs semi-automatic object recognition.

4.2 Commercial Systems

QBIC (http://wwwqbic.almaden.ibm.com) [FBF+94], supports shape queries for semi-manually segmented objects and local features as well as global features. The Virage system (http://www.virage.com) [Gup95] supports feature layout queries and users can give different emphasis to different features. In-

formix data blades (http://www.informix.com, formerly Illustra) enable user defined content processing and analysis routines to be stored in a multimedia database. Data blades for free text, images, sound and video are becoming available by Informix and third party suppliers. Excalibur (http://www.excalib.com) Visual RetrievalWare systems enables queries on gray shape, color shape, texture, and color using adaptive pattern recognition techniques. Excalibur also provides data blades for Informix databases. An example data blade is a scene change detector for video. The data blade detects shots or scenes in video and produces a summary of the video by example frames from each shot. Oracle offers a video server which runs on top of the Oracle Universal Server product and provides concurrent delivery of full-motion video as well as remote control operations. IBM's DB2 system supports video retrieval via "video extenders"(http://www.software.ibm.com/data/db2/extenders). Video extenders allow for the import of video clips and querying these clips based on attributes such as the format, name/number or description of the video as well as last modification time.

4.3 Systems for the World Wide Web

WebSEEk [SC96a] builds several indexes for images and videos based on visual features such as color as well as non-visual features such as key terms assigned subjects and image/video types. In order to classify images and videos into subject categories, a key term dictionary is built from selected terms appearing in a URL (Uniform Resource Locator, the address of a page on the world wide web). The terms are selected based on their frequency of occurrence and whether they are meaningful subject terms. The latter judgment is made manually. For an example, the URL "http://www.chicago.com/people/michael/final98.gif" would produce the following terms: "people", "michael", "final". After the key term dictionary is built, directory portion of the image and video URLs are parsed and analyzed. This analysis produces an initial set of categories of the images and the videos which is then manually verified. Videos are summarized by picking one frame for every second of video and then packaging them as an animated GIF image.

The WebSeer project [SFA96] aims at classifying images based on their visual characteristics. Novel features of WebSeer include (1) image classification such as photographs, graphics, etc., (2) integration of CMU face detector [RBK95], and (3) multiple keyword search on associated text such as a HTTP reference, alternate text field of HTML reference or page title.

Yahoo Image Surfer (http://isurf.yahoo.com) employs Excalibur Visual RetrievalWare for searching

images and video on the WWW.

5 Looking into Future

Information overload has become almost synonymous with information age. Advanced filtering and retrieval methods for all types of multimedia data are highly needed. Current image and video retrieval systems are the results of combining research from various fields. Better collaboration of computer vision, database and user interface techniques will provide more effective and efficient image retrieval. Improved compression and object tracking techniques will increase the accessibility of digital video. One issue that needs to be addressed is the design of generic, customizable user interfaces that can be used for a variety of domains. The ability to customize the schema for image retrieval or effective visualization of video are among the objectives of such interfaces. Incorporation of voice may bring another little explored dimension to image and video retrieval. Systems that combine visual features, sound, text as well as structured descriptions will improve better user interaction.

The performance and effectiveness of multimedia database systems continue to be open issues. In fields such as Geographical Information Systems, there is a dire need for high performance multimedia databases which can support concurrent access for thousands of users while providing powerful query tools and languages. Query and transaction models of multimedia database systems differ from those of the traditional database systems. Research in these areas is likely to gain momentum as more commercial activity rely on highly available multimedia data.

While multimedia-on-demand has been on the agenda of researchers for a while, availability of such systems remains limited. As faster connection methods such as cable modems and digital subscriber lines increasingly become available for the household, a web-based killer application may be the key to increased demand for such systems. The cost issue, however, is likely to remain an obstacle for the near future.

The decreasing cost of multimedia storage and retrieval systems is encouraging medical institutions to transform their existing data into digital form. These institutions also prefer digital capture technologies for future applications. Multimedia systems are certain to play a leading role in tomorrow's clinic. Among the open problems in this domain are the use of multi-modal images for improved diagnosis, automatic feature registration for standardized object recognition and the integration of heterogenous databases.

References

[AHKR96] P. Alshuth, Th. Hermes, J. Kreyb, and M. Roper. Video Retrieval with IRIS. In *Proceedings of ACM Multimedia Conference, Boston MA*, page 421, 1996.

[ATY+95] Aslandogan, Y. A., Thier, C., Yu, T. C., Liu,C., and Nair K. Design, implementation and evaluation of SCORE (a Ssystem for COntent based REtrieval of Pictures). In *Proceedings of IEEE ICDE '95*, pages 280–287, March 1995.

[ATY+97] Aslandogan, Y. Alp, Thier, Charles, Yu, T. Clement, Zou, Jun, and Rishe, Naphtali. Using Semantic Contents and WordNet(TM) in Image Retrieval. In *Proceedings of ACM SIGIR Conference*, 1997.

[Aur98] Aurelio La Corte and Alfio Lombardo and Sergio Palazzo and Giovanni Schembra. Control of perceived quality of service in multimedia retrieval services: Prediction-based mechanisms vs. compensation buffers. *ACM/Springer Multimedia Systems*, 6:102–112, 1998.

[BBK98] Stefan Berchtold, Christian Bohm, and Hans-Peter Kriegel. The Pyramid-Technique: Towards Breaking the Curse of Dimensionality. In *Proceedings of ACM SIGMOD*, pages 142–153, 1998.

[CA96] M. La Cascia and E. Ardizzone. JACOB: Just a content-based query system for video databases. In *Proceedings of ICASSP-96, Atlanta Georgia*, pages 7–10, 1996.

[Cou96] Jonathan D. Courtney. Automatic, Object-Based Indexing for Assisted Analysis of Video Data. In *Proceedings of ACM Multimedia Conference, Boston MA*, pages 423–424, 1996.

[CSM96] Shih-Fu Chang, John R. Smith, and Jianhao Meng. Efficient Techniques for Feature-Based Image/Video Access and Manipulation. In *Proceedings of 33rd Annual Clinic on Library Applications of Data Processing Image Access and Retrieval (Invited Paper)*, March 1996.

[CSW95] S.F. Chang, J. Smith, and H. Wang. Automatic Feature Extraction and Indexing for Content-Based Visual Query . Technical Report CU/CTR 414-95-20, Columbia University, January 1995.

[CSY87] S. K. Chang, Q. Shi, and C. Yan. Iconic indexing by 2-d string. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987.

[CWS95] R. Chellapa, C. L. Wilson, and S. Sirohey. Human and Machine Recognition of Faces: A Survey. *Proceedings of the IEEE*, 83(5), May 1995.

[Ege93] Max J. Egenhofer. What's Special About Spatial? Database Requirements for Vehicle Navigation in Geographic Space. In *Proceedings of ACM SIGMOD*, pages 398–402, 1993.

[FBF+94] Faloutsos C., Barber R., Flickner M., Hafner J., Niblack W., Petkovic D., and Equitz W. Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems*, 3(1):231–262, 1994.

[Gal91] D. Le Gall. MPEG: A Video Compression Standard for Multimedia Applications. *Communications of the ACM*, (34):46–58, 1991.

[Gup95] Gupta, A. Visual Information Retrieval Technology, A VIRAGE Perspective. White paper, Virage Inc., 1995.

[Gut84] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proceedings of ACM SIGMOD Conference*, pages 47–57, June 1984.

[HCA95] D. Hang, B. Cheng, and R. Acharya. Texture-based Image Retrieval Using Fractal Codes . Technical Report 95-19, Department of Computer Science, State Unv. of New York at Buffalo, 1995.

[HCK90] Halin, G., Crehange, M., and Kerekes, P. Machine Learning and Vectorial Matching for an Image Retrieval Model: EXPRIM and the system RIVAGE. In *Proceedings of ACM-SIGIR 1990, Brussles, Belgium*, pages 99–114, 1990.

[HCP95] Wynne Hsu, T. S. Chua, and H.K. Pung. An Integrated Color-Spatial Approach to Content-Based Image Retrieval. In *Proceedings of ACM Multimedia Conference*, pages 305–313, 1995.

[HCT96] Chih-Cheng Hsu, Wesley W. Chu, and Rick K. Taira. A Knowledge-Based Approach for Retrieving Images by Content. *IEEE-TKDE*, 8:522–532, 1996.

[JG95] Jung, G.S. and Gudivada, V. Adaptive Query reformulation in Attribute based Image Retrieval. In *Intelligent Systems*, pages 763–774, 1995.

[JMA79] Ramesh Jain, W. N. Martin, and J. K. Aggarwal. Segmentation Through the Detection of Changes Due to Motion. *Computer Graphics and Image Processing*, 11:13–34, 1979.

[JZL96] A. Jain, Y. Zhong, and S. Lakshmanan. Object Matching Using Deformable Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 408–439, March 1996.

[KDL96] Vikrant Kobla, David Doermann, and King-Ip Lin. Archiving, Indexing and Retrieval of Video in the Compressed Domain. In *SPIE Multimedia Storage and Archival Systems*, pages 78–89, 1996.

[KKOH92] Kato, T., Kurita, T., Otsu, N., and Hirata, K. A Sketch Retrieval Method for Full Color Image Database, Query by Visual Example. In *IEEE-IAPR-11*, pages 530–533, August-September 1992.

[LCHH97] Wen-Syan Li, K.S. Candan, Kyoji Hirata, and Yoshinori Hara. SEMCOG: an object-based image retrieval system and its visual query interface. In *Proceedings of ACM SIGMOD*, pages 521–524, June 1997.

[Lie96] Rainer Lienhart. Automatic Text Recognition for Video Indexing. In *Proceedings of ACM Multimedia Conference, Boston MA*, pages 11–20, 1996.

[Mar97] Marwan Krunz and Satish K. Tripathi. Impact of video scheduling on bandwidth allocation for multiplexed mpeg streams. *ACM/Springer Multimedia Systems*, 5:347–357, 1997.

[Moh96] Rakesh Mohan. Text-based Search of TV News Stories. In *SPIE Multimedia Storage and Archival Systems*, pages 2–13, November 1996.

[NT91] A. Nagasaka and Y. Tanaka. Automatic Video Indexing and Full Video Search for Object Appearances. In *Working Conference on Visual Database Systems*, pages 119–133, 1991.

[OS95] Ogle, V. E. and Stonebraker, M. Chabot: Retrieval from a Relational database of Images. *IEEE Computer*, 28(9), 1995.

[PF95] E.G. M. Petrakis and C. Faloutsos. Similarity Searching in Large Image Databases. Technical Report 3388, Department of Computer Science, University of Maryland, 1995.

[PPS94] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. In *SPIE Paper 2185-05, Storage and Retrieval of Image and Video Databases II, San Jose, CA*, pages 34–47, 1994.

[PSTE95] Dimitris Papadias, Timos Sellis, Yannis Theodorakis, and Max J. Egenhofer. Topological Relations in the World of Minimum Bounding Rectangles: A Study with R-trees. In *Proceedings of ACM SIGMOD*, pages 92–103, 1995.

[PZM96] Greg Pass, Ramin Zabih, and Justin Miller. Comparing Images Using Color Coherence Vectors. In *Proceedings of ACM Multimedia Conference*, pages 65–73, 1996.

[Ran93] Rangan, P. and Vin, H. Efficient storage techniques for digital continuous media. *IEEE TKDE*, 5(4):564–573, 1993.

[RBK95] H. Rowley, S. Baluja, and K. Kanade. Human Face Detection in Visual Scenes. CMU-CS-95 158, Carnegie Mellon University, Computer Science Dept., 1995.

[RS95] Rohini and Srihari. Automatic Indexing and Content Based retrieval of Captioned Images. *IEEE Computer*, 28(9), September 1995.

[Sam89] Hanan Samet. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley, 1989.

[SC96a] John R. Smith and Shih-Fu Chang. Searching for Images and Videos on the World-Wide Web. Technical Report 459-96-25, Columbia Univeristy CTR, 1996.

[SC96b] John R. Smith and Shih-Fu Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of ACM Multimedia Conference, Boston MA*, pages 87–98, 1996.

[SFA96] Michael J. Swain, Charles Frankel, and Vassilis Athitsos. WebSeer: An Image Search Engine for the World Wide Web. Technical report, Univeristy of Chicago, Department of Computer Science, July 1996.

[SHB93] Sonka, M.and Hlavac, V. and Boyle, R. *Image Processing, Analysis, and Machine Vision.* Chapman and Hall, 1993.

[SQ96] Alan F. Smeaton and Ian Qigley. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. In *Proceedings of ACM SIGIR Conference*, 1996.

[SYV97] P. Sistla, C. T. Yu, and R. Venkatasubrahmanian. Similarity Based Retrieval of Videos. In *Proceedings of IEEE ICDE , Birmingham, UK*, pages 181–190, 1997.

[TT90] Tomita, F.and Tsuji Saburo. *Computer Analysis of Visual Textures.* Kluwer Academic Publishers, 1990.

[Voo93] Ellen M. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proceedings of ACM SIGIR Conference*, pages 171–180, 1993.

[WDG95] Ron Weiss, Andrzej Duda, and David K. Gifford. Composition and Search with a Video Algebra. *IEEE Multimedia*, pages 12–25, 1995.

[WKSS96] Howard D. Wactlar, Takeo Kanade, Michael A. Smith, and Scott M. Stevens. Intelligent Access to Digital Video: Informedia Project. *IEEE Computer*, pages 46–52, May 1996.

[YM98] Clement T. Yu and Weiyi Meng. *Principles of Database Query Processing for Advanced Applications.* Data Management Systems. Morgan Kaufmann, 1998.

[ZKS93] H.J. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic Partitioning of Full Motion Video. *Multimedia Systems*, 1(1):10–28, 1993.

[ZTSG95] H.J. Zhang, S. Y. Tan, S. W. Smoliar, and Y. Gong. Automatic Parsing and Indexing of News Video. *Multimedia Systems*, 2:256–266, 1995.